

Finding and Leveraging Structure in Learning Problems

Sivaraman Balakrishnan

CMU-LTI-13-013

August 2013

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Jaime Carbonell (Chair)

John Lafferty

Aarti Singh

Martin Wainwright (External)

Larry Wasserman

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies*

Contents

1	Introduction	7
1.1	Sparse high-dimensional learning	8
1.1.1	Learning generative models for protein fold families	8
1.1.2	Sparse additive functional and kernel CCA	8
1.2	Clustering with noisy and high-dimensional data	9
1.2.1	Noise thresholds for spectral clustering	9
1.2.2	Minimax localization of bi-clusters in large noisy matrices	10
1.2.3	Recovering block structured activation using compressive measurements	10
1.3	Statistical problems in topological data analysis	11
1.3.1	Minimax rates for homology inference	11
1.3.2	Cluster trees on manifolds	12
1.4	Organization of the thesis	12
2	Learning Generative Models for Protein Fold Families	13
2.1	Introduction	13
2.2	Graphical models for protein sequence alignments	15
2.2.1	Modeling Domain Families with Markov Random Fields	15
2.2.2	Structure learning with L_1 Regularization	17
2.3	Results	21
2.3.1	Simulations	22
2.3.2	Evaluating Structure and Parameters Jointly	23
2.3.3	A generative model for the WW domain	25
2.3.4	Allosteric regulation in the PDZ domain	27
2.3.5	Large-scale analysis of families from Pfam	29
2.3.6	Computational efficiency	31
2.4	Discussion	34
2.4.1	Related Work	34
2.4.2	Mutual Information performs poorly in the structure learning task	35
2.4.3	Influence of Phylogeny	35
2.5	Conclusions	37
2.6	Additional experiments	37
2.6.1	Comparison of structures learnt at different regularization levels	37
2.6.2	Receiver operating characteristic curve	39

3	Sparse Additive Kernel and Functional CCA	42
3.1	Introduction	42
3.2	Sparse additive kernel CCA	44
3.3	Sparse additive functional CCA	46
3.4	Marginal Thresholding	48
3.5	Main theoretical results	49
3.6	Experiments	52
3.6.1	Non-linear correlations	52
3.6.2	Marginal thresholding	54
3.6.3	Application to DLBCL data	55
3.7	Discussion	55
3.8	Technical Proofs	56
3.8.1	A derivation of the backfitting algorithm for FCCA	56
3.8.2	Uniform bounds	58
3.8.3	Marginal thresholding	60
3.8.4	Persistence	61
3.9	Additional discussion	62
3.9.1	Discussion of SA-FCCA v/s SA-KCCA	62
3.9.2	Marginal thresholding is needed to get high accuracy in high dimensions	63
3.9.3	Simulation Details	64
3.9.4	Comparison of regularization paths	64
3.9.5	Comparison of SA-FCCA and SA-KCCA on DLBCL data	64
4	Noise Thresholds for Spectral Clustering	66
4.1	Introduction	66
4.2	Related Work	67
4.3	Hierarchical Clustering	68
4.4	K-way Clustering	72
4.5	Minimax Rates	73
4.6	Experimental Results	76
4.6.1	Noise Thresholds and Asymptotic Behavior	77
4.6.2	Examples of Worst Case Behavior	77
4.6.3	Real World Experiments	78
4.7	Proofs	80
4.7.1	Proof of Theorem 4.3.1	80
4.7.2	Proof of Theorem 4.5.1	85
4.7.3	Proof of Theorem 4.5.2	86
4.8	Discussion and open problems	87
4.9	Detailed proofs	88
4.9.1	Proof of Lemma 4.7.1	88
4.9.2	Proof of Lemma 4.7.2	89
4.9.3	Proof of Lemma 4.7.3	92
4.9.4	Davis Kahan	95
4.9.5	Proof of ℓ_∞ Deviations	96

4.10	Proof of Theorem 4.4.1	98
4.11	Proofs of Information Theoretic Limits	107
4.11.1	Lower Bounds	107
4.11.2	Upper Bounds	108
4.11.3	McSherry's Algorithm	112
5	Minimax Localization of Bi-Clusters in Large Noisy Matrices	114
5.1	Introduction	114
5.1.1	Related work	117
5.2	Lower bound	118
5.3	Minimax optimal combinatorial procedure	119
5.4	Computationally efficient biclustering procedures	120
5.4.1	Element-wise thresholding	120
5.4.2	Row/Column averaging	121
5.4.3	Sparse singular value decomposition (SSVD)	121
5.5	Simulation results	123
5.6	Discussion	124
5.7	Technical proofs	125
5.7.1	Proof of Theorem 5.2.1	125
5.7.2	Proof of Theorem 5.3.1	126
5.7.3	Proof of Theorem 5.4.1	127
5.7.4	Proof of Theorem 5.4.2	128
5.7.5	Proof of Theorem 5.4.3	128
5.7.6	Proof of Theorem 5.4.4	130
5.7.7	Proof of Lemma 5.7.1	131
5.7.8	Identifying Large Biclusters Without Normality Assumption	132
5.7.9	Concentration inequalities	133
5.7.10	Convex analysis	135
5.7.11	Nuclear norm and ℓ_1 norm penalty	136
6	Recovering Block-structured Activations Using Compressive Measurements	139
6.1	Introduction	139
6.2	Preliminaries	143
6.3	Detection of contiguous blocks	145
6.3.1	Lower bound	145
6.3.2	Upper bound	146
6.4	Localization from passive measurements	146
6.4.1	Lower bound	146
6.4.2	Upper bound	147
6.5	Localization from active measurements	149
6.5.1	Lower bound	149
6.5.2	Upper bound	149
6.6	Experiments	152
6.7	Technical proofs	153

6.7.1	Proof of Theorem 6.3.1	154
6.7.2	Proof of Theorem 6.3.2	155
6.7.3	Proof of Theorem 6.5.1	155
6.7.4	Proof of Theorem 6.4.1	158
6.7.5	Proof of Theorem 6.4.2	158
6.7.6	Proof of Theorem 6.5.2	161
6.7.7	Proof of Eq. 6.7 and Eq. 6.8	164
6.7.8	Some concentration bounds	164
7	Minimax Rates for Homology Inference	167
7.1	Introduction	167
7.1.1	Related Work	168
7.2	Statistical Model	169
7.3	Homology	171
7.4	Preliminaries	173
7.4.1	Techniques for lower bounds	173
7.4.2	Techniques for upper bounds	176
7.5	Minimax Rates	177
7.5.1	Noiseless Case	177
7.5.2	Clutter Noise	179
7.5.3	Tubular Noise	181
7.5.4	Additive Noise	182
7.6	Tight lower bound	184
7.6.1	Coupon collector lower bound	184
7.6.2	Main result	184
7.7	Technical proofs	185
7.7.1	Key technical lemmas from NSW	185
7.7.2	Additional technical lemmas	195
8	Cluster Trees on Manifolds	199
8.1	Introduction	199
8.1.1	Contributions	200
8.1.2	Related Work	201
8.2	Background and Assumptions	201
8.3	Clustering on Manifolds	203
8.3.1	Technical results	205
8.3.2	Separation and Connectedness	206
8.4	A lower bound instance for the class of RSL algorithms	207
8.5	A modified algorithm for the known manifold case	209
8.6	Cluster tree recovery in the presence of noise	211
8.7	Kernel Density Estimators	213
8.7.1	Assumptions and preliminaries	213
8.7.2	Rates of convergence for the cluster tree	216
8.8	Simulations	217

8.9	Additional proofs	218
8.9.1	Volume estimates for small balls on manifolds	218
8.9.2	Bound on covering number	220
8.9.3	Uniform convergence	220
8.9.4	Sketch of the lower bound instance	221
8.9.5	Clustering with noisy samples	223
8.9.6	Proof of Theorem 8.6.1	223
8.9.7	Proof of Theorem 8.6.2	225
8.9.8	Connection radius for polynomially bounded densities	226
8.10	Discussion	227
9	Conclusions and Future Work	228
9.1	Sparse high-dimensional inference	228
9.1.1	Sparse Maximum Mean Discrepancy	228
9.1.2	Convex relaxations and sparse additive kernel PCA	229
9.1.3	Fast algorithms for additive kernel problems	231
9.2	Other statistical problems in topological data analysis	233
9.2.1	Machine learning with topological features	233
9.3	Clustering with noisy and high-dimensional data	233

List of Figures

2.1	(A) A multiple sequence alignment (MSA) for a hypothetical domain family. (B) The Markov Random Field encoding the conservation in and the coupling in the MSA. The edge between random variables X_1 and X_4 reflects the coupling between positions 1 and 4 in the MSA.	16
2.2	(A) Edge occurrence probability ρ versus F-score for the structure learning methods we propose, and the method proposed in the paper [186]. (B) L_2 norm of the error in the estimated parameters as a function of the weight of the regularization in stage two. The inset shows the case when no regularization is used in stage two. The much higher parameter estimation error in this case highlights the need for regularization in <i>both</i> stages.	23
2.3	(A) Qualitative grouping of edges missed by GREMLIN and the GMRC method (B) Sensitivity of structure learning to size of training set.	24
2.4	WW domain model. Edges returned by GREMLIN overlaid on a circle (a) and on the structure (b) of the WW domain of Transcription Elongation Factor 1 (PDB id: 2DK7) [27]. (c) Coupling profile (see text).	26
2.5	Comparison of Imputation errors on WW and PDZ families. We consider two variants of GREMLIN - with the regularization parameter selected either to produce a model with a smaller number of edges than GMRC (third bar in each group, shown in yellow) or to have zero edges on 20 permuted MSAs (last bar, shown in red). The x-intercept was chosen by estimating a lower bound on the imputation error as described in the text.	27
2.6	PDZ domain model. Edges returned by GREMLIN overlaid on a circle (a) and on the structure (b) of PDZ domain of PSD-95 (PDB id:1BE9). (c) Coupling profile (see text).	28
2.7	Histogram of MSA lengths of the 73 PFAM families in our study.	30
2.8	(A) Histogram of the distance in crystal structure. (B) Degree distribution across all proteins.	30
2.9	(A) Boxplot displaying the effect of coupling on improvement in imputation error at a position when compared to a profile-HMM. The median imputation error shows a near-linear decrease as the number of neighbors learnt by the model increases. (B) Improvement in overall imputation error across all positions for each family.	32

2.10	(A) Number of Positions in the MSA versus runtime of Neighborhood learning (in seconds) (B) Number of sequences in the MSA versus runtime of Neighborhood learning (C) Runtime of Neighborhood learning versus runtime of Parameter learning	33
2.11	(A) Adjacency matrix of a Boltzmann distribution colored by edge strength. (B) Mutual Information between positions induced by this Boltzman distribution. While the mutual information of the strongest edges is highest; a large fraction of the edges have MI comparable to many non-interactions. (C) Shows the weak ability of MI to distinguish between edges and indirect interactions in contrast to GREMLIN . AUC using MI: 0.71; AUC using GREMLIN : 0.98.	36
2.12	F-scores of structures learnt by using L_1 - L_2 norm The figure shows the average and standard deviation of the F-score across 20 different graphs as a function of ρ , the probability of edge-occurrence.	38
2.13	Graph density versus the rank correlation for ranking and selection using (A) BIC (B) AIC.	40
2.14	Receiver operating characteristic (ROC) curve of GREMLIN for the task of distinguishing artificial WW sequences that fold from those that don't.	41
3.1	Test correlations, and precision and recall for identifying relevant variables for the four different methods. SA-FCCA and SA-KCCA find strong correlations in the data, in both linear and non-linear settings. In all five data sets, SA-FCCA and SA-KCCA are always able to find the relevant variables.	52
3.2	DLBCL data : The top row shows two of the functions $f_i(X_i)$ with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ with non-zero norms for Y in blue.	53
3.3	Regularization paths for non-linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.	55
3.4	Regularization paths for linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.	65
3.5	KCCA output on DLBCL data : The top row shows two of the functions $f_i(X_i)$ v/s X_i with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ v/s Y_j with non-zero norms for Y in blue.	65
4.1	An ideal matrix for the hierarchical problem.	69
4.2	Hierachical Spectral Clustering (HS)	70
4.3	k -way Spectral Clustering (K-WAY SPECTRAL)	72
4.4	A variant of the algorithm from the paper of McSherry [138]	75
4.5	Threshold curves for the recovery of one split using HS	76
4.6	Example similarity matrices, red entries are high and blue are low, that result in undesirable behavior for Normalized Laplacians and Adjacency Matrices and Combinatorial Laplacians.	79

4.7	Experiments with real world data. (a): Heatmaps of single linkage (left) and HS (right) on gene expression data with $n = 2048$. (b) Δ -entropy scores on real world data sets.	80
4.8	All sub-matrices corresponding to sub-clusters at level 3	94
4.9	Understanding the combinatorial k -way algorithm union bound	110
5.1	Thresholding: Hamming fraction versus rescaled signal strength.	123
5.2	Averaging: Hamming fraction versus rescaled signal strength.	124
5.3	Sparse SVD: Hamming fraction versus rescaled signal strength.	124
6.1	The collection of blocks \mathcal{D}_1 is shown in solid lines and the collection \mathcal{D}_2 is shown in dashed lines. The collections \mathcal{D}_3 and \mathcal{D}_4 overlap with these and are not shown. The $(k_1 \times k_2)$ block of activation is shown in red.	152
6.2	Probability of success with passive measurements (averaged over 100 simulation runs).	153
6.3	Probability of success with adaptively chosen measurements (averaged over 100 simulation runs).	153
7.1	Relationship between chains C_p , cycles $Z_p = \ker \partial_p$ and boundaries $B_p = \text{im } \partial_{p+1}$. The chains C_p are just collections of simplices. The chains in Z_p are the cycles. The cycles in B_p are the cycles that happen to be boundaries of chains in C_{p+1}	172
7.2	The sum of two 1-cycles is another 1-cycle. Here the cycles are homologous because their sum (in \mathbb{Z}_2) is the boundary of a 2-chain of triangles.	173
7.3	A union of balls and its corresponding Čech complex.	173
7.4	The two manifolds M_1 and M_2 , with $d = 1, D = 2$	174
8.1	Robust Single Linkage (RSL) Algorithm	203
8.2	Spatially Adaptive Robust Single Linkage Algorithm	209
8.3	Figures show the average probability of success across 10 trials for different (n, d, D, ϵ)	218

List of Tables

- 3.1 Test correlation from functions estimated by SA-FCCA for $n = 75$ samples, where $Y_1 = X_1^2$, all other dimensions are Gaussian noise. Random initializations don't work well for all data sizes. Initializing with the non-sparse formulation works well when $n > p$, but fails as $p \geq n$ 49
- 3.2 Test correlations, precision and recall for identifying the correct relevant variables for the four different methods ($n = 150, p_1 = 150, p_2 = 150$). Marginal thresholding was used for selecting relevant variables before running SA-FCCA and SA-KCCA 54
- 3.3 Results for SCCA on linear data $Y_1 = X_1 + \mathcal{N}(0, 1)$ with $n = 100$ samples. As p increases, the performance of the model decreases. 63
- 3.4 Results for SA-FCCA (without marginal thresholding) on quadratic data $Y_1 = X_1^2 + \mathcal{N}(0, 1)$ with $n = 100$ samples. As p increases, the performance of the model decreases. 63
- 5.1 Bi-clustering 116
- 6.1 Summary of known results for the sparse vector case, where the length of the vector is n and the number of active elements is k . The number of measurements is m and μ/σ represents SNR per element of the activated elements. 141
- 6.2 Summary of main findings for the case when $n = n_1 \times n_2$ ($n_1 = n_2$) and $k = k_1 \times k_2$ ($k_1 = k_2$), where the size of the matrix is $n_1 \times n_2$ and the size of the activation block is $k_1 \times k_2$. The number of measurements is m and μ/σ represents SNR per element of the activated block. 142
- 7.1 Summary of our contributions 169

To my parents for their love and support.

Abstract

In the last several years we have witnessed the creation of data at an unprecedented rate and the size of datasets available in various applications has exploded. This data comes from everywhere: sensors used to gather climate information, sky survey telescopes used to collect astronomy data, customers who generate purchase records, and gene expression data from microarrays to name a few. Modern machine learning and statistics has focussed extensively on solving various inference problems involving these datasets. In this thesis we develop robust estimation procedures with theoretical guarantees for a variety of learning problems using *noisy* and *high-dimensional* data.

Learning from noisy and high-dimensional data can be impossible if we do not exploit *structure* available in the data or learning task and in this thesis we focus on understanding the *statistical* and *computational* aspects of finding and leveraging *structure* in these datasets and learning problems.

The challenges we address in this thesis broadly fall into three categories: high-dimensional sparse learning, clustering from noisy high-dimensional data and topological data analysis. In each case our main focus is on developing principled, efficient algorithms that leverage hidden structure and providing rigorous theoretical analysis of their performance. In several cases we also provide (statistical) lower bounds to establish the fundamental statistical limits for the problems we consider.

Acknowledgments

I am greatly indebted to my advisor Jaime Carbonell, for his unconditional support over the years, for giving me the freedom to pursue a variety of topics as my research interests evolved, and for always challenging me to learn more and do better research. I still remember when I first joined graduate school, not knowing what machine learning was, Jaime gave me a thesis on conditional graphical models to read. This thesis is the product of his high expectations and subtle prodding.

My perspectives on research have been deeply influenced by those of my thesis committee, especially those of John Lafferty and Larry Wasserman. I have had many technical conversations with both of them that have only made sense to me years down the road. They are an amazing combination of rigor, intuition and storehouses of knowledge, but beyond all that patient teachers and advisors nonpareil. Aarti Singh has always been an incredible person to work with. She has a knack for putting research in perspective and an ability to focus on everything from the big picture to the smallest detail, that I can only aspire to develop. I am also indebted to Aarti for her feedback on everything from papers and research ideas to talks and research statements. Aarti, John and Larry - I cannot even begin to thank you for everything.

Martin Wainwright's research and papers have inspired many chapters in this thesis. I am thankful for his feedback and insightful questions during my proposal and defense and am really excited by the prospect of working closely with him in the near future.

I am grateful to Carlos Guestrin for the many classes I took under him and for the time I spent as a teaching assistant to him. He is a great teacher and through his classes taught me almost all the machine learning, graphical models and optimization that I know. I am thankful to Alessandro Rinaldo for many interesting research discussions and various pieces of invaluable advice over the years. I am also honored to have had the support, advice and generosity of Prof. Raj Reddy at many crucial junctures.

I am thankful to my many other wonderful co-authors and friends at CMU: Akshay Krishnamurthy for trying to teach me how to work hard and play hard, Min Xu for trying to teach me the value of rigor in theory, and especially Kriti Puniyani and Mladen Kolar for many years of great friendship, great advice, inspiration and collaboration. A special mention goes out to Srivatsan Narayanan for the many hours we spent poring over various textbooks and lecture notes. There have been few, if any, research problems that I don't seek his opinion on and his insightful reasoning and questioning always help me understand my own thoughts more clearly.

I am thankful to Arthur Gretton, Samory Kpotufe, Bernhard Schölkopf and Bharath Sriperumbudur for an extremely fun summer internship at Gatsby and Max Planck. I have had several enlightening discussions with each of them. Arthur has been an extremely accessible friend and teacher over the years and has spent many hours educating me on kernels, and life. I am thankful to Alekh Agarwal, Miroslav Dudik, Dean Foster and John Langford for another fun internship at Microsoft Re-

search. MSR is a vibrant environment and I learned many things from several discussions during my internship. Alekh was the source of much inspiration over several late night coffees and Indian dinners. I am grateful to Amr Ahmed, Martin Azizyan, Xi Chen, Han Liu, Aaditya Ramdas, James Sharpnack, Matus Telgarsky and other members of the Statistical Machine Learning group for their friendship and many research conversations over the years. I am also grateful to Bob Frederking and Stacey Young for their help in all LTI matters.

During my years in graduate school I have been extremely fortunate to have many great friends and companions who have helped color my life. My graduate life without them would have been completely incomplete: Pranjali Awasthi, Ramnath Balasubramanian, Madhavi Ganapathiraju, Jose Gonzalez, Siddharth (GC) Gopal, Varun Gupta, Ravishankar Krishnaswamy, Kaushik Lakshminarayanan, Marco Molinaro, Aditya Prakash, Vyas Sekar, Vivek Seshadri, Harsha Simhadri, Ali Sinop, Ozgur Tastan, Selen Uguroglu, Gaurav Veda, Aravindan Vijayaraghavan. Thank you all!

There are two special people who I need to thank: my brother and Leman Akoglu. My brother for many years has been a fountainhead of wisdom and experience. I have followed his example or advice for everything from schools, to music, movies, and sports. Leman has been an inspiration to me academically since I first met her. Above that though she is my closest confidant, a true friend and the keeper of my sanity.

Finally, I owe the most to my parents to whom this thesis, and every achievement of mine past and future is dedicated. I look to their example at every step of my life. They are my greatest inspiration and I am eternally grateful to them for their unwavering support and belief.

Chapter 1

Introduction

In the past few years we have committed massive resources to the collection and curation of various kinds of data. Sky survey telescopes, users on the internet and genome wide association studies are only a few of the better known sources, that are generating torrential streams of data available in a variety of applications and datasets. These large datasets are often associated with two phenomena that make learning challenging:

1. The curse of dimensionality, which refers to the statistical and algorithmic intractability of systematically learning from high-dimensional data.
2. Large amounts of noise and missing information due to various measurement errors and data corruptions.

To tackle these challenges it is of utmost importance to develop principled statistical procedures that fully exploit structure in the learning problem.

The past decade has witnessed much research on sparse statistical models. Sparsity is an attractive structural assumption that can provide a route to bypass the computational and statistical curses of dimensionality typically associated with large and high-dimensional datasets. However, sparsity is not always the most appropriate notion of structure, and fortunately is not the only means to avoid the curse of dimensionality. This thesis studies sparsity and other notions of structure in an attempt to develop a better understanding and characterization of structured high-dimensional learning problems. We demonstrate that in a variety of problems exploiting structure can turn a computationally/statistically intractable learning problem into a tractable one.

Thesis statement: Finding, understanding and leveraging structure enables the principled development of flexible statistical methods for complex, noisy and high-dimensional learning problems.

In the remainder of this chapter we briefly introduce the main results that appear in this thesis. Broadly, these results are from our investigations in three areas: sparse high-dimensional learning, minimax clustering from noisy and high-dimensional data and topological data analysis.

1.1 Sparse high-dimensional learning

High-dimensional statistical inference deals with problems in which the number of model parameters p is comparable to or larger than the number of samples available n . Traditional procedures are typically not consistent in this regime and recent work has dealt with this unfavorable situation by studying sparse high-dimensional models, where many of the parameters are assumed to be 0. The assumption of sparsity has several theoretical and practical benefits: it leads to more interpretable models, reduces computational cost, and allows for model identifiability even in the high-dimensional regime.

1.1.1 Learning generative models for protein fold families

In Chapter 2 we perform an empirical study of recently proposed algorithms of Lee et al. [124], Ravikumar et al. [158] for learning sparse high-dimensional discrete graphical models. In particular we consider learning graphical models from protein multiple sequence alignments (MSAs). The resulting sparse graphical model encodes both position-specific conservation statistics and correlated mutation statistics between sequential and distant pairs of residues. These graphical models are useful from at least two distinct perspectives:

1. The graphical models we learn are generative and allow for the design of new protein sequences that have the same statistical properties as those in the multiple sequence alignment. Sequences designed this way respect covariance constraints and typically have a much higher success rate, i.e. likelihood of folding.
2. The structure of the graphical model gives insight into both sequential and long-range covariation in the multiple sequence alignment. Long-range interactions are particularly interesting because they can suggest allosteric communication [183]. Allosteric communication is the process by which signals originating at one site in a protein propagate reliably to affect distant functional sites, and one of the fundamental goals of cellular signaling is to understand this mechanism better. The structures we learn can aid in this process by suggesting candidate allosterically-coupled amino acids in a protein.

In addition to formulating the problem of learning interactions from MSAs in the framework of structure learning for graphical models we perform a detailed empirical analysis of covariation statistics on the extensively studied WW and PDZ domains (MSAs). We further apply the method to 71 additional families from the PFAM database [74], and show for instance that the learned models can significantly outperform hidden Markov models in a variety of tasks.

1.1.2 Sparse additive functional and kernel CCA

In Chapter 3 we study sparse non-parametric models for Canonical Correlations Analysis (CCA). Canonical Correlations Analysis (CCA) [97] is a classical tool for finding correlations among the components of two random vectors. In recent years, CCA has been widely applied to the analysis

of genomic data, where it is common for researchers to perform multiple assays on a single set of patient samples. Recent work of Witten et al. [205], Witten and Tibshirani [206] has proposed sparse variants of CCA to address the high dimensionality of such data. However, classical and sparse CCA are based on linear models, and are thus limited in their ability to find general correlations. In this thesis, we present two approaches to high-dimensional nonparametric CCA, building on recent developments in high-dimensional nonparametric regression.

In recent years great progress has been made in understanding sparsity for high-dimensional linear models but many problems have clear nonlinear structure. While fully non-parametric learning seems hopeless in high-dimensions, variants of additive models have been shown to be a useful compromise in many problems [20, 77, 130, 157].

In this thesis we present two approaches to sparse additive non-parametric CCA. We present estimation procedures for both approaches and analyze their theoretical properties in the high-dimensional setting. We further demonstrate the effectiveness of these procedures in discovering nonlinear correlations via extensive simulations, as well as through experiments with genomic data.

1.2 Clustering with noisy and high-dimensional data

Clustering is one of the central pre-occupations of machine learning. Broadly, the goal of clustering is to partition given data objects into groups that share some commonality.

Clustering, partly due to its unsupervised nature, is often considered a difficult topic with empirical results hard to evaluate and theoretical results hard to come by. The ability to discover meaningful clusters in high-dimensional data that is plagued with high noise, outliers and missing observations, can have a significant impact on a wide range of applications. In this thesis we consider three clustering problems in a minimax framework. In each case, we first define the clustering problem we are interested in and then establish upper and information-theoretic lower bounds on the appropriate notion of signal to noise ratio (SNR). In the minimax framework we are able to precisely characterize the fundamental limits and the performance of various popular algorithms and heuristics.

1.2.1 Noise thresholds for spectral clustering

In Chapter 4 we focus our attention on spectral clustering. Spectral clustering algorithms are a family of algorithms that partition data according to the eigenvectors of a similarity matrix formed from the data. Despite considerable empirical success, the theoretical understanding of spectral clustering is somewhat limited. In this thesis we study hierarchical and k -way spectral clustering algorithms on a general class of noisy structured similarity matrices. For hierarchical clustering, we show that recursive application of a simple spectral clustering algorithm can tolerate noise that grows with the number of data points while still recovering the hierarchical clusters

with high probability. For k -way clustering, we derive conditions on the similarity matrix under which spectral clustering perfectly partitions the data, relating the noise variance to the minimum within-cluster similarity, number of clusters, and number of data points. We complement these results with a minimax analysis, identifying the information theoretic limits for the clustering problem with tight upper and lower bounds. We verify our results with experiments on simulated and real data.

1.2.2 Minimax localization of bi-clusters in large noisy matrices

In Chapter 5 we consider the problem of identifying a sparse set of relevant columns and rows in a large data matrix with highly corrupted entries. This problem of identifying groups from a collection of bipartite variables such as proteins and drugs, biological species and gene sequences, malware and signatures, etc is commonly referred to as biclustering or co-clustering. Despite its great practical relevance, and although several ad-hoc methods are available for bi-clustering, theoretical analysis of the problem is largely non-existent. We study bi-clustering in a theoretical model that is closely related to that of structured normal means problems [2, 11, 12], an area of statistics that has recently witnessed much activity.

In this chapter we prove lower bounds on the minimum signal strength needed for successful recovery of a bi-cluster as a function of the noise variance, size of the matrix and bi-cluster of interest. We show that a combinatorial procedure based on the scan statistic achieves this optimal limit. We characterize the SNR required by several computationally tractable procedures for bi-clustering including element-wise thresholding, column/row average thresholding and a convex relaxation approach to sparse singular vector decomposition.

1.2.3 Recovering block structured activation using compressive measurements

In Chapter 6, we consider the problems of detection and localization of a contiguous block of weak activation in a large matrix, from a small number of noisy, possibly adaptive, compressive (linear) measurements. This is closely related to the problem of compressed sensing, where the task is to estimate a sparse vector using a small number of linear measurements. Contrary to results in compressed sensing, where it has been shown that neither adaptivity nor contiguous structure help much, we show that for reliable localization the magnitude of the weakest signals is strongly influenced by both structure and the ability to choose measurements adaptively while for detection neither adaptivity nor structure reduce the requirement on the magnitude of the signal. We characterize the precise tradeoffs between the various problem parameters, the signal strength and the number of measurements required to reliably detect and localize the block of activation. The sufficient conditions are complemented with information theoretic lower bounds.

1.3 Statistical problems in topological data analysis

Recently, there has been considerable interest in developing and understanding the mathematical formalism of topological data analysis (TDA). TDA is a field at the intersection of statistics, computational geometry and topology and aims to incorporate geometric and topological techniques into the study of point clouds, i.e. finite sets of points not necessarily in Euclidean space (although we will not consider the more abstract case in this thesis) equipped with a distance function.

These point clouds are intended to be thought of as finite samples taken from a geometric object, perhaps with noise. For example a basic problem is to compute approximations to the homology, or local coordinate charts of the manifold given finite samples. Understanding the sample complexity of these tasks is a central problem in this area.

The manifold hypothesis, that high-dimensional data often lie on or near a low-dimensional smooth manifold, has been central to much of machine learning, for example to understanding Laplacian based regularization and fast rates of convergence for kernel regression. In tasks of a more exploratory nature however it is of interest to understand the geometry of the manifold itself. Often this geometry is not easy to understand or visualize and one option is to resort to *summaries* of the manifold or the distribution on the manifold. Broadly speaking, TDA is the study of these summaries.

In this thesis we consider two problems in this vein. The first problem is that of finding the homology of a manifold from random samples on or close to the manifold, and the second problem is to understand the cluster tree of a distribution supported on or near the manifold.

1.3.1 Minimax rates for homology inference

Often, high dimensional data lie close to a low-dimensional sub-manifold and it is of interest to understand the geometry of these sub-manifolds. The homology groups of a manifold are important topological invariants that provide an algebraic summary of the manifold. These groups contain rich topological information, for instance, about the connected components, holes, tunnels and sometimes the dimension of the manifold. In this thesis, we consider the statistical problem of estimating the homology of a manifold from noisy samples under several different noise models. We derive upper and lower bounds on the minimax risk for this problem. Our upper bounds are based on estimators which are constructed from a union of balls of appropriate radius around carefully selected points. In each case we establish complementary lower bounds using Le Cam's lemma. Finally, we show tight asymptotic minimax lower bounds by a direct analysis of the likelihood ratio test. Under a variety of noise models our results show that it is possible to infer the homology at ambient dimension independent rates, indicating that often these flexible invariants can be estimated from very few random samples.

1.3.2 Cluster trees on manifolds

In Chapter 8 we investigate the problem of estimating the cluster tree for a density f supported on or near a smooth d -dimensional manifold M isometrically embedded in \mathbb{R}^D . We analyze a modified version of a k -nearest neighbor based algorithm recently proposed by Chaudhuri and Dasgupta [44].

Our main results show that under mild assumptions on f and M , we obtain rates of convergence that depend on d only but not on the ambient dimension D . We also show that similar (albeit non-algorithmic) results can be obtained for kernel density estimators. We sketch a construction of a sample complexity lower bound instance for a natural class of manifold oblivious clustering algorithms. We further briefly consider the known manifold case and show that in this case a spatially adaptive algorithm achieves better rates.

1.4 Organization of the thesis

1. Chapters 2 and 3 consider two sparse high-dimensional learning problems. In Chapter 2 we consider the problem of structure learning in discrete MRFs from protein MSAs. In Chapter 3 we develop flexible sparse non-parametric estimators for high-dimensional canonical correlations analysis and study some theoretical properties of these estimators. The results of these chapters appear in the papers [17, 20].
2. Chapters 4, 5 and 6 consider three clustering problems from noisy high-dimensional measurements. In Chapter 4 we provide a novel analysis of spectral clustering applied to noisy structured similarity matrices, and provide minimax rates for the problem. In Chapter 5 we consider the problem of recovering a sub-matrix of activation from a noisy high-dimensional matrix. We provide minimax rates, study computationally efficient procedures and characterize some of the statistical-computational tradeoffs for this problem. In Chapter 6 we study the problem of recovering block-structured activations using compressive measurements. We characterize the minimax limits for both active and passive measurement schemes in this problem. These chapters are based on the papers [18, 22, 111].
3. Chapters 7 and 8 study problems related to topological data analysis. In Chapter 7 we consider the problem of estimating the homology of a manifold from noisy samples, and in Chapter 8 we consider the problem of learning the cluster tree of a density supported on or near a manifold. In each case we provide simple estimators and analyze their rates of convergence. In the case of homology estimation we also derive minimax lower bounds. The results of these chapters appear in the papers [19, 21].
4. Finally Chapter 9 presents some conclusions and some avenues for future investigation into the topics of the thesis.

Chapter 2

Learning Generative Models for Protein Fold Families

In this chapter we introduce a new approach to learning statistical models from multiple sequence alignments (MSA) of proteins. The method we introduce, called GREMLIN (Generative REGularized ModeLS of proteINs), learns an undirected probabilistic graphical model of the amino acid composition within the MSA. The resulting model encodes both the position-specific conservation statistics *and* the correlated mutation statistics between sequential and long-range pairs of residues. Existing techniques for learning graphical models from multiple sequence alignments either make strong, and often inappropriate assumptions about the conditional independencies within the MSA (e.g., Hidden Markov Models), or else use sub-optimal algorithms to learn the parameters of the model. In contrast, GREMLIN makes no *a priori* assumptions about the conditional independencies within the MSA. We formulate and solve a *convex* optimization problem, thus guaranteeing that we find a *globally optimal* model at convergence. The resulting model is also generative, allowing for the design of new protein sequences that have the same statistical properties as those in the MSA. We perform a detailed analysis of covariation statistics on the extensively studied WW and PDZ domains and show that our method out-performs an existing algorithm for learning undirected probabilistic graphical models from MSA. We then apply our approach to 71 additional families from the PFAM database and demonstrate that the resulting models significantly out-perform Hidden Markov Models in terms of predictive accuracy.

2.1 Introduction

A protein family¹ is a set of evolutionarily related proteins descended from a common ancestor, generally having similar sequences, three dimensional structures, and functions. By examining the statistical patterns of sequence conservation and diversity within a protein family, we can gain insights into the constraints that determine structure and function. These statistical patterns are

¹In this thesis, the expression *protein family* is synonymous with *domain family*.

often learned from multiple sequence alignments (MSA) and then encoded using probabilistic graphical models (e.g., [74, 108, 109, 119, 132]). The well-known database PFAM [74], for example, contains more than 11,000 profile Hidden Markov Models (HMM) [67] learned from MSAs. The popularity of generative graphical models is due in part to the fact that they can be used to perform important tasks such as structure and function classification (e.g., [109, 132]) and to design new protein sequences (e.g., [188]). Unfortunately, existing methods for learning graphical models from MSAs either make unnecessarily strong assumptions about the nature of the underlying distribution over protein sequences, or else use greedy algorithms that are often sub-optimal. A goal of this chapter of the thesis is to introduce a new algorithm that addresses these two issues simultaneously and to demonstrate the superior performance of the resulting models.

A graphical model encodes a probability distribution over protein sequences in terms of a graph and a set of functions. The nodes of the graph correspond to the columns of the MSA and the edges specify the *conditional independencies* between the columns. Each node is associated with a local function that encodes the column-specific conservation statistics. Similarly, each edge is associated with a function that encodes the correlated mutation statistics between pairs of residues.

The task of learning a graphical model from an MSA can be divided into two sub-problems: (i) learning the topology of the graph (i.e., the set of edges), and (ii) estimating the parameters of the functions. The first problem is especially challenging because the number of unique topologies on a graph consisting of n nodes is $O(2^{n^2})$. For that reason, it is common to simply *impose* a topology on the graph, and then focus on parameter estimation. An HMM, for example, has a simple topology where each column is connected to its immediate neighbors. That is, the model assumes each column is conditionally independent of the rest of the MSA, given its sequential neighbors. This assumption dramatically reduces the complexity of learning the model but is not well justified biologically. In particular, it has been shown by Ranganathan and colleagues that it is necessary to model correlated mutations between non-adjacent residues [133, 165, 174].

Thomas et al. [185] demonstrated that correlated mutations between non-adjacent residues can be efficiently modeled using a different kind of graphical model known as a Markov Random Field (MRF). However, when using MRFs one must first identify the conditional independencies within the MSA. That is, one must learn the topology of the model. Thomas and colleagues address that problem using a greedy algorithm, called GMRC, that adds edges between nodes with high mutual information [185, 186, 187, 189]. Unfortunately, their algorithm provides no guarantees as to the optimality of the resulting model.

The algorithm presented in this chapter, called GREMLIN (Generative REgularized ModelS of proteINs), solves the same problem as the paper of Thomas et al. [185] but does so using a method with strong theoretical guarantees. In particular, our algorithm is *consistent*, i.e. it is guaranteed to yield the true model as the data increases, and it has low *sample-complexity*, i.e. it requires less data to identify the true model than any other known approach. GREMLIN also employs *regularization* to penalize complex models and thus reduce the tendency to over-fit the data. Finally, our algorithm is also computationally efficient and easily parallelizable. We demonstrate GREMLIN by performing a detailed analysis on the well-studied WW and PDZ domains and

demonstrate that it produces models with higher predictive accuracy than those produced using the GMRC algorithm. We then apply GREMLIN to 71 other families from the PFAM database and show that our algorithm produces models with consistently higher predictive accuracy than profile HMMs.

2.2 Graphical models for protein sequence alignments

In what follows, we briefly describe our approach to modeling protein multiple sequence alignments using Markov Random Field our approach to learning the statistical patterns within a given multiple sequence alignment. The resulting model is a probability distribution over amino acid sequences for a particular domain family.

2.2.1 Modeling Domain Families with Markov Random Fields

Let X_i be a finite discrete random variable representing the amino-acid composition at position i of the MSA of the domain family taking values in $\{1\dots k\}$ where the number of states, k , is 21 (20 amino acids with one additional state corresponding to a gap). Let $\mathbb{X} = \{X_1, X_2, \dots, X_p\}$ be the multi-variate random variable describing the amino acid composition of an MSA of length p . Our goal is to model $P(\mathbb{X})$, the amino-acid composition of the domain family.

Unfortunately, $P(\mathbb{X})$ is a distribution over a space of size k^p , rendering the explicit modeling of the joint distribution computationally intractable for naturally occurring domains. However, by exploiting the properties of the distribution, one can significantly decrease the number of parameters required to represent this distribution. To see the kinds of properties that we can exploit, let us consider a toy domain family represented by an MSA as shown in Fig. 2.1-(A). A close examination of the MSA reveals the following statistical properties of its composition: (i) the Tyrosine ('Y') at position 2 is conserved across the family; (ii) positions 1 and 4 are co-evolving – sequences with a (S) at position 1 have a Histidine (H) at position 4, while sequences with a Phenylalanine (F) at position 1 have a Tryptophan (W) at position 4; (iii) the remaining positions appear to evolve independently of each other. In probabilistic terms we say that X_1, X_3 are co-varying, and that the remaining X_i 's are statistically independent. We can therefore encode the joint distribution over all positions in the MSA by storing one joint distribution $P(X_1, X_4)$, and the univariate distributions $P(X_i)$, for the remaining positions (since they are all statistically independent of every other variable).

The ability to factor the full joint distribution, $P(\mathbb{X})$, in this fashion has an important consequence in terms of space complexity. Namely, we can reduce the space requirements from 21^7 to $21^2 + 7 * 21$ parameters. This drastic reduction in space complexity translates to a corresponding reduction in time complexity for computations over the distribution. While this simple example utilizes independencies in the distribution; this kind of reduction is possible in the more general case of *conditional independencies*. A Probabilistic Graphical Model (PGM) exploits these (con-

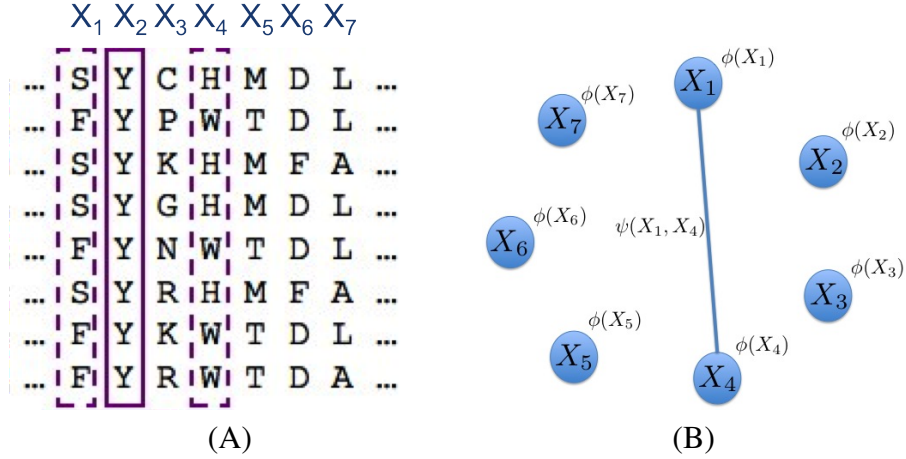


Figure 2.1: (A) A multiple sequence alignment (MSA) for a hypothetical domain family. (B) The Markov Random Field encoding the conservation in and the coupling in the MSA. The edge between random variables X_1 and X_4 reflects the coupling between positions 1 and 4 in the MSA.

ditional) independence properties to store the joint probability distribution using a small number of parameters.

Intuitively, a PGM stores the joint distribution of a multivariate random variable in a graph; while any distribution can be modeled by a PGM with a complete graph, exploiting the conditional independencies in the distribution leads to a PGM with a (structurally) sparse graph. We use a specific type of probabilistic graphical model called a Markov Random Field (MRF). In its commonly defined form with pair-wise log-linear potentials, a Markov Random Field (MRF) can be formally defined as a tuple $\mathcal{M} = (\mathbb{X}, \mathcal{E}, \Phi, \Psi)$ where $(\mathbb{X}, \mathcal{E})$ is an undirected graph over the random variables. \mathbb{X} represents the set of vertices and \mathcal{E} is the set of edges of the graph. The graph succinctly represents conditional independencies through its Markov properties, which state for instance that each node is independent of all other nodes given its neighbors. Thus, graph separation in $(\mathbb{X}, \mathcal{E})$ implies conditional independence. Φ, Ψ are a set of node and edge potentials, respectively, usually chosen to be log-linear functions of the form:

$$\phi_s = [e^{v_1^s} \ e^{v_2^s} \ \dots \ e^{v_k^s}]; \quad \psi_{st} = \begin{bmatrix} e^{w_{11}^{st}} & e^{w_{12}^{st}} & \dots & e^{w_{1k}^{st}} \\ e^{w_{21}^{st}} & e^{w_{22}^{st}} & \dots & e^{w_{2k}^{st}} \\ \dots & \dots & \dots & \dots \\ e^{w_{k1}^{st}} & e^{w_{k2}^{st}} & \dots & e^{w_{kk}^{st}} \end{bmatrix}, \quad (2.1)$$

where s is a position in the MSA, and (s, t) is an edge between the positions s and t in the MSA. ϕ_s is a $(k \times 1)$ vector and ψ_{st} is a $(k \times k)$ matrix. For future notational simplicity we further

define

$$\mathbf{v}^s = [v_1^s \ v_2^s \ \dots \ v_k^s] \quad \text{and} \quad \mathbf{w}^{st} = \begin{bmatrix} w_{11}^{st} & w_{12}^{st} & \dots & w_{1k}^{st} \\ w_{21}^{st} & w_{22}^{st} & \dots & w_{2k}^{st} \\ \dots & \dots & \dots & \dots \\ w_{k1}^{st} & w_{k2}^{st} & \dots & w_{kk}^{st} \end{bmatrix}, \quad (2.2)$$

where \mathbf{v}^s is a $(k \times 1)$ vector and \mathbf{w}^{st} is a $(k \times k)$ matrix. $\mathbf{v} = \{\mathbf{v}^s | s = 1 \dots p\}$ and $\mathbf{w} = \{\mathbf{w}^{st} | (s, t) \in \mathcal{E}\}$ are node and edge “weights”. \mathbf{v} is a collection of p , $(k \times 1)$ vectors and \mathbf{w} is a collection of p , $(k \times k)$ matrices.

The probability of a particular sequence $x = \{x_1, x_2, \dots, x_p\}$ according to \mathcal{M} is defined as:

$$P_{\mathcal{M}}(x) = \frac{1}{Z} \prod_{s \in V} \phi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t), \quad (2.3)$$

where Z , the so-called partition function, is a normalizing constant defined as a sum over all possible assignments to \mathbb{X} . Abusing notation slightly we have,

$$Z = \sum_{X \in \mathbb{X}} \prod_{s \in V} \phi_s(X_s) \prod_{(s,t) \in E} \psi_{st}(X_s, X_t). \quad (2.4)$$

The structure of the MRF for the MSA shown in Fig. 2.1(A) is shown in Fig. 2.1(B). The edge between variables X_1 and X_4 reflects the statistical coupling between those positions in the MSA.

2.2.2 Structure learning with L_1 Regularization

In the previous section we outlined how an MRF can parsimoniously model the probability distribution $P(\mathbf{X})$. In this section we consider the problem of *learning* the MRF from an MSA.

Eq. 2.3 describes the probability of a sequence X for a specific model \mathcal{M} . Given a set of independent sequences $\mathcal{X} = \{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \dots, \mathbf{X}^n\}$, the log-likelihood of the model parameters $\Theta = (\mathcal{E}, \mathbf{v}, \mathbf{w})$ is then:

$$\mathfrak{ll}(\Theta) = \frac{1}{n} \sum_{X^i \in \mathcal{X}} \left[\sum_{s \in V} \log \phi_s(X_s^i) + \sum_{(s,t) \in E} \log \psi_{st}(X_s^i, X_t^i) \right] - \log Z, \quad (2.5)$$

where the term in the braces is the unnormalized likelihood of each sequence, and Z is the global partition function. The problem of learning the structure *and* parameters of the MRF is now simply that of maximizing $\mathfrak{ll}(\Theta)$,

$$MLE(\theta) = \max_{\Theta} \mathfrak{ll}(\Theta). \quad (2.6)$$

This Maximum Likelihood Estimate (MLE) is guaranteed to recover the true parameters as the amount of data increases. However, this formulation suffers from two significant shortcomings:

(i) the likelihood involves the computation of the global partition function which is computationally intractable and requires $O(k^p)$ time to compute, and (ii) in the absence of infinite data, the MLE can significantly over-fit the training data due to the potentially large number of parameters in the model.

An overview of our approach to surmount these shortcomings is as follows: first, we approximate the likelihood of the data with an objective function that is easier to compute, yet retains the optimality property of MLE mentioned above. To avoid over-fitting and learning densely connected structures, we then add a regularization term that penalizes complex models to the likelihood objective. The specific regularization we use is particularly attractive because it has high statistical efficiency.

The general regularized learning problem is then formulated as:

$$\max_{\Theta} \text{pll}(\Theta) - \mathbf{R}(\Theta) \tag{2.7}$$

where the pseudo log-likelihood $\text{pll}(\Theta)$ is an approximation to the exact log-likelihood and $\mathbf{R}(\Theta)$ is a regularization term that penalizes complex models.

While this method can be used to jointly estimate both the structure \mathcal{E} and the parameters \mathbf{v}, \mathbf{w} , it will be convenient to divide the learning problem into two parts: (i) *structure learning* — which learns the edges of the graph, and (ii) *parameter estimation* — learning \mathbf{v}, \mathbf{w} given the structure of the graph. We will use a regularization penalty in the structure learning phase that focuses on identifying the correct set of edges. In the parameter estimation phase, we use these edges and learn \mathbf{v} and \mathbf{w} using a different regularization penalty that focuses on estimating \mathbf{v} and \mathbf{w} accurately. We note that once the set of edges has been fixed, the parameter estimation problem can be solved efficiently. Thus, we will focus on the problem of learning the edges or, equivalently, the set of conditional independencies within the model.

Pseudo Likelihood

The log-likelihood as defined in Eq. 2.5 is smooth, differentiable, and concave. However, maximizing the log-likelihood requires computing the global partition function Z and its derivatives, which in general can take up to $\mathcal{O}(k^p)$ time. While approximations to the partition function based on Loopy Belief Propagation [125] have been proposed as an alternative, such approximations can lead to inconsistent estimates.

Instead of approximating the true-likelihood using approximate inference techniques, we use a different approximation based on a pseudo-likelihood proposed by Besag [28], and used in the papers of Schmidt et al. [167], Wainwright et al. [199]. The pseudo-likelihood is defined as:

$$\begin{aligned}
\text{pll}(\Theta) &= \frac{1}{n} \sum_{X^i \in \mathcal{X}} \sum_{j=1}^p \log(P(X_j^i | X_{-j}^i)) \\
&= \frac{1}{n} \sum_{X^i \in \mathcal{X}} \sum_{j=1}^p \left[\log \phi_j(X_j^i) + \sum_{k \in V'_j} \log \psi_{jk}(X_j^i, X_k^i) - \log Z_j \right]
\end{aligned}$$

where X_j^i is the residue at the j^{th} position in the i^{th} sequence of our MSA, X_{-j}^i denotes the ‘‘Markov blanket’’ of X_j^i , and Z_j is a local normalization constant for each node in the MRF. The set V'_j is the set of all vertices which connect to vertex j in the PGM. The only difference between the likelihood and pseudo-likelihood is the replacement of a global partition function with local partition functions (which are sums over possible assignments to single nodes rather than a sum over all assignments to *all* nodes of the sequence). This difference makes the pseudo-likelihood significantly easier to compute in general graphical models.

The pseudo-likelihood retains the concavity of the original problem, and this approximation makes the problem tractable. Moreover, this approximation is known to yield a consistent estimate of the parameters under fairly general conditions if the generating distribution is in fact a pairwise MRF defined by a graph over \mathbb{X} [82]. That is, under these conditions, as the number of samples increases, parameter estimates using pseudo-likelihood converge to the true parameters.

L1 Regularization

The study of convex approximations to the complexity and goodness of fit metrics has received considerable attention recently [96, 125, 167, 199]. Of these, those based on L_1 regularization are the most interesting because of their strong theoretical guarantees. In particular methods based on L_1 regularization exhibit consistency in both parameters and structure (i.e., as the number of samples increases we are guaranteed to find the true model), and high statistical efficiency (i.e., the number of samples needed to achieve this guarantee is small). See the paper of Tropp [191] for a recent review of L_1 -regularization. Our algorithm uses L_1 -regularization for both structure learning and parameter estimation.

For the specific case of block- L_1 regularization, $R(\Theta)$ usually takes the form:

$$R(\Theta) = \lambda_{node} \sum_{s=1}^p \|\mathbf{v}^s\|_2^2 + \lambda_{edge} \sum_{s=1}^p \sum_{t=s+1}^p \|\mathbf{w}^{st}\|_2, \quad (2.8)$$

where λ_{node} and λ_{edge} are regularization parameters that determine how strongly we penalize higher (absolute) weights. The value of λ_{node} and λ_{edge} control the trade-off between the log-likelihood term and the regularization term in our objective function.

The regularization described above groups all the parameters that describe an edge together in a *block*. The second term in Eq. 2.8 is the sum of the L_2 norms of each block. Since the L_2 norm is always positive, our regularization is exactly equivalent to penalizing the L_1 norm of the vector of norms of each block with the penalty increasing with higher values of λ_{edge} . It is important to distinguish the block- L_1 regularization on the edge weights from the more traditional L_2 regularization on the node weights where we sum the *squares* of the L_2 norms.

The L_1 norm is known to encourage sparsity (by setting parameters to be exactly zero), and the *block* L_1 norm we have described above encourages group sparsity (where *groups* of parameters are set to zero). Since, each group corresponds to all the parameters of a single edge, using the block L_1 norm leads to what we refer to as structural sparsity (i.e. sparsity in the edges). In contrast, the L_2 regularization also penalizes high absolute weights, but does not usually set any weights to zero, and thus does not encourage sparsity.

Optimizing Regularized Pseudo-Likelihood

In the previous two sections we described an objective function, and then a tractable and consistent approximation to it, given a set of weights (equivalently, potentials). However, to solve this problem we still need to be able to find the set of weights that maximizes the likelihood under the block-regularization form of Eq. 2.7. We note that the objective function associated with block- L_1 regularization is no longer smooth. In particular, its derivative with respect to any parameter is discontinuous at the point where the group containing the parameter is 0. We therefore consider an equivalent formulation where the non-differentiable part of the objective is converted into a constraint making the new objective function differentiable,

$$\begin{aligned} & \max_{\Theta, \alpha} \text{pll}(\Theta) - \lambda_{node} \sum_{s=1}^p \|\mathbf{v}^s\|_2^2 - \lambda_{edge} \sum_{s=1}^p \sum_{t=s+1}^p \alpha_{st} \\ \text{subject to:} & \quad \forall (1 \leq s < t \leq p) : \alpha_{st} \geq \|\mathbf{w}^{st}\|_2, \end{aligned}$$

where the constraints hold with equality at the optimal (Θ, α) . Intuitively, α_{st} behaves as a differentiable proxy for the non-differentiable $\|\mathbf{w}^{st}\|_2$, making it possible to solve the problem using techniques from smooth convex optimization. Since the constraints hold with equality at the optimal solution (ie $\alpha_{st} = \|\mathbf{w}^{st}\|_2$), the solutions and therefore, the formulations are identical.

We solve this reformulation through the use of projected gradients. We first ignore the constraints, compute the gradient of the objective, and take a step in this direction. If the step results in any of the constraints being violated we solve an alternative (and simpler) Euclidean projection problem:

$$\begin{aligned} & \min_{\Theta', \alpha'} \left\| \begin{bmatrix} \Theta' \\ \alpha' \end{bmatrix} - \begin{bmatrix} \Theta \\ \alpha \end{bmatrix} \right\|_2^2 \\ \text{subject to:} & \quad \forall (1 \leq s < t \leq p) : \alpha_{st} \geq \|\mathbf{w}^{st}\|_2 \end{aligned}$$

which finds the closest parameter vector to the vector obtained by taking the gradient step (in Euclidean distance), which satisfies the original constraints. In this case the projection problem can be solved extremely efficiently (in linear time) using an algorithm described in Schmidt et al. [167]. Methods based on projected gradients are guaranteed to converge to a stationary point (see Boyd and Vandenberghe [33]), and convexity ensures that this stationary point is globally optimal.

In order to scale the method to significantly larger domains, we can sub-divide the structure learning problem into two steps. In the first step, each node is considered separately to identify its neighbors. This may lead to an asymmetric adjacency matrix, and so in the second step the adjacency matrix is made symmetric. This two-step approach to structure learning has been extensively compared to the single step approach by Hoeffling and Tibshirani [96] and has been found to have almost identical performance. The two-step approach however has several computational advantages. The problem of learning the neighbors of a node is exactly equivalent to solving a logistic regression problem with block- L_1 regularization, and this problem can be solved quickly and with low memory requirements. Additionally, the problem of estimating the graph can now be trivially parallelized across nodes of the graph since these logistic regression problems are completely decoupled. Parameter learning of the graph with just L_2 regularization can then be solved *extremely* efficiently using quasi-Newton methods [129].

2.3 Results

The probabilistic framework defined in Sec. 2.2.1 and the optimization objectives and algorithms defined in Sec. 2.2.2 constitute a method for learning a graphical model from a given MSA. The optimization framework has two major penalty parameters that can be varied (λ_v, λ_e). To understand the effects of these parameters, we first evaluated GREMLIN on artificial protein families whose sequence records were generated from known, randomly generated models. This lets us evaluate the success of the various components of GREMLIN in a controlled setting where the ground truth was known.

Our experiments involve comparing the performance of ranking edges and learning a graph structure using a variety of techniques, including: (i) our algorithm, GREMLIN; (ii) the greedy algorithm of Thomas et al. [185, 186], denoted 'GMRC method'; and (iii) a simpler greedy algorithm that uses the metric suggested in the paper of Lockless and Ranganathan [133], denoted $\Delta\Delta G^{stat}$. We also compare our performance with the Profile Hidden Markov Models [67] used by Finn et al. [74].

We note that the GMRC method only considers edges that meet certain coupling criteria (see the papers of Thomas et al. [185, 186] for details). In particular, we found that it returns sparse graphs (fewer than 100 edges), regardless of choice of run-time parameters. GREMLIN, in contrast, returns a full spectrum from disconnected to completely connected graphs depending on the choice of the regularization parameter. In our experiments, we use our parameter estimation code on their graphs, and compare ourselves to the best graph they return.

In the remainder of this section, we demonstrate that GREMLIN significantly out-performs other algorithms. In particular, we show that GREMLIN achieves higher goodness of fit to the test set, and has lower prediction error than the GMRC method - *even when we learn models of similar sparsity*. Finally, we show that GREMLIN also significantly out performs profile HMM-based models for 71 real protein families, in terms of goodness of fit. These results demonstrate that the use of block-regularized structure learning algorithms can result in higher-quality MRFs than those learnt by the GMRC method, and that MRFs produce higher quality models than HMMs.

2.3.1 Simulations

We generated 32-node graphs. Each node had a cardinality of 21 states, and each edge was included with probability ρ . Ten different values of ρ varying from 0.01 and 0.45 were used; for each value of ρ , twenty different graphs were generated resulting in a total of 200 graphs. For each edge that was included in a graph, edge and node weights were drawn from a Normal distribution (weights $\sim \mathcal{N}(0,1)$). Since each edge involves sampling 441 weights from this distribution, the edges tend to have many small weights and a few large ones. This reflects the observation that in positions with known correlated mutations, a few favorable pairs of amino acids are usually much more frequent than most other pairs. When we sample from our simulated graphs using these parameters, we therefore tend to generate such sequences.

For each of these 200 graphical models, we then sampled 1000 sequences using a Gibbs sampler with a burn-in of 10,000 samples and discarding 1,000 samples between each accepted sequence. These 1000 sequences were then partitioned into two sets: a training set containing 500 sequences and a held-out set of 500 sequences used to test the model. The training set was then used to train a model using the block regularization norm.

We first test our accuracy on structure learning. We measure accuracy by the F-score which is defined as

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

Precision and recall are in turn defined in terms of the number of true positives (tp), false positives (fp) and false negatives (fn) as $\text{precision} = \frac{tp}{tp+fp}$ and $\text{recall} = \frac{tp}{tp+fn}$.

Since the structure of the model directly depends only on the regularization weight on the edges, the structures were learnt for each norm and each training set with different values of λ_e (between 1 and 500), keeping λ_v fixed at 1.

Figure 2.2-A compares our structure learning method with the algorithm in the paper of Thomas et al. [186]. We evaluate their method over a wide range of parameter settings and select the best model. Figure 2.2-A shows that our method significantly out-performs their method for *all* values of ρ . We see that over all settings our best model has an average F-score of *at least* 0.63. We conclude that we are able to infer accurate structures given the proper choice of settings.

Figure 2.2-B, shows the error in our parameter estimates given the true graph as a function of

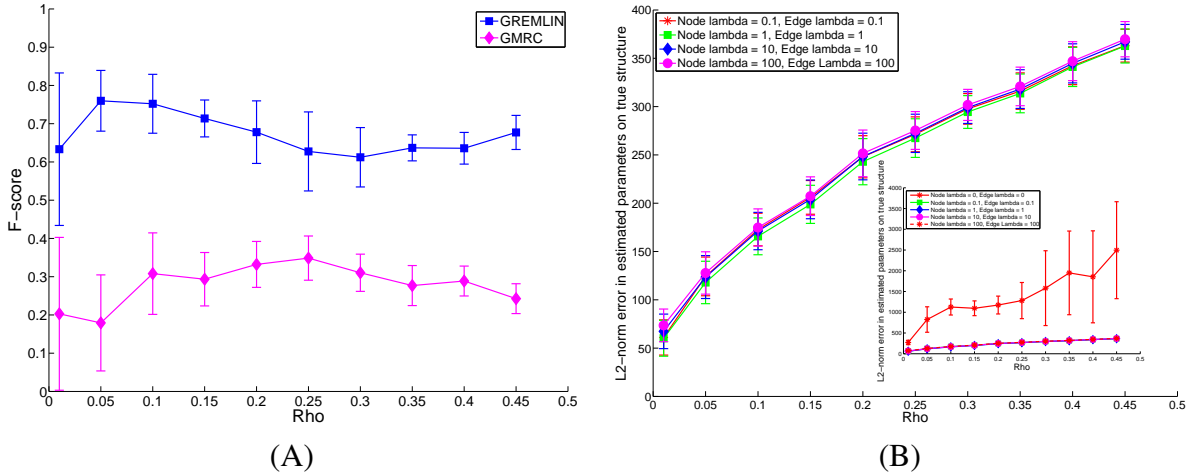


Figure 2.2: (A) Edge occurrence probability ρ versus F-score for the structure learning methods we propose, and the method proposed in the paper [186]. (B) L_2 norm of the error in the estimated parameters as a function of the weight of the regularization in stage two. The inset shows the case when no regularization is used in stage two. The much higher parameter estimation error in this case highlights the need for regularization in *both* stages.

ρ . We also find that parameter estimation is reasonably robust to the choice of the regularization weights, as long as the regularization weights are non-zero.

Fig. 2.3-A shows a qualitative analysis of edges missed by each method (we consider all simulated graphs and the best learnt graph of each method). We divide the missed edges into three groups (weak, intermediate and strong) based on their true L_2 norm. We see again that the three norms perform comparably, significantly out-performing the GMRC method in all three groups.

Finally, Fig. 2.3-B shows the sensitivity of our structure learning algorithms to the size of training set. In particular, we see that for the simulated graphs around 400 sequences results in us learning very accurate structures. However, as few as 50 sequences are enough to infer reasonable structures.

2.3.2 Evaluating Structure and Parameters Jointly

In a simulated setting, structure and parameter estimates can be compared against known ground truth. However, for real domain families we need other evaluation methods. We evaluate the structure and parameters for real domain families by measuring the imputation error of the learnt models. Informally, the imputation error measures the probability of *not* being able to “generate” a complete sequence, given an incomplete one. The imputation error of a column is measured by erasing it in the test MSA, and then computing the probability that the true (known) residues would be predicted by the learnt model. This probability is calculated by performing inference

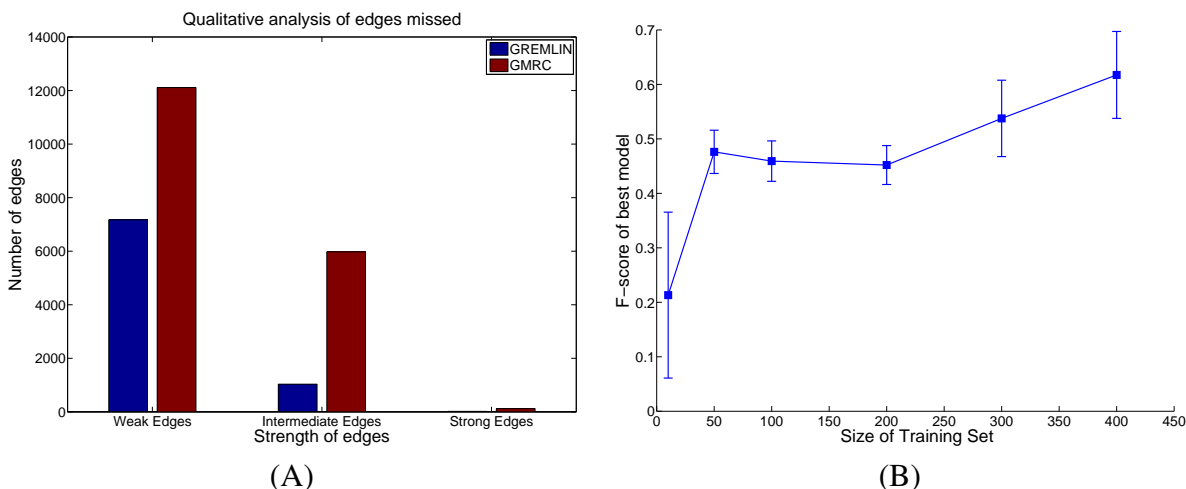


Figure 2.3: (A) Qualitative grouping of edges missed by GREMLIN and the GMRC method (B) Sensitivity of structure learning to size of training set.

on the erased columns, conditioned on the rest of the MSA. The imputation error of a model is the average of its imputation error over columns.

Using imputation error directly for model selection generally gives us models that are too dense. Intuitively, once we have identified the true model, adding extra edges decreases the imputation error by a very small amount, probably a reflection of the finite-sample bias. We evaluated the modified AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) for model selection due to their theoretically appealing properties. In the finite sample case we find that BIC performs well when the true graph is sparse, while AIC performs well when the true graph is dense. We discuss the information criteria in detail in the supplemental material, and provide some general suggestions for their use. Unfortunately, neither method performs well over the entire range of graphs. For this reason, we considered an approach to model selection based on finite sample error control. We chose to control the false discovery rate (FDR) in the following way. Consider permuting the each column of the MSA independently (and randomly). Intuitively, the true graph is now a graph with no edges. Thus, one approach to selecting the regularization parameter is to find the value that yields no edges on the permuted MSA. A more robust method, which we use, is to use the average regularization parameter obtained from multiple random permutations as in the work of Listgarten and Heckerman [128]. In the results that follow we use 20 random permutations.

Given the success of GREMLIN on simulated data, and equipped with a method for model selection described above, we proceed to apply GREMLIN to real protein MSAs. We consider the WW and PDZ families in some detail since the extensive literature on these families allows us to draw meaningful conclusions about the learnt models.

2.3.3 A generative model for the WW domain

The WW domain family (Pfam id: PF00397 [74]) is a small protein interaction module with two highly conserved tryptophans that adopts a curved three-stranded β -sheet structure with a binding site for proline-containing peptides. In the papers [174] and [165], the authors determine, using Statistical Coupling Analysis (SCA), that the residues can be divided into two clusters: the first cluster contains a set of 8 strongly coupled residues and the second cluster contains everything else. Based on this finding, the authors then designed 44 sequences that satisfy co-evolution constraints of the first cluster, of which 12 actually fold *in vitro*. An alternative set of control sequences, which did not satisfy the constraints, failed to fold.

We first constructed an MSA by starting with the PFAM alignment and removing sequences to construct a non-redundant alignment (no pair of sequences was greater than 80% similar). This resulted in an MSA with 700 sequences of which two thirds were used as a training set and the rest were used as a test set. Each sequence in the alignment had 30 positions. The training set was used to learn the model, for multiple values of λ_e . Given the structure of the graph, parameters were learned using $\lambda_v = 1, \lambda_e = 1$. The learnt model is presented in Figure 2.4.

Figure 2.5 compares the imputation errors of our approach (in red and yellow) with the GMRC method of Thomas et al. [186] and Profile HMMs of Eddy [67]. The model in red was learnt using λ_e selected by performing a permutation study. Since this model had more edges than the model learnt by GMRC, we used a higher λ_e to learn a model that had fewer edges than the GMRC model. The x-intercept was based on a loose lower bound on the error and was estimated by computing the imputation error on the test-data of a completely connected model *learnt on the test data*. Due to over-fitting, this is likely to be a very loose estimate of the lower bound. We find that our imputation errors are lower than the methods we compare to (even at comparable levels of sparsity).

To see which residues are affected by these edges, we construct a “coupling profile” (Fig. 2.4-C). We construct a shuffled MSA by taking the natural MSA and randomly permuting the amino acids within the same position (column of MSA) for each position. The new MSA now contains no co-evolving residues but has the same conservation profile as the original MSA. To build a coupling profile, we calculate the difference in the imputation error of sequences in a held-out test set and the shuffled MSA. Intuitively, having a high imputation error difference means that the position was indeed co-evolving with some other positions in the MSA. The other positions would also have a high imputation error difference in the coupling profile.

We also performed a retrospective analysis of the artificial sequences designed by Russ et al. [165]. We attempt to distinguish sequences that folded from those that didn't. Although this is a discriminative test (folded or not) of a generative model, we nevertheless achieve a high AUC of 0.87 (the ROC curve is shown and described in the supplemental material). We therefore postulate that the additional constraints we identify are indeed critical to the stability of the WW fold. In comparing our AUC to the published results of Thomas et al. [186] (AUC of 0.82) and the Profile HMM (AUC of 0.83) we see that we are able to better distinguish artificial sequences that fold from those that don't.

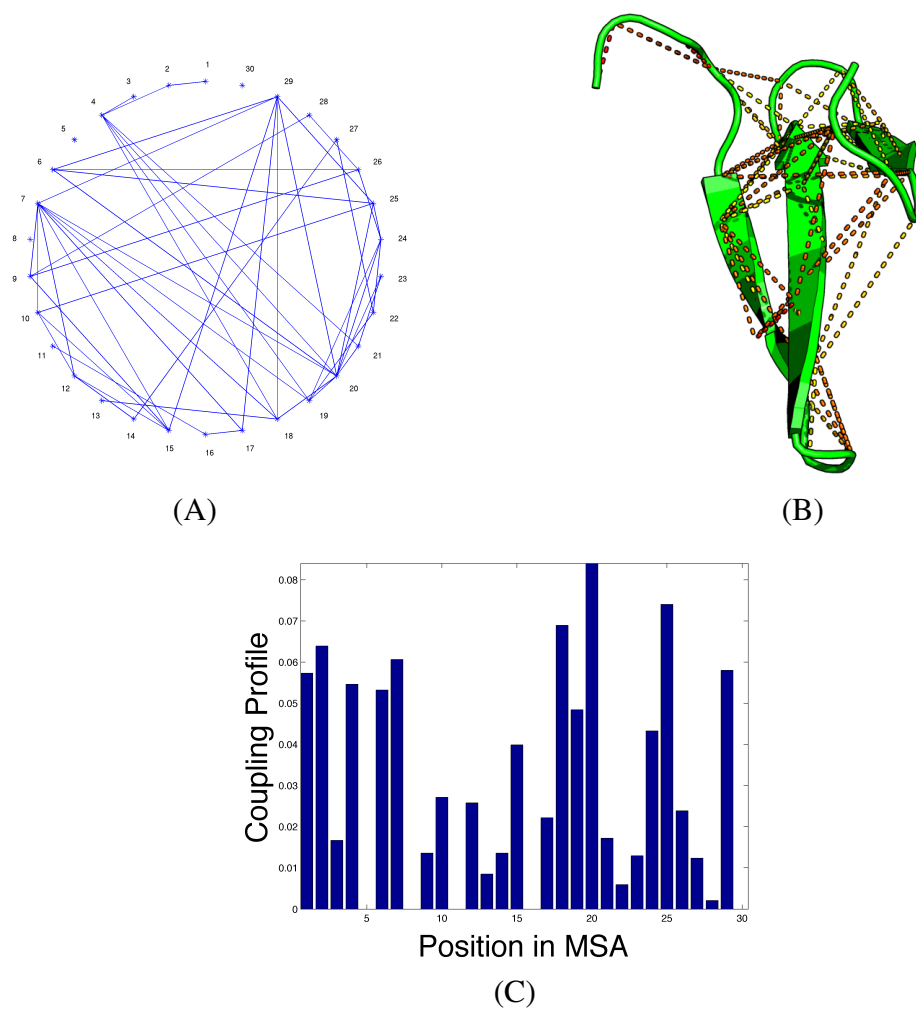


Figure 2.4: WW domain model. Edges returned by GREMLIN overlaid on a circle (a) and on the structure (b) of the WW domain of Transcription Elongation Factor 1 (PDB id: 2DK7) [27]. (c) Coupling profile (see text).

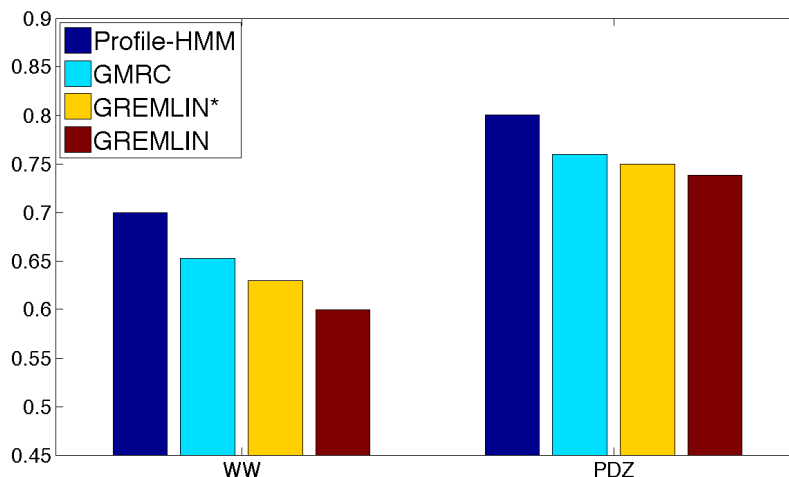


Figure 2.5: Comparison of Imputation errors on WW and PDZ families. We consider two variants of GREMLIN - with the regularization parameter selected either to produce a model with a smaller number of edges than GMRC (third bar in each group, shown in yellow) or to have zero edges on 20 permuted MSAs (last bar, shown in red). The x-intercept was chosen by estimating a lower bound on the imputation error as described in the text.

2.3.4 Allosteric regulation in the PDZ domain

The PDZ domain is a family of small, evolutionarily well represented protein binding motifs. The domain is most commonly found in signaling proteins and helps to anchor trans-membrane proteins to the cytoskeleton and hold together signaling complexes. The PDZ domain is also interesting because it is considered an *allosteric* protein. The domain, and its members have been studied extensively, in multiple studies, using a wide range of techniques ranging from computational approaches based on statistical coupling [133] and Molecular Dynamics simulations [63], to NMR based experimental studies [79].

We use the MSA from Lockless and Ranganathan [133]. The MSA is an alignment of 240 non-redundant sequences, with 92 positions. We chose a random sub-sample with two-thirds of the sequences as the training set and use the rest as a test set. Using this training set, we learnt generative models for each of the block regularizers, and choosing the smallest value of λ_e that gave zero edges for 20 permuted MSAs as explained previously. The resulting model had 112 edges (Fig. 2.6). Figure 2.5 summarizes the imputation errors on the PDZ domain. We again observe that the model we learn is denser than that learnt by GMRC and has lower imputation error. However, even at comparable sparsity GREMLIN out-performs the Profile HMM and GMRC.

The SCA based approach of Lockless and Ranganathan [133] identified a set of residues that were coupled to a residue near the active site (HIS-70) including a residue at a distal site on the other end of the protein (GLY-49 in this case). Since the SCA approach can only deter-

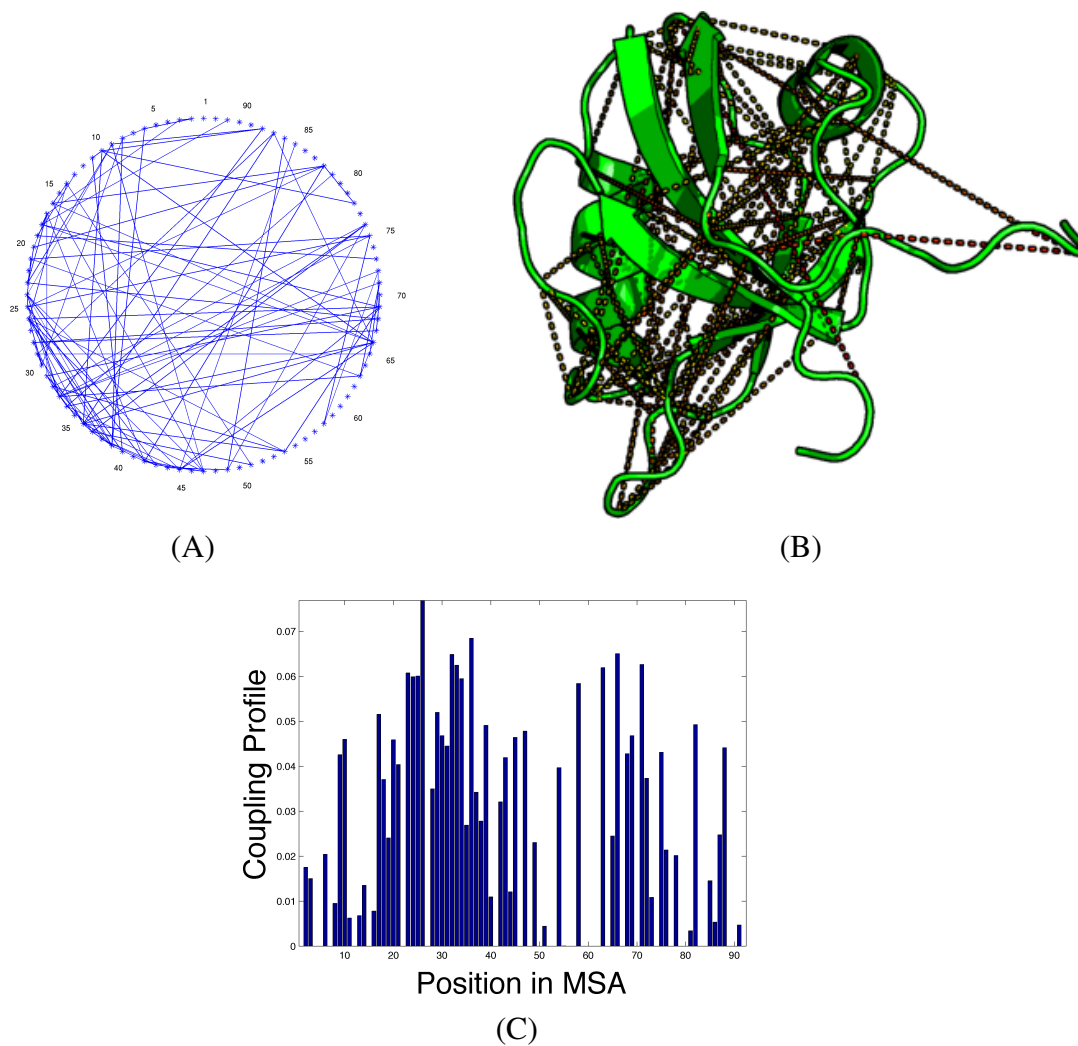


Figure 2.6: PDZ domain model. Edges returned by GREMLIN overlaid on a circle (a) and on the structure (b) of PDZ domain of PSD-95 (PDB id:1BE9). (c) Coupling profile (see text).

mine the presence of a dependence but cannot distinguish between direct and indirect couplings, only a cluster of residues was identified. Our model also identifies this interaction, but more importantly, it determines that this interaction is mediated by ALA-74 with position 74 *directly* interacting with both these positions. By providing such a list of sparse interactions our model can provide a small list of hypotheses to an experimentalist looking for possible mechanisms of such allosteric behavior.

In addition to the pathway between HIS-70 and GLY-49, we also identify residues not on the pathway that are connected to other parts of the protein including, for example ASN-61 of the protein. This position is connected to ALA-88 and VAL-60 in our model, and does not appear in the network suggested by Lockless and Ranganathan [133], but has been implicated by the NMR experiments of Fuentes et al. [79] as being dynamically linked to the active site.

From our studies on the PDZ and WW families we find that GREMLIN produces higher quality models than GMRC and profile HMMs, and identifies richer sets of interactions. In the following section we consider the application of GREMLIN to a larger subset of the PFAM database. Since the greedy algorithm of GMRC does not scale to large families, our experiments are restricted to comparing the performance of GREMLIN with that of profile HMMs.

2.3.5 Large-scale analysis of families from Pfam

We selected all protein families from PFAM [74] that had at least 300 sequences in their seed alignment. We restricted ourselves to such families because the seed alignments are manually curated before depositing and are therefore expected to have higher quality than the whole alignments. We pre-processed these alignments to remove redundant sequences (sequence similarity $> 80\%$) in order to generate non-redundant alignments. From each alignment, we then removed columns that had gaps in more than half the sequences, and then removed sequences in the alignment that had more than insertions at more than 10% of these columns. Finally, we removed sequences that had more than 20% gaps in their alignment. If this post-processing resulted in an alignment with less than 300 sequences, it was dropped from our analysis. 71 families remained at the end of this process. These families varied greatly in their length with the shortest family having 15 positions and the longest having more than 450 positions and the median length being 78 positions. Figure 2.7 shows the distribution of lengths.

For each of these families, we created a random partition of the alignment into training (with 2/3 of the sequences) and test (with 1/3 of the sequences) alignments and trained an MRF using our algorithm. As mentioned earlier, we chose λ_e by performing 20 random permutations of each column and choosing the smallest λ_e that gave zero edges on all 20 permutations. As a baseline comparison, we also trained a profile-HMM using the Bioinformatics toolkit in Matlab on the training alignments. We then used the learnt models to impute the composition of each position of the test MSA and computed the overall and per-position imputation errors for both models. Due to space constraints, we provide the models and detailed analyses for each family on a supporting website (<http://www.cs.cmu.edu/~cjl/gremlin/>) and focus on overall trends in the rest of this section.

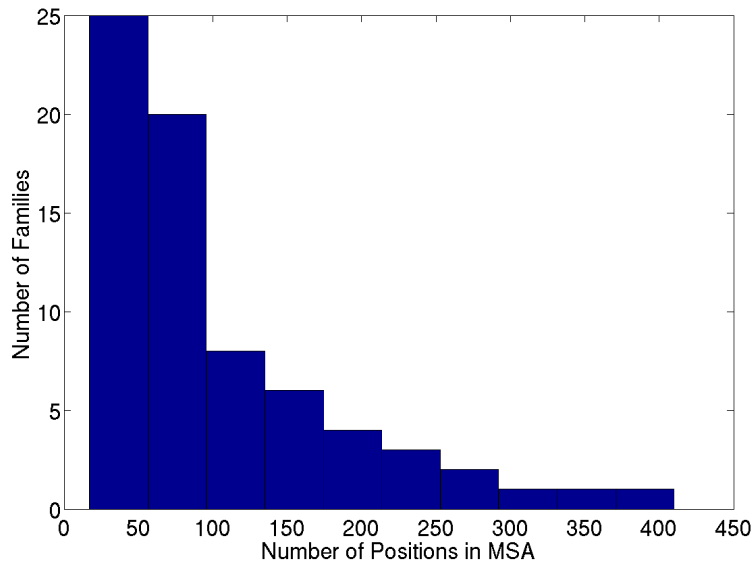


Figure 2.7: Histogram of MSA lengths of the 73 PFAM families in our study.

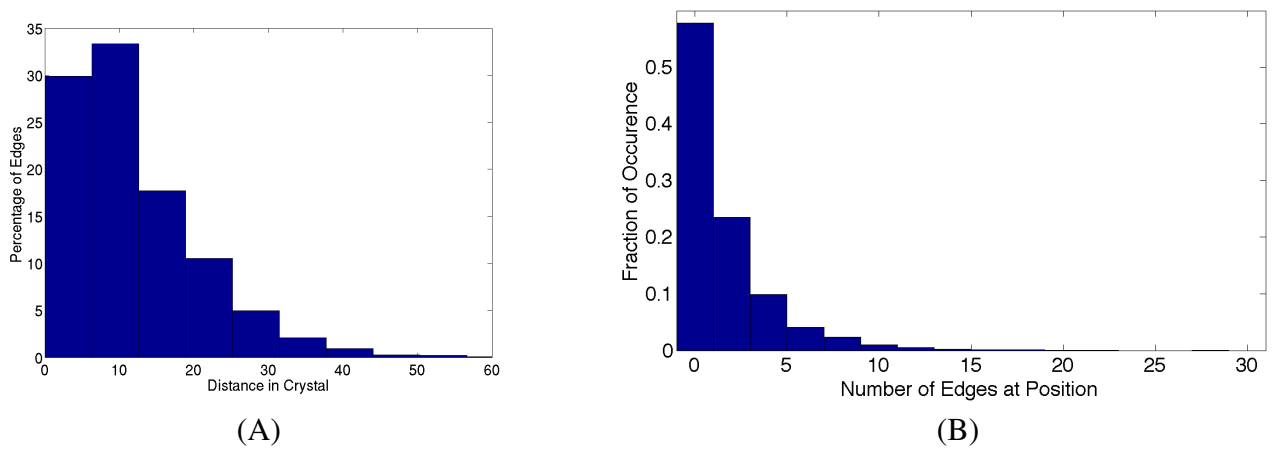


Figure 2.8: (A) Histogram of the distance in crystal structure. (B) Degree distribution across all proteins.

Figure 2.8 shows the histograms of the distance between residues connected by an edge and the degree of the nodes. Approximately 30% of the edges are between residues that are more than 10 Å of each other. That is, GREMLIN learns edges that are different than those that would be obtained from a contact map. Despite the presence of long-range edges, GREMLIN does learn a sparse graph; most nodes have degree less than 5, and the majority have 1 or fewer edges.

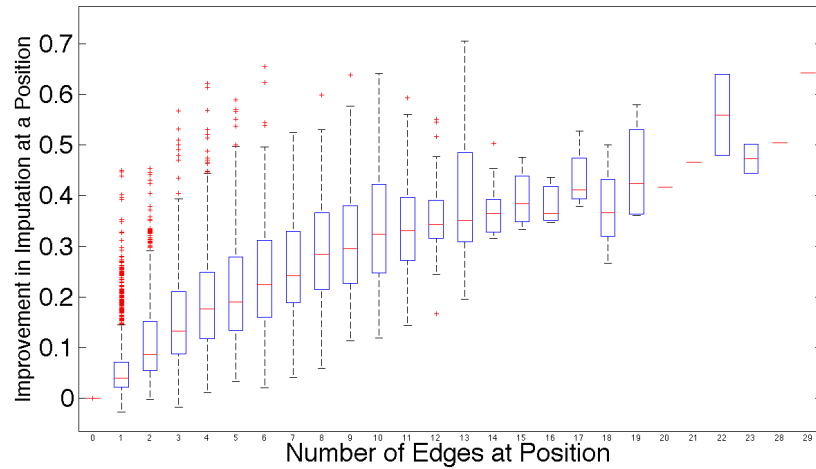
Fig. 2.9-(A) shows a boxplot demonstrating the effect of incorporating co-evolution information according to our model. The y-axis shows the decrease in the per-position imputation error when moving from a profile-HMM model to the corresponding MRF, while the x-axis bins this improvement according to the number of edges in the MRF at that position. In each box, the central red line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually with red '+' marks. As the figure shows, moving from a profile-HMM model to an MRF never hurts: for positions with 0 edges, there is no difference in imputation; for positions with at least one edge, the MRF model *always* results in lower error. While this is not completely surprising given that the MRF has more parameters and is therefore more expressive, it is not obvious that these parameters can be learnt from such little data. Our results demonstrate that this is indeed possible. While there are individual variations within each box, the median improvement in imputation error shows a clear linear relationship to the number of neighbors of the position in the model. This linear effect falls off towards the right in the high-degree vertices where the relationship is sub-linear. Fig. 2.9-(B) shows the effect of this behavior on the improvement in overall imputation error across all positions for a family.

2.3.6 Computational efficiency

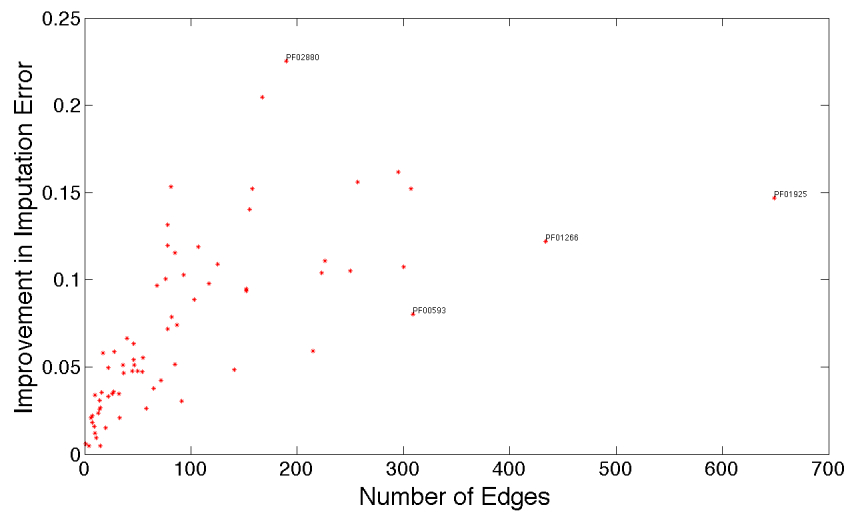
In this subsection we briefly discuss the computational efficiency of GREMLIN. The efficiency of GREMLIN was measured based on the running time (i.e. CPU seconds until a solution to the convex optimization problem is found). GREMLIN was run on a 64 node cluster. Each node had 16GB DRAM and 2xquad-cores (each with 2.8-3 GHZ), allowing us to run 512 jobs in parallel with an average of 2GB RAM per job.

Fig. 2.10 shows a plot of the running time for a given λ_e on all the PFAM MSAs. Fig. 2.10-(A) plots the running time for learning the neighbors of a position, against the number of columns (positions) in the MSA (A) while 2.10-(B) plots it against number of rows (sequences) in the training MSA. In both, the average running time *per column* is shown in red circles. While learning the neighbors at a position, since GREMLIN is run in parallel for each column of the MSA, the actual time to completion for each protein depends on the maximum running time across these columns. This number is shown in blue squares. Fig. 2.10-(C) plots the running time for parameter learning against the maximum running time to learn the neighbors at a position. Recall that this task is performed serially. As the figure demonstrates, GREMLIN takes roughly similar amounts of time in its parallel stage (neighborhood learning) as it does in its serial stage (parameter learning).

The plots show that the running time has an increasing trend as the size of the MSA increases

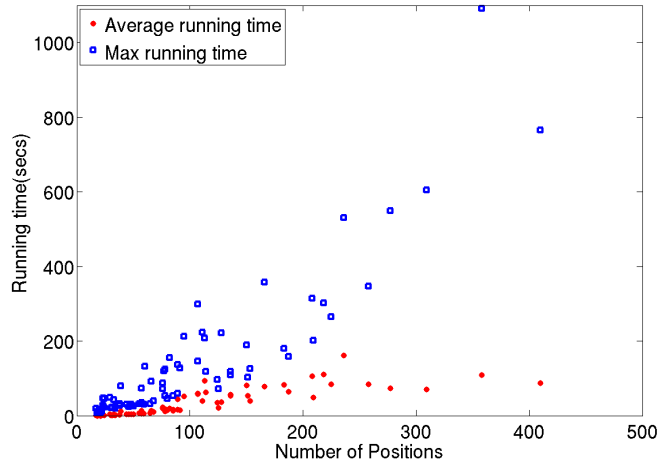


(A)

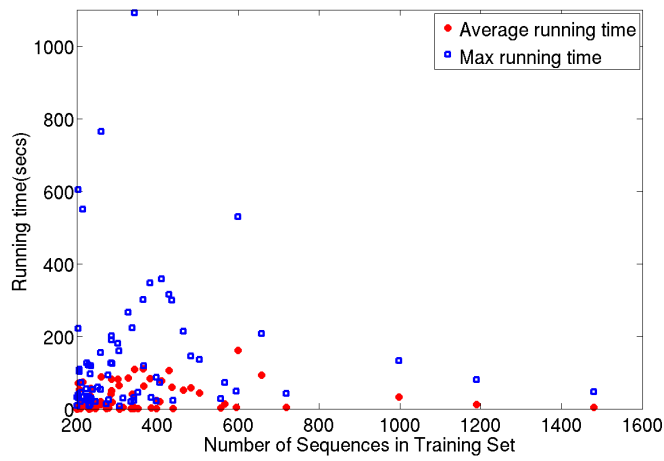


(B)

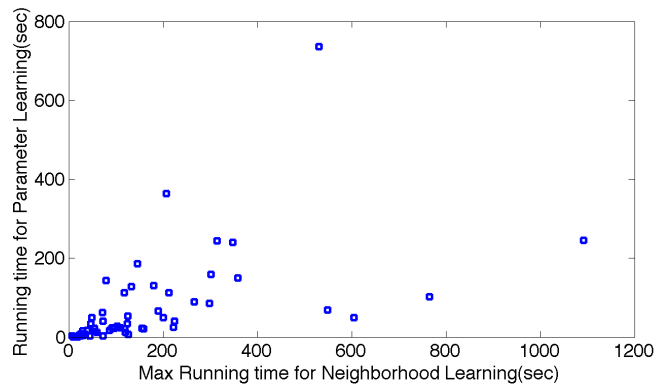
Figure 2.9: (A) Boxplot displaying the effect of coupling on improvement in imputation error at a position when compared to a profile-HMM. The median imputation error shows a near-linear decrease as the number of neighbors learnt by the model increases. (B) Improvement in overall imputation error across all positions for each family.



(A)



(B)



(C)

Figure 2.10: (A) Number of Positions in the MSA versus runtime of Neighborhood learning (in seconds) (B) Number of sequences in the MSA versus runtime of Neighborhood learning (C) Runtime of Neighborhood learning versus runtime of Parameter learning

(number of positions and number of sequences). Also, the dependence of the running time on the number of columns is stronger than its dependence on the number of rows. This is consistent with the analysis in the paper of Wainwright et al. [199] which shows that a similar algorithm for structure learning with a pure L_1 penalty has a computational complexity that scales as $\mathcal{O}(\max(n, p)p^3)$, where n corresponds to the number of rows and p to the number of columns in the MSA.

2.4 Discussion

2.4.1 Related Work

The study of co-evolving residues in proteins has been a problem of much interest due to its wide utility. Much of the early work focused on detecting such pairs in order to predict contacts in a protein in the absence of a solved structure (see the papers [5, 84]) and to perform fold recognition. The pioneering work of Lockless and Ranganathan [133] used an approach to determine probabilistic dependencies that they called SCA and observed that analyzing such patterns could provide insights into the allosteric behavior of the proteins and be used to design new sequences [174]. Others, [72, 76, 78], have since developed similar methods. By focusing on co-variation or probabilistic *dependencies* between residues, such methods conflate direct and indirect influences and can lead to incorrect estimates. In contrast, Thomas et al. [186] developed an algorithm for learning a Markov Random Field over sequences. Their constraint-based algorithm proceeds by identifying conditional independencies and adding edges in a greedy fashion. However, the algorithm can provide no guarantees on the correctness of the networks it learns. They then extended this approach to incorporate interaction data to learn models over pairs of interacting proteins [187] and also develop a sampling algorithm for protein design using such models [189]. More recently, Weigt et al. [202] use a similar approach to determine residue contacts at a protein-protein interface. Their method uses a gradient descent approach using Loopy Belief Propagation to approximate likelihoods. Additionally, their algorithm does not regularize the model and may therefore be prone to over-fitting. In contrast, we use a Pseudo-Likelihood as our objective function thereby avoiding problems of convergence that Loopy BP based methods can face and regularize the model using block regularization to prevent over-fitting.

Block regularization is most similar in spirit to the group LASSO [210] and the multi-task LASSO [7]. LASSO [190] is the problem of finding a linear predictor, by minimizing the squared loss of the predictor with an L_1 penalty. It is well known that the shrinkage properties of the L_1 penalty lead to sparse predictors. The group LASSO extends this idea by grouping the weights of some features of the predictor using an L_2 norm, Yuan and Lin [210] show that this leads to sparse selection of groups. The multi-task LASSO solves the problem of multiple separate (but similar) regression problems by grouping the weight of a single feature across the multiple tasks. Intuitively, we solve a problem similar to a group LASSO, replacing the squared loss with an approximation to the negative log-likelihood, where we group all the feature weights of an edge in an undirected graphical model. Thus, sparse selection of groups gives our graphs the property

of structural sparsity.

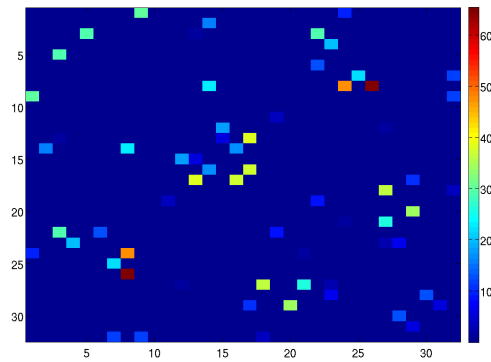
Lee et al. [124] introduced structure learning in MRFs with a pure L_1 penalty, but do not go further to explore block regularization. They also use a different approximation to the likelihood term, using Loopy Belief Propagation. Schmidt et al. [167] apply block-regularized structure learning to the problem of detecting abnormalities in heart motion. They also developed an efficient algorithm for tractably solving the convex structure learning problem based on projected gradients.

2.4.2 Mutual Information performs poorly in the structure learning task

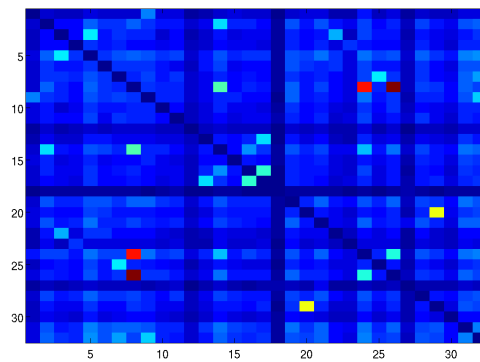
One of the key advantages of a graphical model based approach to modeling protein families is that the graph reveals which interactions are direct and which are indirect. One might assume that alternative quantities, like Mutual Information, might yield similar results. We now demonstrate with an example that a simple Mutual Information based metric cannot distinguish well between direct and indirect interactions. Fig. 2.11-(A) shows the adjacency matrix of a Probabilistic Graphical Model. The elements of the matrix are color-coded by the strength of their interaction: blue represents the weakest interaction (of strength 0, i.e. a non-interaction) and red the strongest interaction in this distribution. Fig. 2.11-(B) shows the mutual information induced between the variables by this distribution as measured from 500 sequences sampled from the graphical model (the diagonal elements of the mutual information matrix have been omitted to highlight the information between different positions). While it may appear visually that (B) shares a lot of structure with (A), it isn't actually the case. In particular, the edges with the highest mutual information indeed tend to be direct interactions; however a large fraction of the direct interactions might not have high MI. This is demonstrated in Fig. 2.11-(C) where MI is used as a metric to classify edges into direct and indirect interactions. The blue line shows the ROC curve using MI as a metric and has only moderate discriminatory power for this task (AUC: 0.71). In contrast, our approach, shown in red, is much more successful at discriminating between direct and indirect interactions: the AUC of our approach is a near-perfect 0.98.

2.4.3 Influence of Phylogeny

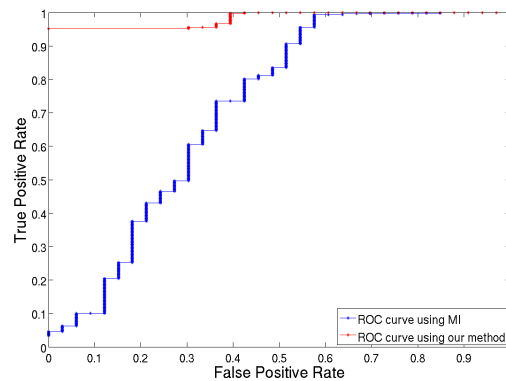
One limitation associated with a sequence-only approach to learning a statistical model for a domain family is that the correlations observed in the MSA can be inflated due to phylogeny [73, 154]. A pair of co-incident mutations at the root of the tree can appear as a significant dependency even though they correspond to just once co-incident mutation event. To test if this was the case with the WW domain, we constructed a phylogenetic tree from the MSA using Junes-Cantor measure of sequence dissimilarity. In the case of WW, this resulted in a tree with two clear sub-trees, corresponding to two distinct (nearly equal-sized) clusters in sequence space. Since each sub-tree had a number of sequences, we re-learned MRFs for each sub-tree separately. The resulting models for each sub-tree did not vary significantly from our original models – a case that would have occurred if there were co-incident mutations at the root that lead to



(A)



(B)



(C)

Figure 2.11: (A) Adjacency matrix of a Boltzmann distribution colored by edge strength. (B) Mutual Information between positions induced by this Boltzman distribution. While the mutual information of the strongest edges is highest; a large fraction of the edges have MI comparable to many non-interactions. (C) Shows the weak ability of MI to distinguish between edges and indirect interactions in contrast to GREMLIN . AUC using MI: 0.71; AUC using GREMLIN : 0.98.

spurious dependencies. Indeed the only difference between the models was in the C-terminal end was an edge between positions 1 and 2 that was present in sequences from the first sub-tree but was absent in the second sub-tree. This occurred because in the second sub-tree, these positions were completely conserved due to which our model was not able to determine the dependency between them. While this does not eliminate the possibility of confounding due to phylogeny, we have reason to believe that our dependencies are robust to significant phylogenetic confounding in this family. A similar analysis for the PDZ domain, found 3 sub-trees, and again we found that the strongest dependencies were consistent across models learnt on each sub-tree separately. Nevertheless, we believe that incorporating phylogenetic information into our method is an important direction for future research.

2.5 Conclusions

In this chapter we have proposed a new algorithm for discovering and modeling the statistical patterns contained in a given MSA. Overall, we find that by employing sound probabilistic modeling and convex structure (and parameter) learning, we are able to find a good balance between structural sparsity (simplicity) and goodness of fit. One of the key advantages of a graphical model approach is that the graph reveals the direct and indirect constraints that can further our understanding of protein function and regulation.

MRFs are generative models, and can therefore be used design new protein sequences via sampling and inference. However, we expect that the utility of our model in the context of protein design could be greatly enhanced by incorporating structure based information which explicitly models the physical constraints of the protein.

Finally, we note that there are a number of other ways to incorporate phylogenetic information directly into our model. For example, given a phylogenetic clustering of sequences, we can incorporate a single additional node in the graphical model reflecting the cluster to which the sequence belongs. This would allow us to distinguish functional coupling from coupling caused due to phylogenetic variations.

2.6 Additional experiments

2.6.1 Comparison of structures learnt at different regularization levels

Fig. 2.12 shows our performance in predicting the true structure by using L_1-L_2 (Fig. 2.12) The accuracy is measured using the F-score (the harmonic mean of precision and recall) of the edge set. We observe that for all settings of ρ GREMLIN learns fairly accurate graphs at some value of λ_e .

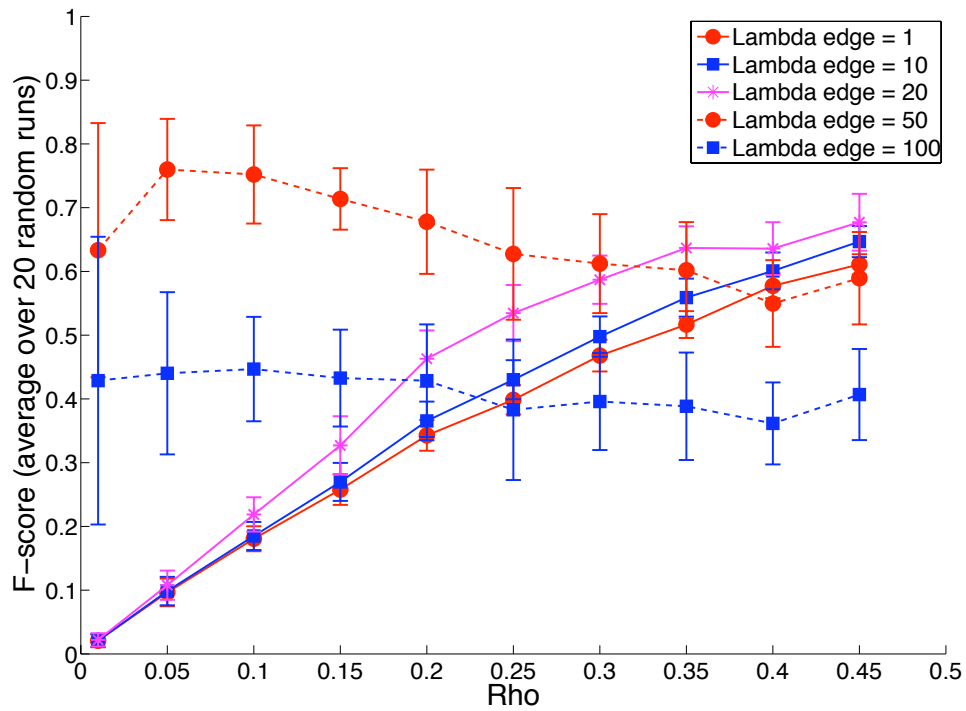


Figure 2.12: F-scores of structures learnt by using L_1 - L_2 norm. The figure shows the average and standard deviation of the F-score across 20 different graphs as a function of ρ , the probability of edge-occurrence.

Model selection using information criteria

We consider modifications to two widely used model selection strategies. The Bayesian Information Criterion (BIC) [168], is used to select parsimonious models and is known to be asymptotically consistent in selecting the true model. The Akaike Information Criterion (AIC) [4], typically selects denser models than the BIC, but is known to be asymptotically consistent in selecting the model with lowest predictive error (risk). In general, they do not however select the same model [207].

We use the following definitions:

$$\begin{aligned}\text{pseudo-BIC}(\lambda) &= -2\text{pll}(\lambda) + \log(n)\text{df}(\lambda) \\ \text{pseudo-AIC}(\lambda) &= -2\text{pll}(\lambda) + 2\text{df}(\lambda)\end{aligned}$$

Where we use the pseudo log-likelihood approximation to the log-likelihood. While it may be expected that using the pseudo log-likelihood instead of the true log-likelihood may in fact lead to inconsistent selection a somewhat surprising result of Csiszar and Talata [51] shows that in the case of BIC using pseudo log-likelihood is in fact also consistent for model selection. Although we aren't aware of the result, we expect a similar result to hold for the risk consistency of the pseudo-AIC.

We evaluate the likelihood on the *training* sample to score the different models. n is the number of training sequences.

Estimating the degrees of freedom of a general estimator is quite hard in practice. This has lead to use of various heuristics in practice. For the LASSO estimator which uses a pure-L1 penalty, it is known that the number of non-zeros in the regression vector is a good estimate of the degrees of freedom. A natural extension when using a *block*-L1 penalty is the number of non-zero blocks (i.e. edges). Since this does not differentiate between weak and strong edges, we used the block-L1 norm as an estimate of the degress of freedom. In our simulations, we find that choice often results in good model selection.

Figure 2.13 shows the performance of the two model selection strategies at different sparsity levels. We evaluate the performance by learning several graphs (at different levels of regularization) and comparing the Spearman rank-correlation between the F-score of the graphs and their rank. We can clearly see that when the true graph is sparse the modified BIC has a high rank-correlation, whereas when the true graph is dense the modified AIC does well, with neither method providing reliable model selection for all graphs.

2.6.2 Receiver operating characteristic curve

We consider the task of distinguishing artificial sequences that were found to take the WW fold from those that did not. All sequences and their labels (folded in vivo or not) are from the paper of Russ et al. [165]. The ROC curve (Figure 2.14) is obtained by varying a threshold on scores

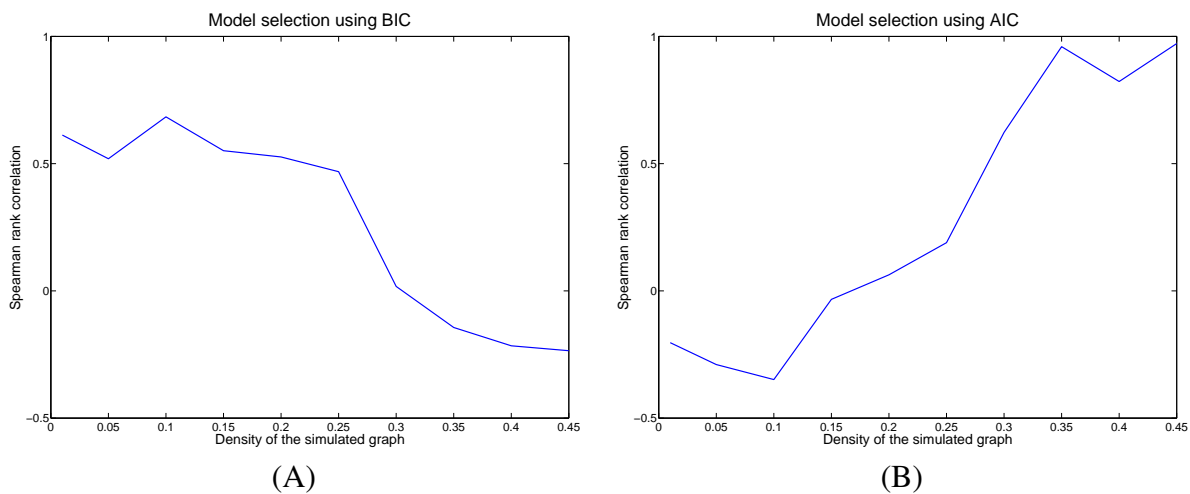


Figure 2.13: Graph density versus the rank correlation for ranking and selection using (A) BIC (B) AIC.

(we use the unnormalized likelihood as the score). Sequences above the threshold are predicted to fold. For each threshold we calculate the sensitivity and specificity and show the resulting curve.

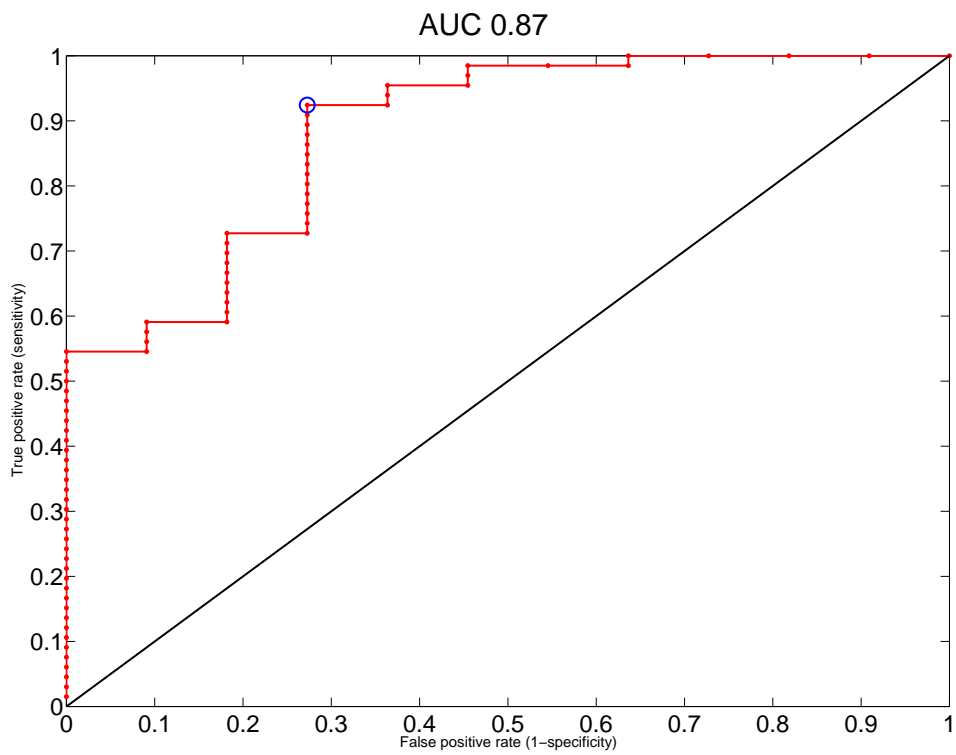


Figure 2.14: Receiver operating characteristic (ROC) curve of GREMLIN for the task of distinguishing artificial WW sequences that fold from those that don't.

Chapter 3

Sparse Additive Kernel and Functional CCA

Canonical Correlations Analysis (CCA) is a classical tool for finding correlations among the components of two random vectors. In recent years, CCA has been widely applied to the analysis of genomic data, where it is common for researchers to perform multiple assays on a single set of patient samples.

Recent work of Witten et al. [205], Witten and Tibshirani [206] has proposed sparse variants of CCA to address the high dimensionality of such data. However, classical and sparse CCA are based on linear models, and are thus limited in their ability to find general correlations. In this chapter, we present two approaches to high-dimensional nonparametric CCA, building on recent developments in high-dimensional nonparametric regression. We present estimation procedures for both approaches, and analyze their theoretical properties in the high-dimensional setting. We demonstrate the effectiveness of these procedures in discovering nonlinear correlations via extensive simulations, as well as through experiments with genomic data.

3.1 Introduction

Canonical correlation analysis [97], is a classical method for finding correlations between the components of two random vectors $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$. Given a set of n paired observations $(X_1, Y_1), \dots, (X_n, Y_n)$, we form the design matrices $\mathbb{X} \in \mathbb{R}^{n \times p_1}$ and $\mathbb{Y} \in \mathbb{R}^{n \times p_2}$ and find vectors $u \in \mathbb{R}^{p_1}$ and $v \in \mathbb{R}^{p_2}$ that are solutions to the optimization

$$\begin{aligned} & \arg \max_{u,v} \frac{1}{n} u^T \mathbb{X}^T \mathbb{Y} v \\ \text{s.t. } & \frac{1}{n} u^T \mathbb{X}^T \mathbb{X} u \leq 1 \quad \frac{1}{n} v^T \mathbb{Y}^T \mathbb{Y} v \leq 1, \end{aligned} \tag{3.1}$$

where the columns of \mathbb{X} and \mathbb{Y} have been standardized to have mean zero and standard deviation one. This is the sample version of the problem of maximizing the correlation between the linear combinations $u^T X$ and $v^T Y$, assuming the random variables have mean zero.

CCA can serve as a valuable dimension reduction tool, allowing one to quickly zoom in on interesting phenomena shared by multiple data sets. This tool is increasingly attractive in genomic data analysis, where researchers perform multiple assays per item. For instance, data including DNA copy number (or comparative genomic hybridization, CGH), gene expression, and single nucleotide polymorphism (SNP) information can be collected on a common set of patients. Witten et al. [205] present examples of recent studies involving such data.

When the data are high dimensional, as is often the case for genomic data, the classical formulation of CCA is not meaningful, since the sample covariance matrices $\mathbb{X}^T \mathbb{X}$ and $\mathbb{Y}^T \mathbb{Y}$ are singular. This has motivated different approaches to *sparse* CCA, which regularizes Eq. 3.1 by suitable sparsity-inducing ℓ_1 penalties [49, 149, 205, 206]. Sparsity can lead to more interpretable models, reduced computational cost, and favorable statistical properties for high dimensional data. Existing methods for CCA are, however, restricted in that they attempt to find linear combinations of the variables—interesting correlations need not be linear. The need for this flexibility motivates the nonparametric approaches we consider in this chapter.

The general nonparametric analogue of Eq. 3.1 is

$$\begin{aligned} \arg \max_{f,g} \quad & \frac{1}{n} \sum_{i=1}^n f(X_i)g(Y_i) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n f^2(X_i) \leq 1 \quad \frac{1}{n} \sum_{i=1}^n g^2(Y_i) \leq 1, \end{aligned} \tag{3.2}$$

where f and g are restricted to belong to an appropriate class of smooth functions. Bach and Jordan [14] introduce a version of this called kernel CCA by applying the “kernel trick” to the CCA problem. Kernel CCA allows flexible nonparametric modeling of correlations, solving Eq. 3.2 with additional regularization to enforce smoothness of the functions f and g in appropriate reproducing kernel Hilbert spaces. However, this general nonparametric model suffers from the curse of dimensionality, as the number of samples required for consistency grows exponentially with the dimension. It is thus necessary to further restrict the complexity of possible functions. We consider the class of additive models where f and g can be written as

$$f(x_1, x_2, \dots, x_{p_1}) = \sum_{j=1}^{p_1} f_j(x_j) \tag{3.3}$$

$$g(y_1, y_2, \dots, y_{p_2}) = \sum_{k=1}^{p_2} g_k(y_k), \tag{3.4}$$

in terms of univariate component functions [91]. In the regression setting, such models no longer require the sample size to be exponential in the dimension; however, they only have strong statistical properties in low dimensions. Recently, several authors have shown how sparse additive models for regression can be efficiently estimated even when $p > n$ [115, 139, 156, 157].

In this chapter we propose two additive nonparametric formulations of CCA, one over a family of RKHSs and another over Sobolev spaces without a reproducing kernel. In the low-dimensional setting where we do not enforce sparsity, the formulation over Sobolev spaces is closely related to the Alternating Conditional Expectations (ACE) formulation of nonparametric regression due to Breiman and Friedman [34]. In addition to formulating algorithms for the optimizations, we provide risk consistency guarantees for the global risk minimizer in the high dimensional regime where $\min(p_1, p_2) > n$.

An important consideration is that sparse nonparametric CCA is biconvex, but not jointly convex in f and g . This is true even for the linear CCA model, which is a special case of the model we propose. In the absence of the sparsity constraints the linear problem reduces to a generalized eigenvalue problem which can be efficiently solved. This remains true in the nonparametric case as well. Over an RKHS, the problem without sparsity is a generalized eigenvalue problem where Gram matrices replace the data covariance matrices. In the population setting over the Sobolev spaces we consider, Breiman and Friedman [34] show that the problem reduces to an eigenvalue problem with respect to conditional expectation operators.

Returning to the nonconvex sparse CCA problem, Witten et al. [205] and Parkhomenko et al. [149] suggest using the solution to the nonsparse version of the problem to initialize sparse CCA; Chen and Liu [49] use several random initializations. As we show in simulations, both approaches can lead to poor results, even in the linear case. To address this issue, we propose and study a simple marginal thresholding step to reduce the dimensionality, in the spirit of the diagonal thresholding of Johnstone and Lu [104] and the SURE screening of Fan and Song [71]. This results in a three step procedure where after preprocessing we use the nonsparse version of our problem to determine a good initialization for the sparse formulation.

In Sections 3.2 and 3.3 we briefly describe the additive Sobolev and RKHS function spaces over which we work, introduce our two nonparametric CCA formulations, and discuss their optimization. In Section 3.4 we address the non-convexity of the formulations and initialization strategies. In Section 3.5 we summarize the theoretical guarantees of these procedures when $p_1, p_2 > n$ and in Section 3.6 we describe some simulations and real data experiments.

3.2 Sparse additive kernel CCA

Recall the linear CCA problem Eq. 3.1. We will now derive its additive generalization over RKHSs. Let $\mathcal{F}_j \subset L_2(\mu(x_j))$ be a reproducing kernel Hilbert space of univariate functions on the domain of X_j , and let $\mathcal{G}_k \subset L_2(\mu(y_k))$ be a reproducing kernel Hilbert space of univariate functions on the domain Y_k , for each $j = 1, \dots, p_1$ and $k = 1, \dots, p_2$. We assume that $\mathbb{E}[f_j(X_j)] = 0$ and $\mathbb{E}[g_k(Y_k)] = 0$ for all $f_j \in \mathcal{F}_j$, and $g_k \in \mathcal{G}_k$ for each j and k . This is necessary to enforce model identifiability. In practice, we will always work with centered Gram matrices to enforce this (see [14]).

Denote by

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{p_1} f_j(x_j) \mid f_j \in \mathcal{F}_j \right\},$$

and

$$\mathcal{G} = \left\{ g = \sum_{k=1}^{p_2} g_k(y_k) \mid g_k \in \mathcal{G}_k \right\},$$

the sets of additive functions of x and y , respectively.

We are given n independent tuples of the form $(X_i, Y_i)_{i=1}^n$ where $X_i = \{X_{i1}, \dots, X_{ip_1}\}$ and $Y_i = \{Y_{i1}, \dots, Y_{ip_2}\}$, and positive definite kernel functions on each covariate of X and Y . We denote the Gram matrix for the j^{th} X covariate by K_{x_j} and for the k^{th} Y covariate by K_{y_k} .

We will need to regularize the CCA problem to enforce smoothness and sparsity of the functions. The two norms

$$\|f_j\|_{\mathcal{F}_j} = \sqrt{\langle f_j, f_j \rangle_{\mathcal{F}_j}} \quad \text{and} \quad \|f_j\|_2 = \sqrt{\frac{1}{n} \sum_{i=1}^n f_j^2(X_{ij})}$$

play an important role in our approach. We can now formulate the *sparse additive kernel CCA* (SA-KCCA) problem as

$$\begin{aligned} \max_{f \in \mathcal{F}, g \in \mathcal{G}} \quad & \frac{1}{n} \sum_{i=1}^n f(X_i)g(Y_i) \quad \text{subject to} \\ & \frac{1}{n} \sum_{i=1}^n f^2(X_i) + \gamma_f \sum_{j=1}^{p_1} \|f_j\|_{\mathcal{F}_j}^2 \leq 1 \quad \sum_{j=1}^{p_1} \|f_j\|_2 \leq C_f \\ & \frac{1}{n} \sum_{i=1}^n g^2(Y_i) + \gamma_g \sum_{k=1}^{p_2} \|g_k\|_{\mathcal{G}_k}^2 \leq 1 \quad \sum_{k=1}^{p_2} \|g_k\|_2 \leq C_g. \end{aligned} \tag{3.5}$$

for given regularization parameters γ_f, γ_g, C_f and C_g . As with the group LASSO, constraining $\sum_j \|f_j\|_2$ encourages sparsity amongst the functions f_j [157]. As stated, this is an infinite dimensional optimization problem over Hilbert spaces. However, a straightforward application of the representer theorem shows that it is equivalent to the following finite dimensional optimization

problem:

$$\begin{aligned}
& \max_{\alpha, \beta} \frac{1}{n} \left(\sum_{j=1}^{p_1} K_{xj} \alpha_j \right) \left(\sum_{k=1}^{p_2} K_{yk} \beta_k \right) \quad \text{subject to} \\
& \frac{1}{n} \left(\sum_{j=1}^{p_1} K_{xj} \alpha_j \right)^T \left(\sum_{j=1}^{p_1} K_{xj} \alpha_j \right) + \gamma_f \sum_{j=1}^{p_1} \alpha_j^T K_{xj} \alpha_j \leq 1 \\
& \frac{1}{n} \left(\sum_{k=1}^{p_2} K_{yk} \beta_k \right)^T \left(\sum_{k=1}^{p_2} K_{yk} \beta_k \right) + \gamma_g \sum_{k=1}^{p_2} \beta_k^T K_{yk} \beta_k \leq 1 \\
& \sum_{j=1}^{p_1} \sqrt{\frac{1}{n} \alpha_j^T K_{xj}^T K_{xj} \alpha_j} \leq C_f, \quad \sum_{k=1}^{p_2} \sqrt{\frac{1}{n} \beta_k^T K_{yk}^T K_{yk} \beta_k} \leq C_g.
\end{aligned} \tag{3.6}$$

Here α is an $(n \times p_1)$ matrix, α_j is its j^{th} column, β is an $(n \times p_2)$ matrix and β_k is its k^{th} column.

The problem Eq. 3.6 is not convex. However, if we fix the function g (or equivalently the coefficients β) the problem is convex in f (equivalently α), and vice-versa. This *biconvexity* leads to a natural optimization strategy for Eq. 3.6 which we describe below. However, this procedure only guarantees convergence to a local optimum and in practice we still need to be able to find a good initialization.

In the absence of the sparsity penalty the problem becomes an additive form of kernel CCA [14]. One could also consider alternative formulations that, for instance, separate the smoothness and variance constraints. One attractive feature of our formulation is that without the sparsity constraint the problem can be reduced to a generalized eigenvalue computation which can be solved optimally. This leads us to a strategy of biconvex optimization that mirrors the linear algorithm of Witten et al. [205]; specifically, initialize by solving the problem without the sparsity constraints, fix α and optimize for β and vice-versa until convergence. As our experiments will show this is indeed a good strategy when $p_1, p_2 < n$. However, new ideas, to be described in Section 3.4, are necessary to scale this to the high dimensional setting where $p_1, p_2 > n$.

3.3 Sparse additive functional CCA

We now formulate an optimization problem for sparse additive functional CCA (SA-FCCA), and derive a scalable backfitting procedure for this problem. Here we work directly over the Hilbert spaces $L_2(\mu(x))$ and $L_2(\mu(y))$. We will denote by \mathcal{S}_j the subspace of $\mu(x_j)$ measurable functions with mean 0, with the usual inner product $\langle f_j, f'_j \rangle = \mathbb{E} (f_j(X_j) f'_j(X_j))$, and similarly \mathcal{T}_k for the functions of y .

To enforce smoothness we consider functions lying in a ball in a second order Sobolev space. We further assume the functions are uniformly bounded, and the measures μ are supported on a

compact subset of a Euclidean space with Lebesgue measure λ . For a fixed uniformly bounded, orthonormal basis ψ_{jk} with respect to λ we have

$$\mathcal{F}_j = \left\{ f_j \in \mathcal{S}_j : f_j = \sum_{k=0}^{\infty} \beta_{jk} \psi_{jk}, \sum_{k=0}^{\infty} \beta_{jk}^2 k^4 \leq C^2 \right\}$$

and similarly for \mathcal{G}_k . We will call these the *smooth* functions, and denote by \mathcal{F} and \mathcal{G} the set of smooth additive functions over the respective Hilbert spaces.

Our formulation of *sparse additive functional CCA* is the optimization

$$\max_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n f(X_i)g(Y_i) \quad (3.7)$$

$$\begin{aligned} \text{s.t.} \quad & \frac{1}{n} \sum_{j=1}^{p_1} \sum_{i=1}^n f_j^2(X_{ij}) \leq 1, & \sum_{j=1}^{p_1} \|f_j\|_2 \leq C_f \\ & \frac{1}{n} \sum_{k=1}^{p_2} \sum_{i=1}^n g_k^2(Y_{ik}) \leq 1, & \sum_{k=1}^{p_2} \|g_k\|_2 \leq C_g, \end{aligned}$$

where the $\|\cdot\|_2$ norm is defined as in additive kernel CCA. This problem is superficially similar to Eq. 3.2; however, there are three important differences. First, we don't regularize for smoothness but instead work directly over a Sobolev space of smooth functions. Secondly, we do not constrain the variance of the function f . Instead, in the spirit of "diagonal penalized CCA" of Witten et al. [205] we constrain the sum of the variances of the individual f_j s. This choice is made primarily because it leads to backfitting updates that have a particularly simple and intuitive form. Perhaps most importantly, we can no longer appeal to the representer theorem since we are not working over RKHSs.

We study the population version of this problem to derive a biconvex backfitting procedure to directly optimize this criterion. The sample version of the algorithm is described in Algorithm 1, and a complete derivation is part of the supplementary material. To gain some intuition for this procedure we describe one special case of the population algorithm, where g is fixed and both constraints on f are tight. Consider the Lagrangian problem

$$\max_f \min_{\lambda \geq 0, \gamma \geq 0} \mathbb{E}[f(X)g(Y)] - \lambda(\|f\|_2^2 - 1) - \gamma(\|f\|_1 - C_f).$$

The norms are defined as $\|f\|_1 = \sum_{j=1}^{p_1} \sqrt{\mathbb{E}(f_j^2(x_j))}$ and $\|f\|_2^2 = \sum_{j=1}^{p_1} \mathbb{E}(f_j^2(x_j))$. For simplicity, consider the case when $\lambda, \gamma > 0$, and denote $a \equiv g(Y)$.

We now can derive a coordinate ascent style procedure where we optimize over f_j holding the other functions fixed. The Fréchet derivative w.r.t. f_j in the direction η gives one of the KKT conditions $\mathbb{E}[(a - 2\lambda f_j - \gamma \nu_j)\eta] = 0$ for all η in the Hilbert space \mathcal{H}_j , where the subdifferential is $\nu_j = \frac{f_j}{\sqrt{\mathbb{E}(f_j^2)}}$ if $\sqrt{\mathbb{E}(f_j^2)}$ is not 0, and is the set $\{u_j \in \mathcal{H}_j \mid \mathbb{E}(u_j^2) \leq 1\}$ if $\sqrt{\mathbb{E}(f_j^2)} = 0$.

Using iterated expectations the KKT condition can be written as $\mathbb{E}[(\mathbb{E}(a | X_j) - 2\lambda f_j - \gamma \nu_j)\eta] = 0$. Denote $E(a | X_j) \equiv P_j$. In particular, if we consider $\eta = \mathbb{E}[(\mathbb{E}(a | X_j) - 2\lambda f_j - \gamma \nu_j)]$, we can see that $\mathbb{E}[(\mathbb{E}(a | X_j) - 2\lambda f_j - \gamma \nu_j)] = 0$, i.e., $\mathbb{E}(a | X_j) - 2\lambda f_j - \gamma \nu_j = 0$ almost everywhere.

Then if $\sqrt{\mathbb{E}(P_j^2)} \leq \gamma$, we have $f_j = 0$, and we arrive at the following soft thresholding update:

$$f_j = \frac{1}{2\lambda} \left[1 - \frac{\gamma}{\sqrt{\mathbb{E}(P_j^2)}} \right]_+ P_j.$$

Now, going back to the constrained version, we need to select γ and λ so that the two constraints are tight. To get the sample version of this update we replace the conditional expectation P_j by an estimate $S_j a$, where S_j is a locally linear smoother.

Algorithm 1 Biconvex backfitting for SA-FCCA

input $\{(X_i, Y_i)\}$, parameters C_f, C_g , initial $g(Y_i)$

1. Compute smoothing matrices S_j and T_k .
2. Fix g . For each j , set $f_j \leftarrow \frac{S_j g}{\lambda}$ where $\lambda = \sqrt{\sum_{j=1}^{p_1} (g^T S_j^T S_j g)}$
3. **if** $\sum_{j=1}^{p_1} \|f_j\|_2 \leq C_f$, **break**
 else let \mathcal{F}_m denote the functions with maximum $\|\cdot\|_2$ norm. Set all other functions to 0.
 For each $f \in \mathcal{F}_m$, set $f \leftarrow \frac{C_f f}{\|\mathcal{F}_m\| \|f\|_2}$. **If** $\sum_{j=1}^{p_1} \|f_j\|_2^2 \leq 1$, **break**
 else set $f_j \leftarrow \left(1 - \frac{\gamma}{\sqrt{\|S_j g\|_2}} \right)_+ \frac{S_j g}{\lambda}$ where $\lambda = \sqrt{\sum_{j=1}^{p_1} \left\| \left(1 - \frac{\gamma}{\sqrt{\|S_j g\|_2}} \right)_+ S_j g \right\|_2^2}$ and γ
 is chosen so that $\sum_{j=1}^{p_1} \sqrt{g^T S_j^T S_j g} = C_f$
4. Center by setting each $f_j \leftarrow f_j - \text{mean}(f_j)$.
5. Fix f and repeat above to update g . Iterate both updates till convergence.

output Final functions f, g

3.4 Marginal Thresholding

The formulations of SA-KCCA and SA-FCCA above are not jointly convex, but are biconvex. Hence, iterative optimization algorithms may not be guaranteed to reach the globally optimal solution. To address this issue, we first run the algorithms without any sparsity constraint. The resulting nonsparse collections of functions are then used as initializations for the algorithm that incorporates the sparsity penalties. While such initialization works well for low dimensional problems, as p increases, the performance of the estimator goes down (Figure 3.1). To extend the algorithms to the high dimensional scenario, we propose marginal thresholding as a screening method to reject irrelevant variables and run the SA-FCCA and SA-KCCA models on the

Init	p=10	p=25	p=50
Random	0.05	0.009	-0.02
Non-sparse	0.97	0.62	0.26

Table 3.1: Test correlation from functions estimated by SA-FCCA for $n = 75$ samples, where $Y_1 = X_1^2$, all other dimensions are Gaussian noise. Random initializations don't work well for all data sizes. Initializing with the non-sparse formulation works well when $n > p$, but fails as $p \geq n$.

reduced dimensionality problem. For each pair of variables X_i and Y_j , we fit marginal functions to that pair by optimizing the criteria in either Equation Eq. 3.6 or Equation Eq. 3.7 *without* the sparsity constraints since we only consider one X and one Y covariate at a time. We then compute the correlation on held out data. This constructs a matrix M of size $p_1 \times p_2$ with (i, j) entry of the matrix representing an estimate of the marginal correlation between $f_i(X_i)$ and $g_j(Y_j)$. We then threshold the entries of M to obtain a subset of variables on which to run SA-FCCA and SA-KCCA. Theorem 3.5.3 discusses the theoretical properties of marginal thresholding as a screening procedure, and Section 3.6.2 presents results on marginal thresholding for high dimensional problems.

3.5 Main theoretical results

In this section we will characterize both the functional and kernel marginal thresholding procedures and study the theoretical properties of the estimators Eq. 3.6 and Eq. 3.7. We will state the main theorems and defer all proofs to the supplementary material.

The theoretical characterization of these procedures relies on *uniform* large deviation inequalities for the covariance between functions. For simplicity in this section we will assume all the univariate spaces are identical. In the RKHS case we restrict our attention to functions in a ball of a *constant* radius in the Hilbert space associated with a reproducing kernel K . In the functional case the univariate space is a second order Sobolev space where the integral of the square of the second derivative is bounded by a *constant*. With some abuse of notation we will denote these spaces \mathcal{C} . We are interested in controlling the quantity

$$\Theta_n = \sup_{f_j, g_k} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|$$

where $f_j, g_k \in \mathcal{C}, j \in \{1, \dots, p_1\}, k \in \{1, \dots, p_2\}$.

All results extend to the case when each covariate is endowed with a possibly distinct function space.

Lemma 3.5.1 (Uniform bound over RKHS). *Assume $\sup_x |K(x, x)| \leq M < \infty$, for functions*

$$f_j(x) = \sum_{i=1}^n \alpha_{ij} K_x(x, X_{ij}), g_k(y) = \sum_{i=1}^n \beta_{ik} K_y(y, Y_{ik})$$

$$\mathbb{P} \left(\Theta_n \geq \underbrace{\zeta + C \sqrt{\frac{\log((p_1 p_2)/\delta)}{n}}}_{\epsilon} \right) \leq \delta$$

where C is a constant depending only on M , and

$$\zeta = \max_{j,k} \frac{2}{n} \mathbb{E}_{X \sim x_j, Y \sim y_k} \sqrt{\sum_{i=1}^n K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})}$$

Note that ζ is independent of the dimensions p_1 and p_2 and that under the assumption that K is bounded, $\zeta = O(1/\sqrt{n})$. In some cases however this term can be much smaller. The second term depends only logarithmically on p_1 and p_2 and this *weak* dependence is the main reason our proposed procedures are consistent even when $p_1, p_2 > n$.

Lemma 3.5.2 (Uniform bound for Sobolev spaces). Assume $\|f\|_\infty \leq M \leq \infty$, then

$$\mathbb{P} \left(\Theta_n \geq \underbrace{\frac{C_1}{\sqrt{n}} + C_2 \sqrt{\frac{\log((p_1 p_2)/\delta)}{n}}}_{\epsilon} \right) \leq \delta$$

where C_1 and C_2 depend only on M .

Lemma 3.5.1 is proved via a Rademacher symmetrization argument of Bartlett and Mendelson [24] (see also [87]) while Lemma 3.5.2 is based on a bound on the bracketing integral of the Sobolev space (see [157]). The Rademacher bound gives a distribution dependent bound which can in some cases lead to faster rates.

We are now ready to characterize the marginal thresholding procedure described in Section 3.4. To study marginal thresholding we need to define *relevant* and *irrelevant* covariates. For each covariate X_j , denote

$$\alpha_j = \sup_{f_j, g_k \in \mathcal{C}, k \in \{1, \dots, p_2\}} \mathbb{E}(f_j(X_j) g_k(Y_k))$$

with $\mathbb{E}(f_j^2) \leq 1, \mathbb{E}(g_k^2) \leq 1$. A covariate X_j is considered irrelevant if $\alpha_j = 0$ and relevant if $\alpha_j > 0$. Similarly, for each Y_k we associate β_k defined analogously.

Now, assume that for every pair of covariates, we find the maximizer of the SA-FCCA or SA-KCCA objective over the given sample, over the appropriate class \mathcal{C} and with $\mathbb{E}(f_j^2) \leq 1, \mathbb{E}(g_k^2) \leq 1$. Recall that for marginal thresholding we do not enforce sparsity. The global maximization of the SA-KCCA objective can be efficiently carried out since it is equivalent to a generalized eigenvalue problem. For SA-FCCA however, the backfitting procedure is only guaranteed to find the global maximizer in the population setting.

Theorem 3.5.3. Given $\mathbb{P}(\Theta_n \geq \epsilon) \leq \delta$.

1. With probability at least $1 - \delta$, marginal thresholding at ϵ has no false inclusions.
2. Further, if we have that α_j or $\beta_k \geq 2\epsilon$ then under the same $1 - \delta$ probability event marginal thresholding at ϵ correctly includes the relevant covariate X_j or Y_k .

The importance of Lemmas 3.5.1 and 3.5.2 is that they provide values at which to threshold the marginal covariances. In particular, notice that the minimum sample covariance that can be reliably detected, with no false inclusions, falls rapidly with n and approaches zero even when $p_1, p_2 > n$.

In the spirit of early results on the LASSO of Greenshtein and Ritov [85], Juditsky and Nemirovski [105] we will establish the risk consistency or *persistence* of the empirical maximizers of the two objectives. Although we cannot guarantee that we find these empirical maximizers due to the non-convexity this result shows that with good initialization the formulations Eq. 3.6 and Eq. 3.7 can lead to solutions which have good statistical properties in high dimensions.

For SA-KCCA we will assume that our algorithm maximizes

$$\frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^{p_1} \mu_j f_j(X_{ij}) \right] \left[\sum_{k=1}^{p_2} \gamma_k g_k(Y_{ik}) \right]$$

over the classes

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p_1} \mu_j f_j(x_j), \mathbb{E}f_j = 0, \mathbb{E}f_j^2 = 1, \right. \\ \left. \|\mu\|_1 \leq C_f, \|\mu\|_2^2 + \gamma_f \sum_{j=1}^{p_1} \|f_j\|_{\mathcal{H}}^2 \leq 1 \right\}$$

$$\mathcal{G} = \left\{ g : g(x) = \sum_{k=1}^{p_2} \gamma_k g_k(y_k), \mathbb{E}g_k = 0, \mathbb{E}g_k^2 = 1, \right. \\ \left. \|\gamma\|_1 \leq C_g, \|\gamma\|_2^2 + \gamma_g \sum_{k=1}^{p_2} \|g_k\|_{\mathcal{H}}^2 \leq 1 \right\}$$

and for SA-FCCA we will assume that our algorithm maximizes the same objective over the same class without the RKHS constraint but which are instead in a Sobolev ball of constant radius. Denote these solutions (\hat{f}, \hat{g}) .

We will compare to an *oracle* which maximizes the population covariance

$$\text{cov}(f, g) \equiv \mathbb{E} \left[\sum_{j=1}^{p_1} \mu_j f_j(x_j) \right] \left[\sum_{k=1}^{p_2} \gamma_k g_k(y_k) \right]$$

Denote this maximizer by (f^*, g^*) . Our main result will show that these procedures are *persistent*, i.e., $\text{cov}(f^*, g^*) - \text{cov}(\hat{f}, \hat{g}) \rightarrow 0$ even if $p_1, p_2 > n$.

Theorem 3.5.4 (Persistence). *If $p_1 p_2 \leq e^{n^\xi}$ for some $\xi < 1$ and $C_f C_g = o(n^{(1-\xi)/2})$, then SA-FCCA and SA-KCCA are persistent over their respective function classes.*






Model	Test correlation				Precision/Recall		
	SA-FCCA	SA-KCCA	SCCA	KCCA	SA-FCCA	SA-KCCA	SCCA
 $Y = X^2$	0.96	0.99	0.05	0.44	1/1	1/1	0.28/0.14
 $Y = \text{abs}(X)$	0.98	0.99	0.06	0.35	1/1	1/1	0/0
 $Y = \cos(X)$	0.94	0.99	0.071	0.04	1/1	1/1	0.1/0.1
 $\log(Y) = \sin(X)$	0.91	0.93	0.22	0.09	1/1	1/1	0.71/0.66
 $Y = X$	0.99	0.99	0.99	0.98	1/1	1/1	1/1

Figure 3.1: Test correlations, and precision and recall for identifying relevant variables for the four different methods. SA-FCCA and SA-KCCA find strong correlations in the data, in both linear and non-linear settings. In all five data sets, SA-FCCA and SA-KCCA are always able to find the relevant variables.

3.6 Experiments

3.6.1 Non-linear correlations

We compare SA-FCCA and SA-KCCA with two models, sparse additive linear CCA (SCCA) [205] and kernel CCA (KCCA) [14]. Figure 3.1 shows the performance of each model, when run on data with $n = 150$ samples in $p_1 = 15, p_2 = 15$ dimensions, where only one relevant variable is present in X and Y (the remaining dimensions are Gaussian random noise). We report two metrics to measure whether the correct correlations are being captured by the different methods - (a) test correlation on 200 samples, using the estimated functions, and (b) precision and recall in identifying the correct variables involved in the correlation estimation. Each result is averaged over 10 repeats of the experiment. Since KCCA uses all data dimensions in finding correlations, its precision and recall are not reported.

When the relationship between the relevant variables is linear, all methods identify the correct variables and have high test correlation. While KCCA should be able to identify non-linear correlations, since it is strongly affected by the curse of dimensionality, it has poor test correlation even in $p = 15$ dimensions.

Both SA-FCCA and SA-KCCA correctly identify the relevant variables in all cases, and have high test correlation.

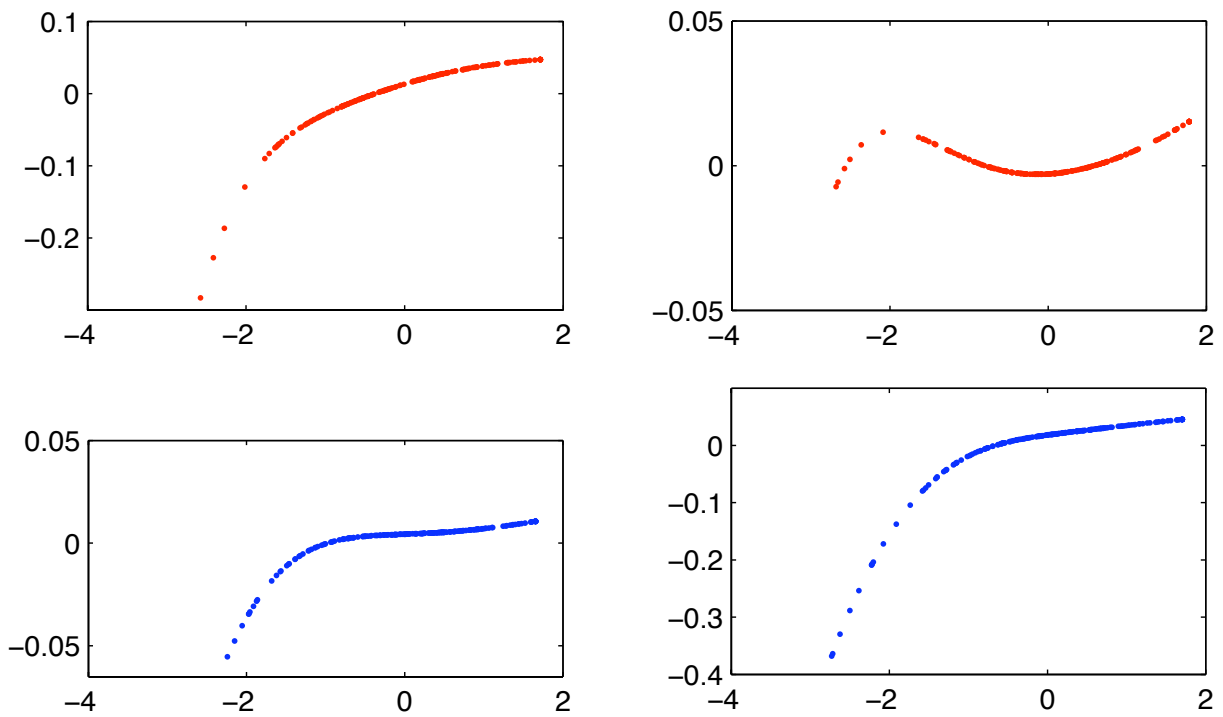


Figure 3.2: DLBCL data : The top row shows two of the functions $f_i(X_i)$ with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ with non-zero norms for Y in blue.

3.6.2 Marginal thresholding

We now test the efficiency of marginal thresholding by running an experiment for $n = 150$, $p_1 = 150$, $p_2 = 150$. We generate multiple relevant variables as:

$$f_i(X_i) = \cos\left(\frac{\pi}{2}X_i\right), \quad i \in \{1, 3\}, \quad f_i(X_i) = X_i^2, \quad i \in \{2, 4\}$$

$$Y_j = \sum_{i=1; i \neq j}^4 f_i(X_i) + \mathcal{N}(0, 0.1^2) \quad j \in \{1, 2, 3, 4\}.$$

Thus, there are four relevant variables in each data set. X and Y are sampled from a uniform distribution, and standardized before computing $f_i(X_i)$. Each $f_i(X_i)$ is also standardized before computing Y_j . We repeat the experiment by generating data 10 times, and report results in Table 3.2. Bandwidth in the different methods was selected using a plug-in estimator of the median distance between points in a single dimension. The sparsity and smoothness parameters for all methods were tuned using permutation tests, as described in Witten et al. [205], assuming that $C_f = C_g = C$, and $\gamma_f = \gamma_g = \gamma$.

We ran marginal thresholding by splitting the data into equal sized train and held out data, fitting marginal functions on the train data, computing functional correlation on the held out data, and picking a threshold so that $n/5$ elements of the thresholded correlation matrix are non-zero. We found that in all experiments, marginal thresholding always selected the relevant variables for the subsampled data. Table 3.2 shows the precision, recall and test correlations for the different methods. As can be expected, SA-FCCA and SA-KCCA are able to correctly identify the relevant variables, and the estimated functions have high correlation on test data.

We visualize the effect of the parameter tuning by plotting regularization paths, as the sparsity parameter is varied ($n=100$, $p_1=p_2=12$). For SA-FCCA and SA-KCCA, the norm of each function is plotted, and for sparse linear CCA, the absolute values of the entries of u and v are shown. Figure 3.3 shows how, unlike SCCA, SA-FCCA and SA-KCCA are able to separate the relevant and non-relevant variables over the entire range of the sparsity parameter.

Method	Test correlation	Precision	Recall
SA-FCCA	0.94	1	0.785
SA-KCCA	0.98	0.95	0.8
SCCA	0.02	0.02	0.36
KCCA	0.07	N/A	N/A

Table 3.2: Test correlations, precision and recall for identifying the correct relevant variables for the four different methods ($n = 150$, $p_1 = 150$, $p_2 = 150$). Marginal thresholding was used for selecting relevant variables before running SA-FCCA and SA-KCCA

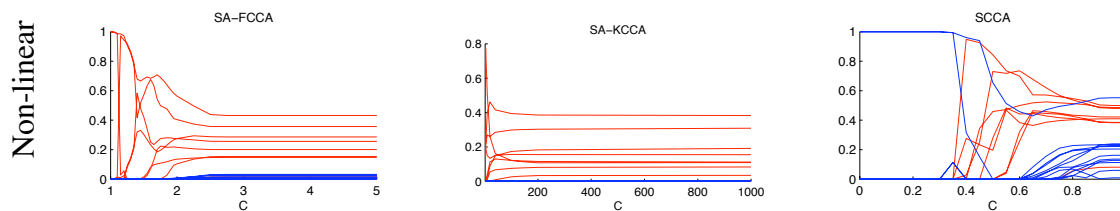


Figure 3.3: Regularization paths for non-linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.

3.6.3 Application to DLBCL data

We apply our non-linear CCA models to a data set of comparative genomic hybridization (CGH) and gene expression measurements from 203 diffuse large B-cell lymphoma (DLBCL) biopsy samples [127]. We obtained 1500 CGH measurements from chromosome 1 of the data, and 1500 gene expression measurements from genes on chromosome 1 and 2 of the data. The data was standardized, and Winsorized so that the data lies within two times the mean absolute deviation.

We used marginal thresholding to reduce the dimensionality of the problem, and then ran SA-FCCA. Permutation tests were used to pick an appropriate bandwidth and sparsity parameter, as described in Witten et al. [205]. We found that the model picked interesting non-linear relationships between CGH and gene expression data. Figure 3.2 shows the functions extracted by the SA-FCCA model from this data. Even though this data has been previously analyzed using linear models, we do not necessarily expect gene expression measurements from Affymetrix chips to be linearly correlated with array CGH measurements, even if the specific CGH mutation is truly affecting the gene expression. Further, the extracted functions in Figure 3.2 suggest that the changes in gene expression are dependent on the CGH measurements via a saturation function - as the copy number increases, the gene expression increases, until it saturates to a fixed level, beyond which increasing the copy numbers does not lead to an increase in expression. From a systems biology view point, such a prediction seems reasonable since single CGH mutations will not affect other pathways that are required to be activated for large changes in gene expression.

3.7 Discussion

In this chapter we introduced two proposals for nonparametric CCA and demonstrated their effectiveness both in theory and practice. Several interesting questions and extensions remain. CCA is often run on more than two data sets, and one is often interested in more than just the *principal* canonical direction. Chen and Liu [49] have proposed group sparse linear CCA for situations when a grouping of the covariates is known. These extensions all have natural non-parametric analogues which would be interesting to explore. As in the case of regression [115],

the KCCA formulation considered in this chapter can also be generalized to involve multiple kernels and kernels over groups of variables in a straightforward way.

While thresholding marginal correlations one can imagine exploiting the structure in the correlations. In particular, in the $(p_1 \times p_2)$ marginal correlations matrix we are looking for a *bicluster* of high entries in the matrix. Leveraging this structure could potentially allow us to detect weaker marginal correlations. Finally, an important application of kernel CCA is as a contrast function in independence testing. The additive formulations we have proposed allow for independence testing over more restricted alternatives but can be used to construct *interpretable* tests of independence. We discuss this further in chapter 9.

3.8 Technical Proofs

3.8.1 A derivation of the backfitting algorithm for FCCA

In this section we derive the biconvex backfitting algorithm for FCCA. In particular, consider the case when g is fixed and let a denote the vector of $(g(Y_1), \dots, g(Y_n))^T$ in the sample setting, and let it denote the function g in the population setting.

It is instructive to first consider the population setting. The optimization problem becomes

$$\begin{aligned} \max_{f \in \mathcal{F}} \quad & \mathbb{E}[f(X)a] \\ \text{subject to} \quad & \|f\|_2^2 \leq 1 \\ & \|f\|_1 \leq C_f. \end{aligned}$$

The norms are defined as $\|f\|_1 = \sum_{j=1}^{p_1} \sqrt{\mathbb{E}(f_j^2(x_j))}$ and $\|f\|_2^2 = \sum_{j=1}^{p_1} \mathbb{E}(f_j^2(x_j))$.

Consider the Lagrange problem,

$$\max_f \min_{\lambda \geq 0, \gamma \geq 0} \mathbb{E}[f(X)a] - \lambda(\|f\|_2^2 - 1) - \gamma(\|f\|_1 - C_f).$$

The Fréchet derivative w.r.t. f_j along the direction η gives one of the KKT conditions $\mathbb{E}[(a - 2\lambda f_j - \gamma \nu_j)\eta] = 0$ for all η in the Hilbert space \mathcal{H}_j , where $\nu_j = \frac{f_j}{\sqrt{\mathbb{E}(f_j^2)}}$ if $\sqrt{\mathbb{E}(f_j^2)}$ is not 0, and is the set $\{u_j \in \mathcal{H}_j | \mathbb{E}(u_j^2) \leq 1\}$ if $\sqrt{\mathbb{E}(f_j^2)} = 0$.

Using iterated expectations the KKT condition can be written as $\mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma \nu_j)\eta] = 0$. Now, if we denote $E(a|X_j) = P_j$. In particular, if we consider $\eta = \mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma \nu_j)]$,

we can see that

$$\mathbb{E}[(\mathbb{E}(a|X_j) - 2\lambda f_j - \gamma \nu_j)] = 0.$$

This implies that

$$\begin{aligned} \mathbb{E}(a|X_j) - 2\lambda f_j - \gamma \nu_j &= 0 \text{ almost everywhere} \\ \text{i.e. } P_j - 2\lambda f_j &= \gamma \nu_j. \end{aligned}$$

If $\sqrt{\mathbb{E}(P_j^2)} \leq \gamma$, we have $f_j = 0$, and we arrive at the following

$$\begin{aligned} f_j \left(2\lambda + \frac{\gamma}{\sqrt{\mathbb{E}(f_j^2)}} \right) &= P_j \text{ if } \sqrt{\mathbb{E}(P_j^2)} > \gamma \\ f_j &= 0 \text{ otherwise} \end{aligned}$$

and this gives the following soft threshold update:

$$f_j = \frac{1}{2\lambda} \left[1 - \frac{\gamma}{\sqrt{\mathbb{E}(P_j^2)}} \right]_+ P_j.$$

We analyze the Lagrangian in the 3 cases (i.e. all constraints are tight, only the 2-norm constraints are tight, and only the 1-norm constraints are tight).

1. When only the 2-norm constraint is tight, $\gamma = 0$ and λ is selected to make the 2-norm be 1.
2. When only the 1-norm constraint is tight, we use the equation above with $\lambda = 0$ and see that only the f_j s with the largest $\sqrt{\mathbb{E}(P_j^2)}$ are non-zero.
3. When both constraints are tight, we use the soft-threshold update with λ and γ selected to make both constraints tight.

Now, we can define the algorithm in the finite sample case, as an analog of the algorithm for the basic problem in the linear case. For a fixed g , FCCA problem can be solved using the following algorithm.

1. Test for case 1 by setting $f_j(X_j) = \frac{S_j a}{\lambda}$ for each j , where $\lambda^2 = \frac{1}{n} \sum_{j=1}^p \|S_j a\|_2^2$. If the solution satisfies $\|f\|_1 \leq c_1$ this is the required f .

2. Test for case 2, in this case we find $\|S_j a\|_2$ for each j , and find all k such that $\|S_k a\|_2 \geq \|S_j a\|_2$ for all j . Denote the cardinality of this set ϕ . Set

$$f_k(X_k) = \frac{C_f S_k a}{\phi \|S_k a\|}$$

for all k such that $\|S_k a\|_2 \geq \|S_j a\|_2$ for all j , and all other $f_j = 0$. If $\|f\|_2 \leq 1$ this is the required f .

3. If neither of the above cases are satisfied then in this case $f_j(X_j) = \frac{S_\gamma(S_j a)}{\lambda}$ where $\lambda^2 = \frac{1}{n} \sum_{j=1}^p \|S_\gamma(S_j a)\|_2^2$ for each j . where γ is chosen so that $\|f\|_1 = C_f$.

Here S_j is a linear smoother and is used to estimate the conditional expectation of a given X_j , i.e. if $P_j = \mathbb{E}(a|X_j)$ then $\widehat{P}_j = S_j a$.

3.8.2 Uniform bounds

We will first prove Lemma 3.5.1 and then give a proof sketch for Lemma 3.5.2.

Proof. We will limit our attention to functions

$$f_j \in B_{\mathcal{H}}(1)$$

since the general case for a constant radius follows by a simple rescaling argument. We have the condition

$$\sup_x |K(x, x)| \leq c < \infty.$$

This also implies the uniform boundedness of the univariate functions by a simple argument.

$$\sup_x |f_j(x)| = \sup_x |\langle f_j, K(\cdot, x) \rangle| \leq \sup_x \|f_j\|_{\mathcal{H}} \sqrt{K(x, x)}$$

Thus, we have

$$\sup_x |f_j(x)| \leq C$$

for some absolute constant C .

Recall that we wish to uniformly control

$$\Omega_n = \sup_{f_j, g_k \in \mathcal{C}, j \in \{1, \dots, p_1\}, k \in \{1, \dots, p_2\}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|.$$

Let us first analyze

$$\Theta_n = \sup_{f_j, g_k \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|$$

which is just Ω_n for a fixed pair j, k . The bound on Ω_n will then follow from a union bound.

Deviation from its expectation is a simple consequence of the boundedness of functions and McDiarmid's inequality, i.e. for some absolute constant C , we have

$$\mathbb{P}(\Theta_n - \mathbb{E}\Theta_n > t) \leq \exp\left(\frac{-nt^2}{C}\right).$$

Now, we need to understand the expectation. A symmetrization argument gives us

$$\mathbb{E}\Theta_n \leq 2\mathcal{R}(\mathcal{C})$$

where

$$\mathcal{R}(\mathcal{C}) = \mathbb{E}_{X,Y,\sigma} \left(\sup_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_{ij}) g_k(Y_{ik}) \right).$$

A bound on $\mathcal{R}(\mathcal{C})$ is given by Lemma 16 in the paper of Gretton et al. [87]. They show,

$$\mathbb{E}_{X,Y,\sigma} \left(\sup_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(X_{ij}) g_k(Y_{ik}) \right) \leq \frac{1}{n} \mathbb{E}_{X,Y} \sqrt{\sum_{i=1}^n K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})}.$$

This gives us a bound on Θ_n , and to get a bound on Ω_n we just union bound over the $p_1 p_2$ possible choices for j, k .

Defining,

$$\zeta = \max_{j,k} \frac{2}{n} \mathbb{E}_{X \sim x_j, Y \sim y_k} \sqrt{\sum_{i=1}^n K(X_{ij}, X_{ij}) K(Y_{ik}, Y_{ik})},$$

we have for some absolute constant C

$$\mathbb{P} \left(\Theta_n \geq \zeta + C \sqrt{\frac{\log((p_1 p_2)/\delta)}{n}} \right) \leq \delta.$$

□

For SA-FCCA we have a different class of functions. Ravikumar et al. [157] show the following result for uniformly bounded (by a constant) f and g in a second order Sobolev space, for an absolute constant C ,

$$\omega \equiv \mathbb{E} \left(\sup_{f_j, g_k \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| \right) \leq \frac{C}{\sqrt{n}}.$$

Since, the functions are uniformly bounded we can now use McDiarmid's inequality to get for some C'

$$\mathbb{P} \left(\sup_{f_j, g_k \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| - \omega \geq t \right) \leq \exp\left(\frac{-t^2 n}{C'}\right).$$

Now, applying the union bound over j and k we get the desired lemma. Again, defining

$$\Omega_n = \sup_{f_j, g_k \in \mathcal{C}, j \in \{1, \dots, p_1\}, k \in \{1, \dots, p_2\}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right|,$$

we get

$$\mathbb{P} \left(\Omega_n \geq \frac{C_1}{\sqrt{n}} + C_2 \sqrt{\frac{\log((p_1 p_2)/\delta)}{n}} \right) \leq \delta$$

3.8.3 Marginal thresholding

In this section we prove the following result:

Theorem 3.8.1. *Given*

$$\mathbb{P} \left(\sup_{f_j, g_k \in \mathcal{C}, j \in \{1, \dots, p_1\}, k \in \{1, \dots, p_2\}} \left| \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) - \mathbb{E}(f_j(X_j) g_k(Y_k)) \right| \geq \epsilon \right) \leq \delta$$

with probability at least $1 - \delta$, marginal thresholding at ϵ has no false inclusions. Further, if we have that α_j or $\beta_k \geq 2\epsilon$ then under the same $1 - \delta$ probability event marginal thresholding at ϵ correctly includes the relevant covariate X_j or Y_k .

Proof. The first part is straightforward. In particular, we know for any irrelevant X_j for any Y_k and $f_j, g_k \in \mathcal{C}$, $\mathbb{E} f_j(X_j) g_k(Y_k) = 0$, and in the at least $1 - \delta$ probability event we have

$$\max_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) < \epsilon.$$

For the second part, consider a particular relevant covariate X_j , denote

$$\theta^* = \max_{f_j, g_k \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}).$$

It suffices to show that if $\alpha_j \geq 2\epsilon \implies \theta^* \geq \epsilon$.

Denote, $(f_j^*, g_k^*) = \arg \sup_{f_k, g_k \in \mathcal{C}} \mathbb{E}(f_j(X_j) g_k(Y_k))$. Then in the at least $1 - \delta$ probability event we have,

$$\theta^* \geq \frac{1}{n} \sum_{i=1}^n f_j^*(X_{ij}) g_k^*(Y_{ik}) \geq \mathbb{E}(f_j^*(X_j) g_k^*(Y_k)) - \epsilon \geq \epsilon.$$

□

3.8.4 Persistence

We will show the high dimensional persistence of the global optimizers of the SA-FCCA and SA-KCCA objectives.

We will prove the result for SA-FCCA and give a proof sketch for SA-KCCA.

Let us assume that the SA-FCCA estimator is chosen to maximize the objective

$$\frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^{p_1} \mu_j f_j(X_{ij}) \right] \left[\sum_{k=1}^{p_2} \gamma_k g_k(Y_{ik}) \right]$$

over the classes

$$\mathcal{F} = \left\{ f : f(x) = \sum_{j=1}^{p_1} \mu_j f_j(x_j), \mathbb{E} f_j = 0, \quad \mathbb{E} f_j^2 = 1, \|\mu\|_1 \leq C_f, \|\mu\|_2^2 \leq 1 \right\}$$

$$\mathcal{G} = \left\{ g : g(x) = \sum_{k=1}^{p_2} \gamma_k g_k(y_k), \mathbb{E} g_k = 0, \quad \mathbb{E} g_k^2 = 1, \|\gamma\|_1 \leq C_g, \|\gamma\|_2^2 \leq 1 \right\}.$$

An analogous role to risk in classification/regression problems is played by the (negative) covariance,

$$\text{cov}(f, g) = \mathbb{E} \left[\sum_{j=1}^{p_1} \mu_j f_j(X_j) \right] \left[\sum_{k=1}^{p_2} \gamma_k g_k(Y_k) \right].$$

Theorem 3.8.2. *If $p_1 p_2 \leq e^{n^\xi}$ for some $\xi < 1$. Then,*

$$\text{cov}(f_n^*, g_n^*) - \text{cov}(\hat{f}_n, \hat{g}_n) = O_P \left(\frac{C_f C_g}{n^{(1-\xi)/2}} \right). \quad (3.8)$$

If $C_f C_g = o(n^{(1-\xi)/2})$ the FCCA procedure described is persistent, i.e. $\text{cov}(f_n^, g_n^*) - \text{cov}(\hat{f}_n, \hat{g}_n) \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. We can write

$$\text{cov}(f, g) = \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \mu_j \gamma_k \mathbb{E}[f_j(X_j) g_k(Y_k)] \quad (3.9)$$

and

$$\hat{C}(f, g) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \mu_j \gamma_k f_j(X_{ij}) g_k(Y_{ik}). \quad (3.10)$$

Now, we have (using Holder's inequality)

$$|\hat{C}(f, g) - \text{cov}(f, g)| \leq \|\mu\|_1 \|\gamma\|_1 \max_{jk} \left[\frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) \right] - \mathbb{E}(f_j(X_j) g_k(Y_k)). \quad (3.11)$$

Now, we are almost done. Using Lemma 3.5.2 we know that we can uniformly bound

$$\left[\frac{1}{n} \sum_{i=1}^n f_j(X_{ij}) g_k(Y_{ik}) \right] - \mathbb{E}(f_j(X_j) g_k(Y_k))$$

over all f_j, g_k in our function class and over all $j \in \{1, \dots, p_1\}, k \in \{1, \dots, p_2\}$. In particular, this quantity is $O_P \left(\sqrt{\frac{\log(p_1 p_2)}{n}} \right)$.

Now, this gives us that

$$|\widehat{C}(f, g) - \text{cov}(f, g)| = O_P \left(C_f C_g \sqrt{\frac{\log(p_1 p_2)}{n}} \right) = O_P \left(\frac{C_f C_g}{n^{(1-\xi)/2}} \right). \quad (3.12)$$

Using this we have,

$$\text{cov}(\widehat{f}_n, \widehat{g}_n) \geq \widehat{C}(\widehat{f}_n, \widehat{g}_n) - O_P \left(\frac{C_f C_g}{n^{(1-\xi)/2}} \right) \geq \widehat{C}(f_n^*, g_n^*) - O_P \left(\frac{C_f C_g}{n^{(1-\xi)/2}} \right) \geq \text{cov}(f_n^*, g_n^*) - O_P \left(\frac{C_f C_g}{n^{(1-\xi)/2}} \right) \quad (3.13)$$

and the result follows. \square

The proof for the persistence of SA-KCCA follows an almost identical argument. We make two minor modifications. As described in the main text we bound the Rademacher term as $O(1/\sqrt{n})$ by only using the boundedness of the kernel. We can then follow the proof of this theorem exactly, replacing the use of Lemma 3.5.2 with Lemma 3.5.1.

3.9 Additional discussion

3.9.1 Discussion of SA-FCCA v/s SA-KCCA

SA-FCCA and SA-KCCA offer different advantages and disadvantages and neither is completely dominated by the other. The methods are two instances of the same approach, which is to use a nonparametric additive model.

From an optimization perspective, SA-KCCA works over RKHS, leading to an optimization problem over a finite parameter space for which strong convergence guarantees can be made. For SA-FCCA however, we use backfitting, which is typically known to converge only in the population setting. From a statistical perspective, stronger results are known for the kernel version in the regression setting. From a practitioner's perspective, these algorithms perform comparably statistically. Computationally, the SA-FCCA algorithm is considerably more simple - after some pre-computations, the coordinate descent back-fitting algorithm only requires matrix-vector multiplications in each iteration, and typically converges in a small number of iterations. SA-KCCA requires us to optimize a second order cone-program which, although convex, is not amenable to fast coordinate descent algorithms.

p	5	10	25	50	75	100	150
Test correlation	0.9999	1.0000	1.0000	0.6846	0.9079	0.4967	0.2918
Precision	1.0000	1.0000	1.0000	0.7000	0.9000	0.5000	0.3000
Recall	1.0000	1.0000	1.0000	0.7000	0.9000	0.5000	0.3000

Table 3.3: Results for SCCA on linear data $Y_1 = X_1 + \mathcal{N}(0, 1)$ with $n = 100$ samples. As p increases, the performance of the model decreases.

p	5	10	25	50	75	100	150
Test correlation	0.9672	0.9717	0.6178	0.2564	0.2040	0.0294	0.0959
Precision	1.0000	1.0000	0.6000	0.4000	0.2000	0	0.2000
Recall	1.0000	1.0000	0.6000	0.4000	0.2000	0	0.2000

Table 3.4: Results for SA-FCCA (without marginal thresholding) on quadratic data $Y_1 = X_1^2 + \mathcal{N}(0, 1)$ with $n = 100$ samples. As p increases, the performance of the model decreases.

There is a clear dichotomy here from a statistical/optimization theory perspective, we would recommend the SA-KCCA formulation but from a practical perspective we would recommend the SA-FCCA formulation.

Computational costs: The computational cost of each inner loop optimization of SA-FCCA when it is done to an accuracy of ϵ is $O(n^2 \max(p_1, p_2)/\epsilon)$ using the algorithm we propose. SA-KCCA using a standard interior point solver has complexity $O(n^3 \max(p_1, p_2)^3 \log(1/\epsilon))$. SA-FCCA also requires a pre-computation of smoother matrices which takes $O(n^3 \max(p_1, p_2))$. These methods typically require a small number of outer-loop iterations to converge.

It is also worth noting that these non-parametric methods are more computationally intensive than both sparse linear CCA which requires $O(n^2/\epsilon)$ for each inner loop iteration, and kernel CCA which requires $O(n^2 \log n)$ in total after computing the Gram matrices.

Notice also that in linear CCA we are learning $p_1 + p_2$ parameters, in kernel CCA we are learning $2n$ parameters, while in SA-KCCA we are learning the much larger $n(p_1 + p_2)$ parameters. A direct comparison of the number of parameters in SA-FCCA is subtle, since at least from a degrees of freedom perspective this depends on the smoothness of the target function.

3.9.2 Marginal thresholding is needed to get high accuracy in high dimensions

We show that for both linear SCCA (Table 3.3) and non-linear SA-FCCA (Table 3.4) models to measure correlation, the models do not have good performance when $p \sim n$. Hence, using a screening procedure to extract variables of interest before running CCA is essential.

3.9.3 Simulation Details

This section describes how the simulated data was generated for the experiments in Section 3.6.1. The algorithm requires a function $f(x)$ that defines the relationship between X and Y . Four different functions were used, as defined in the results (Figure 3.1).

Algorithm 2 Generate simulated data

input n, p_1, p_2 , function $f(x)$.

1. Pick relevant feature r_x and r_y of X and Y randomly from $\{1, \dots, p_1\}$ and $\{1, \dots, p_2\}$ resp.
2. For $j = 1 \dots p_1$
 For $i = 1 \dots n$
 $X(i, j) = \mathcal{N}(0, 1)$;
3. For $j = 1 \dots p_2$
 For $i = 1 \dots n$
 if ($j == r_y$)
 $Y(i, r_y) = f(X(i, r_x)) + \mathcal{N}(0, 0.1^2)$;
 else
 $Y(i, j) = \mathcal{N}(0, 1)$;

output X, Y

3.9.4 Comparison of regularization paths

As the sparsity parameter is varied, different number of features are selected. We plot the regularization paths obtained by varying the sparsity parameter for linear data (Figure 3.4). The linear data was selected in a similar manner to Section 3.6.2 with $n = 100, p_1 = p_2 = 12$, so that X and Y have 4 relevant variables each.

For SA-FCCA and SA-KCCA, the norm of each function is plotted, and for sparse linear CCA, the absolute values of u and v are shown, as a function of the sparsity parameter. Figure 3.4 shows that when the true relationship between the variables is linear, all three models separate the relevant and irrelevant variables. Note that the bandwidth of SA-FCCA and SA-KCCA were not tuned in this problem, so both models are capable of extracting the correct linear relationships without adjusting the bandwidth heavily.

3.9.5 Comparison of SA-FCCA and SA-KCCA on DLBCL data

We ran SA-KCCA on the DLBCL data, on which SA-FCCA results were reported in Section 3.6.3. We observed that the same co-variates were picked as relevant by both SA-FCCA and SA-KCCA. The functions extracted by SA-KCCA are shown in Figure 3.5. Note that the functions appear to be mirror-images of the ones extracted by SA-FCCA in Figure 3.2. Since

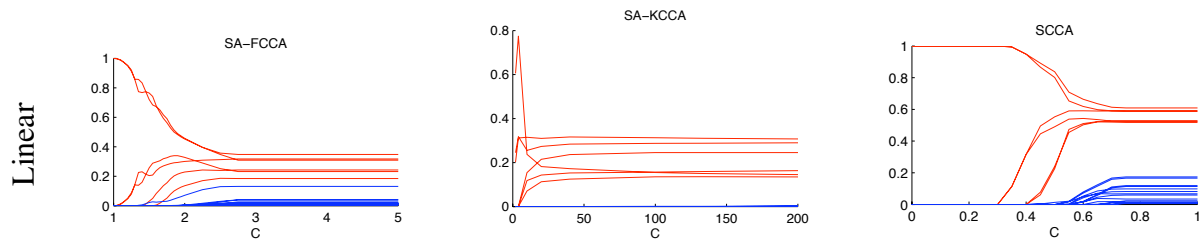


Figure 3.4: Regularization paths for linear correlations in the data, for SA-FCCA, SA-KCCA and SCCA resp. The paths for the relevant variables (in X and Y) are shown in red, the irrelevant variables are shown in blue.

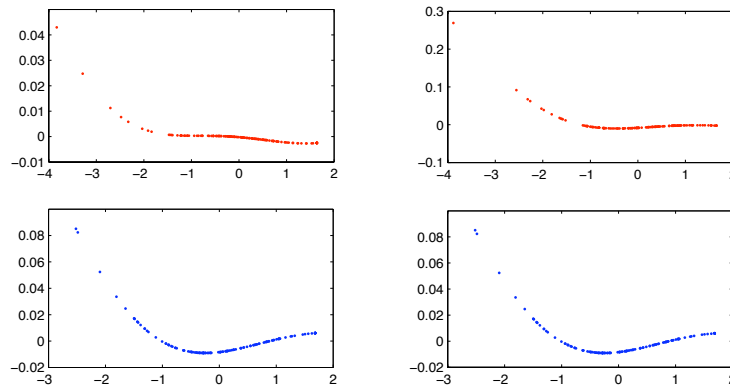


Figure 3.5: KCCA output on DLBCL data : The top row shows two of the functions $f_i(X_i)$ v/s X_i with non-zero norms for X in red, and the bottom row shows two functions $g_j(Y_j)$ v/s Y_j with non-zero norms for Y in blue.

a mirror image of the function still preserves the non-linear correlations, we conclude that SA-FCCA and SA-KCCA work comparably in such predictions.

Chapter 4

Noise Thresholds for Spectral Clustering

Spectral clustering algorithms are a family of algorithms that partition data according to the eigenvectors of a similarity matrix formed from the data. Despite considerable empirical success, the theoretical understanding of spectral clustering is somewhat limited. In this chapter we study k -way and hierarchical spectral clustering algorithms on a general class of noisy *structured* similarity matrices. For hierarchical clustering, we show that recursive application of a simple spectral clustering algorithm can tolerate noise that grows with the number of data points while still recovering the hierarchical clusters with high probability. For k -way clustering, we derive conditions on the similarity matrix under which spectral clustering perfectly partitions the data, relating the noise variance to the minimum within-cluster similarity, number of clusters, and number of data points. We complement these results with a minimax analysis, identifying the information theoretic limits for the clustering problem with tight upper and lower bounds. We verify our results with experiments on simulated and real world data.

4.1 Introduction

Clustering, a fundamental and ubiquitous problem in machine learning, is the task of organizing data points into homogenous groups using a given measure of similarity. Two popular forms of clustering are k -way, where an algorithm directly partitions the data into k disjoint sets, and *hierarchical*, where the algorithm organizes the data into a hierarchy of groups. Popular algorithms for the k -way problem include k -means, spectral clustering, and density-based clustering, while *agglomerative* methods that merge clusters from the bottom up are popular for the latter problem.

Spectral algorithms are a family of clustering algorithms which embed the data points by projection onto a few eigenvectors of a similarity matrix or data graph, constructed from the data points, and uses this *spectral embedding* to find a good clustering.

In this chapter, we study the statistical performance of spectral clustering, focusing on the robustness to noise. To obtain quantitative results, we introduce a class of structured similarity matrices seeded with the true clustering but corrupted with noise, and characterize when spectral clustering correctly recovers the true clustering in terms of the model parameters. In particular, we are interested in quantifying the relationship between the number of data points, the number of clusters, and a signal-to-noise ratio, which parameterizes our family of similarity matrices.

The main contributions of this chapter are:

1. We leverage results from perturbation theory in a novel analysis of a spectral algorithm for hierarchical clustering to understand its behavior in the presence of noise. We provide strong guarantees on its correctness; in particular, we show that the amount of noise spectral clustering tolerates can grow rapidly with the size of the smallest cluster we want to resolve.
2. We sharpen existing results on k -way spectral clustering. In contrast with earlier work, we provide precise error bounds through a careful characterization of a k -means style algorithm run on the spectral embedding of the data.
3. We also address the issue of optimal noise thresholds via the use of minimax theory. In particular, we establish tight *information-theoretic upper and lower bounds* for cluster resolvability for both the k -way and hierarchical settings that we consider.

4.2 Related Work

There are several high-level justifications for the success of spectral clustering. The algorithm has deep connections to various graph-cut problems, random walks on graphs, electric network theory, and via the graph Laplacian to the Laplace-Beltrami operator. See the survey paper of von Luxburg [195] for an overview.

Several authors (see the work of von Luxburg et al. [196] and references therein) have shown various forms of asymptotic convergence for the Laplacian of a graph constructed from random samples drawn from a distribution on or near a manifold. These results however often do not easily translate into precise guarantees for successful recovery of clusters, which is the emphasis of our work.

There has also been some theoretical work on spectral algorithms for cluster recovery in random graph models. McSherry [138] studies the “cluster-structured” random graph model in which the probability of adding an edge can vary depending on the clusters the edge connects. He considers a specialization of this model, the planted partition model, which specifies only two probabilities, one for inter-cluster edges and another for intra-cluster edges. In this case, we can view the observed adjacency matrix as a random perturbation of a low rank “expected” adjacency matrix which encodes the cluster membership. McSherry shows that one can recover the clusters from a low rank approximation of the observed (noisy) adjacency matrix. These results show that low-rank matrices have spectra that are robust to noise. Our results however, show that we

can obtain similar insensitivity (to noise) guarantees for a class of interesting structured *full-rank* matrices, indicating that this robustness extends to a much broader class of matrices.

More recently, Rohe et al. [163] analyze spectral clustering in the stochastic block model (SBM), which is an example of a structured random graph. They consider the *high-dimensional* scenario where the number of clusters k grows with the number of data points n and show that under certain assumptions the *average* number of mistakes made by spectral clustering $\rightarrow 0$ with increasing n . Our work on hierarchical clustering also has the same high-dimensional flavor since the number of clusters we resolve grows with n . However, in the hierarchical clustering setting, errors made at the bottom level propagate up the tree and we need to make precise arguments to ensure that the *total* number of errors $\rightarrow 0$ with increasing n (see Theorem 4.3.1).

Since Rohe et al. [163] and McSherry [138] consider random graph models, the “noise” on each entry has *bounded* variance. We consider more general noise models and study the relation between errors in clustering and noise variance. Another related line of work is on the problem of spectrally separating mixtures of Gaussians ([1, 35, 107]).

In a seminal paper, Ng et al. [145] studied k -way clustering and showed that the eigenvectors of the graph Laplacian are stable in 2-norm under small perturbations. This justifies the use of k -means in the perturbed subspace since ideally without noise, the spectral embedding by the top k eigenvectors of the graph Laplacian reflects the true cluster memberships. However, closeness in 2-norm does not translate into a strong bound on the *total number* of errors made by spectral clustering.

More recently, Huang et al. [98] have studied the misclustering rate of spectral clustering under the somewhat unnatural assumption that every coordinate of the Laplacian’s eigenvectors are perturbed by identically distributed noise. In contrast, we specify our noise model as an additive perturbation to the similarity matrix, making no direct assumptions on how this affects the spectrum of the Laplacian. We show that the eigenvectors are stable in ∞ -norm and use this result to precisely bound the misclustering rate of our algorithm.

In this chapter of the thesis we analyze spectral clustering using the unnormalized Laplacian which, as we demonstrate, is well suited to the homogenous degree models we consider. Since the publication of our paper [22] there has been an increased interest in spectral clustering in degree inhomogenous models like the *degree corrected* stochastic block model (also known as the *extended planted partition model*) considered by Chaudhuri et al. [45], Chen et al. [48], Jin [101]. Chaudhuri et al. [45] in particular shows that in this situation using a modified normalized graph Laplacian is more appropriate. Although beyond the scope of this thesis we expect that many of our techniques will be useful in these problems, particularly in obtaining ∞ -norm perturbation bounds for the normalized Laplacian.

4.3 Hierarchical Clustering

Our first set of results focus on binary hierarchical clustering, which is formally defined as:

$[\alpha_{s \cdot LL}, \beta_{s \cdot LL}]$	$[\alpha_{s \cdot L}, \beta_{s \cdot L}]$	$[\alpha_s, \beta_s]$		\dots
$[\alpha_{s \cdot L}, \beta_{s \cdot L}]$	$[\alpha_{s \cdot LR}, \beta_{s \cdot LR}]$			
$[\alpha_s, \beta_s]$		$[\alpha_{s \cdot RL}, \beta_{s \cdot RL}]$	$[\alpha_{s \cdot R}, \beta_{s \cdot R}]$	\dots
		$[\alpha_{s \cdot R}, \beta_{s \cdot R}]$	$[\alpha_{s \cdot RR}, \beta_{s \cdot RR}]$	
		\vdots		

Figure 4.1: An ideal matrix for the hierarchical problem.

Definition 1. A **hierarchical clustering** \mathcal{T} on data points $\{X_i\}_{i=1}^n$ is a collection of clusters (subsets of the points) such that $C_0 = \{X_i\}_{i=1}^n \in \mathcal{T}$ and for any $C_i, C_j \in \mathcal{T}$, either $C_i \subset C_j$, $C_j \subset C_i$ or $C_i \cap C_j = \emptyset$.

A **binary hierarchical clustering** \mathcal{T} is a hierarchical clustering such that for each non-atomic $C_k \in \mathcal{T}$, there exists two proper subsets $C_i, C_j \in \mathcal{T}$ with $C_i \cap C_j = \emptyset$ and $C_i \cup C_j = C_k$.

We label each cluster by a sequence of *Ls* and *Rs* so that $C_{s \cdot L}$ and $C_{s \cdot R}$ partitions C_s , $C_{s \cdot LL}$ and $C_{s \cdot LR}$ partitions $C_{s \cdot L}$ and so on.

A large class of clustering algorithms operate exclusively on similarities (or distances) between the data points and are agnostic to the representation of the points themselves. In practice, one typically specifies an appropriate similarity (distance) function between the data points. Ideally, at all levels of the hierarchy, points within a cluster are more similar to each other than to points outside of that cluster. In this chapter we work directly with a family of structured similarity matrices, so that we can circumvent the problem of selecting a good similarity metric, a task which usually requires the application of some domain knowledge.

We work with the noisy hierarchical block matrix, which captures the intuition of our ideal similarity matrix, where between-cluster similarity is higher than within-cluster similarity, but allows for deviations from the ideal situation. These matrices can be decomposed into an ideal matrix and a noise term:

Definition 2. A similarity matrix W is a **noisy hierarchical block matrix** (noisy HBM) if $W \triangleq A + R$ where A is ideal and R is a perturbation matrix, defined as follows:

- An **ideal similarity matrix** is characterized by an interval $[\alpha_s, \beta_s] \subset \mathbb{R}^+$ for each cluster C_s (see Figure 4.1) such that the between-subcluster similarities, i.e. $A_{x,y}$ for $x \in C_{s \cdot L}, y \in C_{s \cdot R}$ must lie between $[\alpha_s, \beta_s]$. Additionally, the within-subcluster similarities of each of the two subclusters must be larger than β_s : $\min \alpha_{s \cdot R}, \alpha_{s \cdot L} > \beta_s$.
- A symmetric $(n \times n)$ matrix R is a **perturbation matrix** with parameter σ if
 1. $\mathbb{E}(R_{ij}) = 0$,
 2. the upper triangular and diagonal entries of R are independent and sub-Gaussian, that is $\mathbb{E}(\exp(tR_{ij})) \leq \exp(\frac{\sigma^2 t^2}{2})$.

Finally, we define the combinatorial Laplacian matrix, which will be the focus of our spectral

<p>INPUT: (noisy) $n \times n$ similarity matrix W</p> <ol style="list-style-type: none"> 1. Compute Laplacian $L = D - W$ 2. $v_2 \leftarrow$ smallest non-constant eigenvector of L 3. $C_1 \leftarrow \{i : v_2(i) \geq 0\}, C_2 \leftarrow \{j : v_2(j) < 0\}$ 4. $\mathcal{C} \leftarrow \{C_1, C_2\} \cup \text{HS}(W_{C_1}) \cup \text{HS}(W_{C_2})$ <p>OUTPUT: \mathcal{C}</p>

Figure 4.2: Hierarchical Spectral Clustering (HS)

algorithm and the subsequent analysis.

Definition 3. The *combinatorial Laplacian* L of a matrix W is defined as $L \triangleq D - W$ where D is a diagonal matrix with $D_{ii} \triangleq \sum_{j=1}^n W_{ij}$.

We note that other analyses of spectral clustering have studied other Laplacian matrices, particularly, the *normalized Laplacians* defined as $L_n \triangleq D^{-1}L$ and $L_n \triangleq D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. However, as we show in Section 4.6, the normalized Laplacian can mis-cluster points even for an ideal noiseless similarity matrix making it unsuitable for our problem.

We first state the following general assumptions, which we place on the *ideal* similarity matrix A :

Assumption 1. For all i, j , $0 < A_{ij} \leq \beta^*$ for some constant β^* .

Assumption 2. (*Balanced clusters*) There is a constant $\eta \geq 1$ such that at every split of the hierarchy

$$\frac{|C_{\max}|}{|C_{\min}|} \leq \eta,$$

where $|C_{\max}|, |C_{\min}|$ are the sizes of the biggest and smallest clusters respectively.

Assumption 3. (*Range Restriction*) For every cluster s ,

$$\min\{\alpha_{s-L}, \alpha_{s-R}\} - \beta_s > \eta(\beta_s - \alpha_s).$$

It is important to note that these assumptions are placed *only* on the ideal matrices. The noisy HBMs can, and with high probability will, violate these assumptions.

We assume that the entries of A are strictly greater than 0 for technical reasons; we believe, as confirmed empirically, that this restriction is not necessary for our results to hold. Assumption 2 says that at every level the largest cluster is only a constant fraction larger than the smallest. This can be relaxed albeit at the cost of a worse rate. Our proofs explicitly maintain the dependence on η . For the ideal matrix, the Assumption 3 ensures that at every level of the hierarchy, the gap between the within-cluster similarities and between-cluster similarities is larger than the range of between-cluster similarities. Earlier papers of McSherry [138], Rohe et al. [163] assume that the ideal similarities are constant within a block in which case the assumption is trivially satisfied by the definition of the ideal matrix. However, more generally this assumption is necessary to show that the entries of the eigenvector are safely bounded away from zero. If this assumption is violated by the ideal matrix, then the eigenvector entries can decay as fast as $O(1/n)$ (see Section

4.6 for more details), and our analysis shows that such matrices will no longer be as robust to noise.

Other analyses of spectral clustering often directly make less interpretable assumptions about the spectrum. For instance, Ng et al. [145] assume conditions on the eigengap of the normalized Laplacian and this assumption implicitly creates constraints on the entries of the ideal matrix A that can be hard to make explicit.

To state our theorems concisely we will define an additional quantity γ_S^* . Intuitively, γ_S^* quantifies how close the ideal matrix comes to violating Assumption 3 over a set of clusters S .

Definition 4. For a set of clusters S , define

$$\gamma_S^* \triangleq \min_{s \in S} \min \{ \alpha_{s \cdot L}, \alpha_{s \cdot R} \} - \beta_s - \eta(\beta_s - \alpha_s).$$

We, as well as previous works of Ng et al. [145], Rohe et al. [163], rely on results from perturbation theory to bound the error in the observed eigenvectors in 2-norm. Using this approach, the straightforward way to analyze the number of errors is pessimistic since it assumes the difference between the two eigenvectors is concentrated on a few entries. However since the perturbation is in fact generated by a random process it is unlikely to be adversarially concentrated. We formalize this intuition to *uniformly* bound the perturbations on every entry and get a stronger guarantee.

Our main result for hierarchical spectral clustering gives conditions on the noise scale factor σ under which Algorithm HS will recover all clusters $s \in \mathcal{S}_m$, where \mathcal{S}_m is the set of all clusters of size at least m .

Theorem 4.3.1. Suppose that $W = A + R$ is an $(n \times n)$ noisy HBM where A satisfies Assumptions 1, 2, and 3. Suppose that the scale factor of R increases at

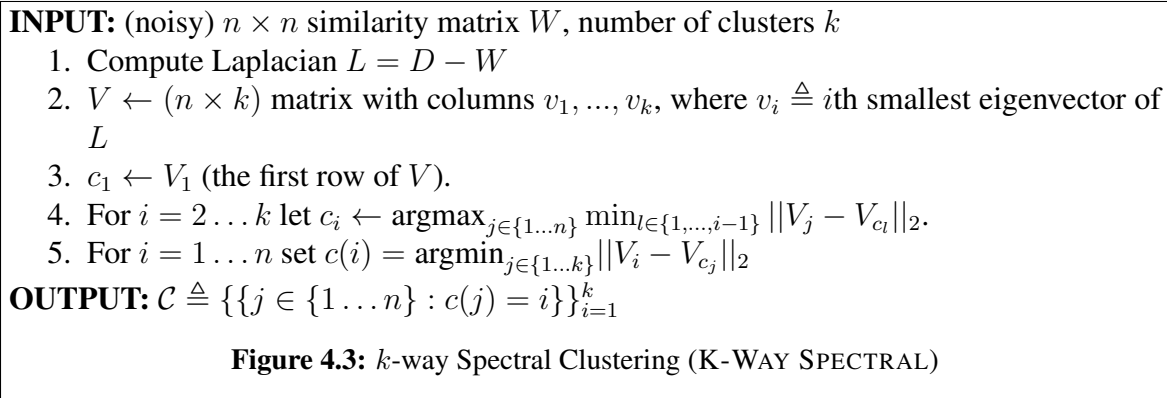
$$\sigma = o \left(\min \left(\kappa^{*5} \sqrt{\frac{m}{\log n}}, \kappa^{*4} \sqrt[4]{\frac{m}{\log n}} \right) \right)$$

where $\kappa^* = \min \left(\alpha_0, \frac{\gamma_{\mathcal{S}_m}^*}{1+\eta} \right)$, $m > 0$ ¹. Then for all n large enough, with probability at least $1 - 6/n$, Algorithm HS, on input M , will exactly recover all clusters of size at least m .

A few remarks are in order:

1. It is impossible to resolve the entire hierarchy, since small clusters can be irrecoverably buried in noise. The amount of noise that algorithm HS can tolerate is directly dependent on the size of the smallest cluster we want to resolve.
2. It is easy to see that in resolving only the first level of the hierarchy, the amount of noise Algorithm HS can tolerate is (pessimistically) $o(\kappa^{*5} \sqrt[4]{n/\log n})$ which grows rapidly with the number of objects to be clustered n .
3. Under this scaling between n and σ , it can be shown that popular agglomerative algorithms such as single linkage will fail with high probability. We verify this negative result through experiments (see Section 4.6).

¹Recall $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$



4. In the noiseless case, when the similarities are constant in each block, the higher eigenvectors correspond to the Haar wavelets and reveal the structure of the hierarchy. Similar results hold for the ideal HBM matrices we use. Rather than using all the eigenvectors of the original matrix we recursively use the *second* eigenvector the submatrices of W . Intuitively, an algorithm that uses all of the eigenvectors of W , rather than the eigenvectors of the submatrices performs poorly because eigenvectors of W are affected by noise on the *entire* matrix, rather than the noise on just the particular submatrix of W under consideration.
5. Since we assume that β^* does not grow with n , both the range $(\beta_s - \alpha_s)$ and the gap $(\min\{\alpha_{s,L}, \alpha_{s,R}\} - \beta_s)$ must decrease with n and hence that $\gamma_{\mathcal{S}_m}^*$ must decrease as well. For example, if we have uniform ranges and gaps across all levels, then $\gamma_{\mathcal{S}_m}^* = \Theta(1/\log n)$. For constant α_0 , for n large enough $\kappa^* = \frac{\gamma_{\mathcal{S}_m}^*}{1+\eta}$. We see that in our analysis $\gamma_{\mathcal{S}_m}^*$ is a crucial determinant of the noise tolerance of spectral clustering.

4.4 K-way Clustering

In the k -way case, we consider the following similarity matrix which is studied by Ng et al. [145].

Definition 5. W is a *noisy k-Block Diagonal* matrix if $W \triangleq A + R$ where R is a perturbation matrix and A is an ideal matrix for the k -way problem. An ideal matrix for the k -way problem has within-cluster similarities larger than $\beta_0 > 0$ and between cluster similarities 0.

We extend the intuition behind Theorem 4.3.1 to the k -way setting. Some arguments are considerably more subtle since spectral clustering uses the *subspace* spanned by the k smallest eigenvectors of the Laplacian. Our results improve those of Ng et al. [145] to provide a coordinate-wise bound on the perturbation of the subspace, and use this to make precise guarantees for Algorithm K-WAY SPECTRAL, which includes an iteration of k -means style algorithm for cluster assignment.

Theorem 4.4.1. Suppose that $W = A + R$ is an $(n \times n)$ *noisy k-Block Diagonal* matrix where

A satisfies Assumptions 1 and 2. Suppose that the scale factor of R increases at rate

$$\sigma = o\left(\frac{\beta_0}{k} \left(\frac{n}{k \log n}\right)^{1/4}\right)$$

then with probability $1 - 8/n$, for all n large enough, K-WAY SPECTRAL will exactly recover the k clusters.

Notice that we assume that in the *ideal* case between cluster similarities are exactly 0. In current work we are investigating extending these results to the case when the *ideal* between cluster similarities are strictly smaller than the within cluster similarities (but not necessarily 0). The extension is however substantially more involved.

4.5 Minimax Rates

Theorems 4.3.1 and 4.4.1 show that the spectral clustering can tolerate a high amount of noise while still recovering the clusters. This guarantee leaves open the question of optimality:

Are the spectral algorithms optimal in their dependence on the various parameters of the problem?

In this section we present tight minimax upper and lower bounds for the hierarchical and k -way problems under the assumptions of known cluster sizes and block-constant activations. The modification to non block-constant activations is straightforward. Indeed, in a precise sense the constant similarities considered in this section can be seen as the “worst case”, and the performance of the combinatorial procedures we consider only improves in the non block-constant case. Adapting to unknown cluster sizes is more involved and beyond the scope of this chapter.

For the hierarchical problem, we establish the minimax rate in a simplification of the noisy HBM where $\epsilon_s \triangleq \alpha_s = \beta_s$ for all s :

$$\gamma \triangleq \min_{s \in \mathcal{S}} \min\{\epsilon_{s \cdot L}, \epsilon_{s \cdot R}\} - \epsilon_s$$

quantifies the gap between inter and intra-cluster similarities across all of the clusters. We will also assume that the matrix is perfectly balanced ($\eta = 1$). Our first minimax result is a lower bound which establishes a condition on (n, σ, γ) under which *any* method will make an error in identifying the correct clusters.

Theorem 4.5.1. *There exists a constant $\alpha \in (0, 1/8)$ such that if*

$$\sigma \geq \gamma \sqrt{\frac{2m}{\alpha \log nm}}$$

the probability that any estimator fails to recover all clusters of size $\geq m$ remains bounded away from 0 as $n \rightarrow \infty$.

Under the conditions of this theorem, γ and κ^* from Theorem 4.3.1 coincide, provided the inter-cluster similarities remain bounded away from 0 by at least a constant. Theorem 4.3.1 then implies that spectral clustering requires

$$\sigma \leq \min \left(\gamma^5 \sqrt{\frac{n}{C \log n}}, \gamma^4 \sqrt[4]{\frac{n}{C \log n}} \right)$$

for a large enough constant C . The noise threshold for spectral clustering does not match the lower bound. To establish that Theorem 4.5.1 is indeed the minimax rate, we need to demonstrate a procedure (that is not necessarily computationally efficient) that matches it. For this, we analyze a combinatorial procedure that solves the NP-hard problem of finding the minimum cut of size exactly $n/2$ by searching over all subsets. A recursive application of this algorithm can be used for hierarchical clustering.

More formally, for any subcluster C_s , denote the submatrix corresponding to C_s by W^s . For a given index set I_s define:

$$S(W^s, I_s) = \sum_{i \in I, j \in I} W_{ij}^s + \sum_{i \in I^c, j \in I^c} W_{ij}^s - \sum_{i \in I, j \in I^c} W_{ij}^s - \sum_{i \in I^c, j \in I} W_{ij}^s.$$

At each subcluster C_s , our algorithm exactly minimizes $S(W^s, I_s)$ subject to $|I_s| = |C_s|/2$.

This algorithm is strongly related to spectral clustering with the combinatorial Laplacian, which solves a *relaxation* of the balanced minimum cut problem.

Theorem 4.5.2. *There exists a constant C such that if*

$$\sigma < \gamma \sqrt{\frac{m}{C \log n}}$$

the combinatorial procedure described above succeeds with probability at least $1 - \frac{1}{n}$ which goes to 1 as $n \rightarrow \infty$.

This theorem and the lower bound together establish the minimax rate. It however, remains an open problem to tighten the analysis of spectral clustering in this chapter to match this rate.

In the k -way setting, the lower bound follows a similar proof to the hierarchical case. For the upper bound we use a combinatorial procedure that finds and removes *one cluster* at a time. The algorithm will find a set of $m \triangleq n/k$ objects that maximizes the difference between within-cluster and between cluster similarity:

$$\hat{I} = \operatorname{argmax}_{I \subset [n], |I|=m} S(W, I)$$

where

$$S(W, I) \triangleq \sum_{i, j \in I} W_{ij} + \sum_{i, j \notin I} W_{ij} - \sum_{i \in I, j \notin I} W_{ij} - \sum_{i \notin I, j \in I} W_{ij}.$$

The search is repeated $k - 1$ times (each time removing the indices in \hat{I}) to find the k clusters.

INPUT: Noisy similarity matrix W , number of clusters k .

1. Randomly divide the columns of W into two parts W_1 and W_2 of size $n/2$ and define $P_{W_1} = Q_{W_1}Q_{W_1}^T, P_{W_2} = Q_{W_2}Q_{W_2}^T$, where Q_{W_1} are the top k left singular vectors of W_1 and Q_{W_2} are the top k left singular vectors of W_2 .
2. Compute $\widehat{W} = [P_{W_2}W_1|P_{W_1}W_2]$.
3. Run the version of k -means above directly on the columns of \widehat{W} to recover the k clusters.

OUTPUT: \mathcal{C}

Figure 4.4: A variant of the algorithm from the paper of McSherry [138]

Theorem 4.5.3. *The minimax rate for k -way clustering is*

$$\sigma \asymp \gamma \sqrt{\frac{n}{k \log(n/k)}}.$$

Under the added restriction of block constant ideal similarities, the analysis of McSherry [138] yields an efficient algorithm (in Figure 4.4) that achieves the minimax dependence between n, γ and σ for the k -way problem. Note that the algorithm is not minimax optimal in terms of k , which we assume here is a constant. The algorithm and its analysis rely crucially on the fact that the noiseless matrix is rank k ; consequently our result does not directly extend to the non-constant block similarities analyzed in Theorem 4.3.1. In this sense our analysis of spectral clustering and the combinatorial procedures are much more general.

The analysis of this algorithm closely follows that of McSherry [138]. As in that proof, we will analyze the algorithm under the assumption that each of the k clusters is exactly bisected in W_1 and W_2 . It is straightforward to show that each cluster is approximately bisected with high probability for k constant and large n . Although it is possible to modify the analysis for the more realistic approximate bisection case (see McSherry [138] for a discussion), the assumption that the clusters are exactly bisected eases the analysis considerably.

Theorem 4.5.4. *If:*

$$\gamma \geq C_1 \sigma k \sqrt{\frac{k}{n}} + c_2 \sigma k \sqrt{\frac{\log(n/\delta)}{n}}$$

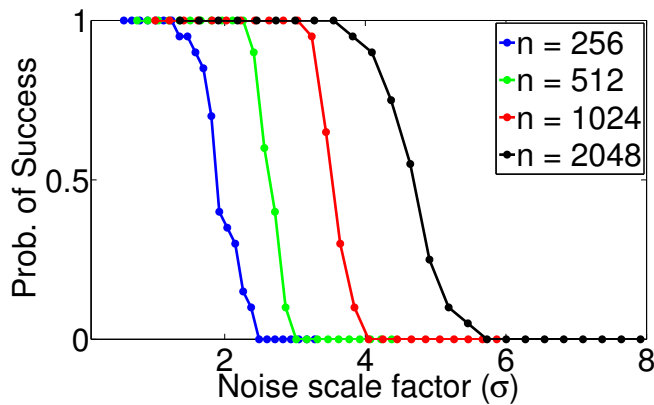
then with probability at least $1 - \delta$ the algorithm succeeds in recovering the k clusters.

This rate is optimal except in its dependence on k . For constant k , the second term dominates and we recover the minimax rate, i.e.

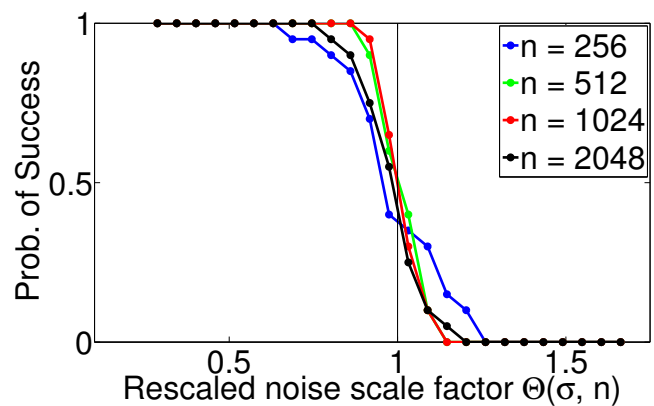
$$\gamma \geq C \sigma \sqrt{\frac{\log(n/\delta)}{n}}$$

suffices to recover the clusters.

To summarize, in this section we have given information theoretic lower bounds for the hierarchical and k -way problems and analyzed combinatorial procedures that achieve these lower



(a)



(b)

Figure 4.5: Threshold curves for the recovery of one split using HS

bounds. For the special case of constant block similarities in the k -way case we have analyzed a computationally efficient algorithm that achieves the lower bound.

4.6 Experimental Results

We evaluate our algorithms and theoretical guarantees on simulated matrices, synthetic phylogenies, and finally on two real biological datasets. Our experiments focus on the effect of noise on spectral clustering in comparison with agglomerative methods such as single, average, and complete linkage.

4.6.1 Noise Thresholds and Asymptotic Behavior

One of our primary interests is to empirically validate the relation between the scale factor σ and the sample size n derived in our theorems. For a range of scale factors and noisy HBMs of varying size, we empirically compute the probability with which spectral clustering recovers the first split of the hierarchy. From the probability of success curves (Figure 4.5(a)), we can conclude that spectral clustering can tolerate noise that grows with the size of the clusters.

We further verify the dependence between σ and n for recovering the first split. We observe that when we rescale the x-axis of the curves in Figure 4.5(a) by $\sqrt{\log(n)/n}$ the curves line up for different n (Figure 4.5(b)). This shows that empirically, at least for the first split, spectral clustering appears to achieve the minimax rate for the problem. It is an important open question to show that, at least for this special case, spectral clustering with the combinatorial Laplacian achieves the minimax rate.

4.6.2 Examples of Worst Case Behavior

Here we demonstrate the undesirable spectral properties of both the combinatorial and normalized laplacians, in addition to the adjacency matrix. We use concrete examples of similarity matrices whose second eigenvector does not immediately produce the correct clustering. Additionally, we motivate our Range Restriction, by showing that if this condition is not satisfied, the entries of the eigenvector decay at $O(\frac{1}{n})$ instead of $O(\sqrt{\frac{1}{n}})$.

First, we turn to the drawbacks of using the spectrum of the adjacency matrix. McSherry [138] shows that in the planted partition model, the eigenvectors of the adjacency matrix are enough to identify the clusters. However, in the more general HBM, this is not the case. Consider a matrix with small off-diagonal entries, larger entries on the diagonal blocks, and 2 very high entries in this block (See Figure 4.6(a)). This is an ideal matrix and the second eigenvector of the combinatorial Laplacian exactly identifies the true clustering, yet the eigenvector of the adjacency matrix fails to convey any meaningful information (See Figure 4.6(e)).

The normalized Laplacian can also fail to identify the clusters of an ideal hierarchical matrix. For example, on a similarity matrix like the one in Figure 4.6(b), the second eigenvector of the normalized laplacian identifies the clustering at the second level of the hierarchy rather than the first, as shown in Figure 4.6(f). We conjecture that different conditions will guarantee that correctness of a spectral method using the normalized laplacian, but we instead focus on the combinatorial Laplacian and our definition of ideal matrices.

The combinatorial Laplacian also has its shortcomings, most notably that it is highly influenced by outliers in the data. If even one data point disrupts the structure of the matrix, as in Figure 4.6(c), the second eigenvector of the combinatorial Laplacian becomes highly spiked and it can no longer tolerate even small perturbations (see Figure 4.6(g)).

A related example demonstrates the necessity of the Assumption 3. Consider the matrix shown in Figure 4.6(d), which is an ideal matrix that violates the range restriction. In this case, the

eigenvector again becomes highly spiked (Figure 4.6(h)), and moreover, the entries decay at a rate of $O(1/n)$ (not shown), which is too sharp for our results to hold.

4.6.3 Real World Experiments

We apply hierarchical clustering methods to a yeast gene expression data set and one phylogenetic data set from the PFAM database [74]. To evaluate our methods, following Eriksson et al. [69], we use a Δ -entropy metric defined as follows: Given a permutation π and a similarity matrix W , we compute the rate of decay off of the diagonal as

$$\hat{s}_d \triangleq \frac{1}{n-d} \sum_{i=1}^{n-d} W_{\pi(i), \pi(i+d)}$$

for $d \in \{1, \dots, n-1\}$. Next, we compute the entropy

$$\hat{E}(\pi) \triangleq - \sum_{i=1}^{n-1} \hat{p}_\pi(i) \log \hat{p}_\pi(i)$$

where

$$\hat{p}_\pi(i) \triangleq \left(\sum_{d=1}^n \hat{s}_d \right)^{-1} \hat{s}_i.$$

Finally, we compute Δ -entropy as

$$\hat{E}_\Delta(\pi) = \hat{E}(\pi_{random}) - \hat{E}(\pi).$$

A good clustering will have a large amount of the probability mass concentrated at a few of the $\hat{p}_\pi(i)$ s, thus yielding a high $\hat{E}_\Delta(\pi)$. On the other hand, poor clusterings will specify a more uniform distribution and will have lower Δ -entropy.

We first compare HS to single linkage on yeast gene expression data from DeRisi et al [62]. This dataset consists of 7 expression profiles, which we use to generate Pearson correlations that we use as similarities. We sampled gene subsets of size $n = 512, 1024, \text{ and } 2048$ and ran both algorithms on the reduced similarity matrix. We report Δ -entropy scores in Table 4.7(b). These scores quantitatively demonstrate that HS outperforms single linkage and additionally, we believe the clustering produced by HS (Figure 4.7(a)) is qualitatively better than that of single linkage.

Finally, we run HS on real phylogeny data, specifically, a subset of the PDZ domain (PFAM Id: PF00595). We consider this family because it is a highly-studied domain of evolutionarily well-represented protein binding motifs. Using alignments of varying length, we generated similarity matrices and computed Δ -entropy of clusterings produced by both HS and Single Linkage. The results for three sequence lengths (Table 4.7(b)) show that HS and Single Linkage are comparable.

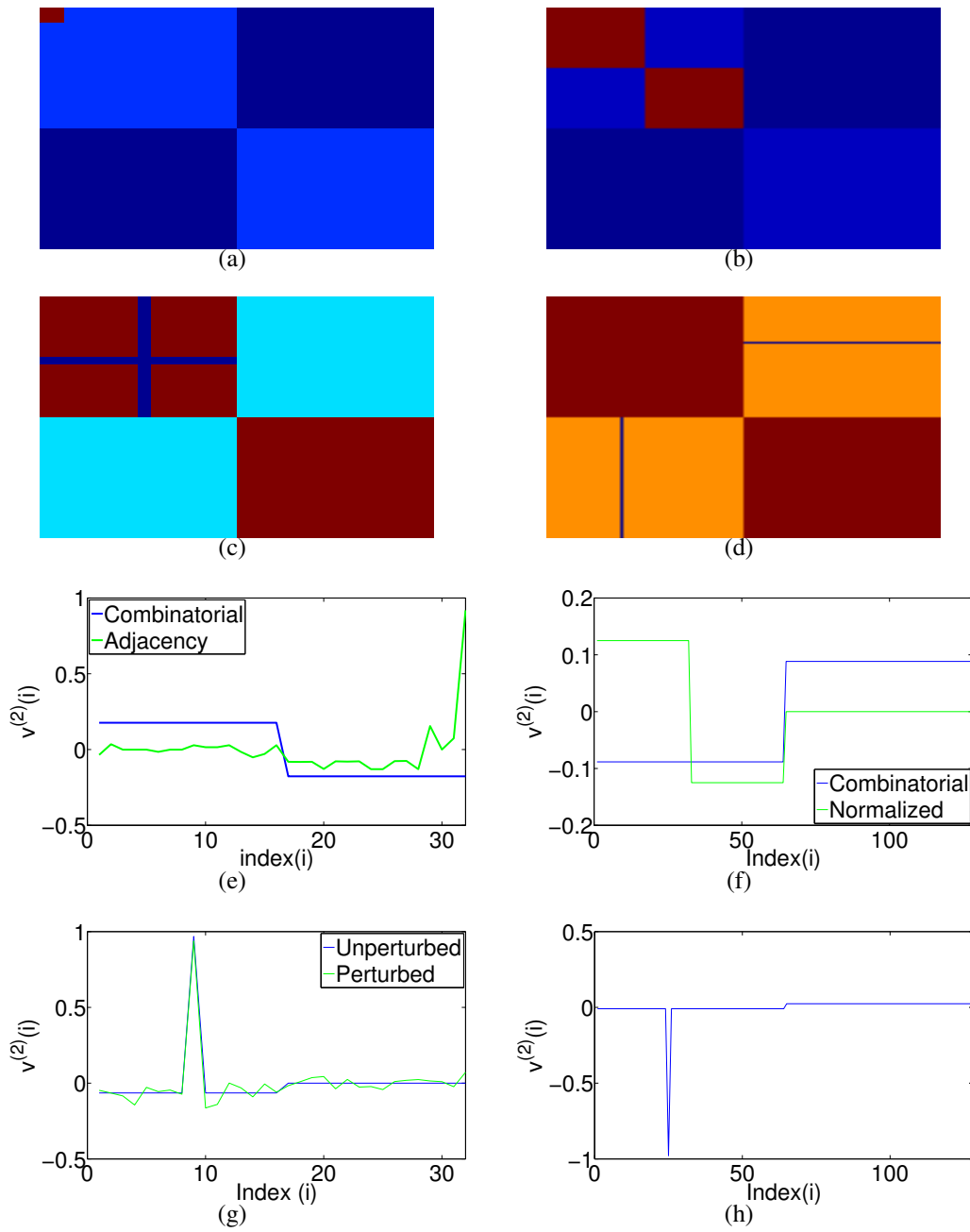
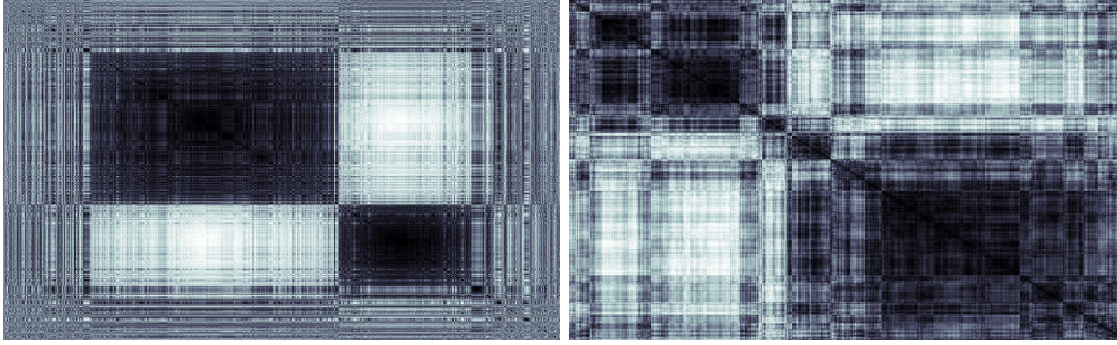


Figure 4.6: Example similarity matrices, red entries are high and blue are low, that result in undesirable behavior for Normalized Laplacians and Adjacency Matrices and Combinatorial Laplacians.



(a)

Dataset	HS	Agglomerative
Gene (n = 2048)	0.0775	0.0203
Gene (n = 1024)	0.1006	0.0312
Gene (n = 512)	0.0785	0.0280
Phylogeny (l = 100)	0.0067	0.0063
Phylogeny (l = 200)	0.0066	0.0069
Phylogeny (l = 300)	0.0066	0.0060

(b)

Figure 4.7: Experiments with real world data. (a): Heatmaps of single linkage (left) and HS (right) on gene expression data with $n = 2048$. (b) Δ -entropy scores on real world data sets.

4.7 Proofs

In this section we outline proof sketches for Theorem 4.3.1 as well as Theorems 4.5.1 and 4.5.2. Detailed proofs of these and the other theorems appear in Section 4.9.

4.7.1 Proof of Theorem 4.3.1

The analysis of the hierarchical spectral algorithm can be compartmentalized into several sections that we outline here:

1. Noiseless Spectral Clustering: We show that Algorithm HS will perfectly cluster a noiseless Hierarchical Block Matrix (HBM).
2. Derive spectral properties of noiseless matrices: We study the spectral properties of a related matrix, the Constant Block Matrix (CBM), and use it to understand the spectral properties of the HBM. This analysis is entirely deterministic.
3. Bound spectral norm of noise matrices: We analyze the noise matrices and show that, with high probability, they have small spectral norm uniformly across all levels of the hierarchy.

4. Davis-Kahan for Laplacians: We next use a variant of the well-known Davis-Kahan $\sin \theta$ theorem to bound the ℓ_2 -norm deviation between the eigenvectors of the HBM and the noisy HBM in terms of the spectral norm of the noise matrices.
5. ℓ_∞ -norm deviation bounds: We observe that due to the independence and randomness of the noise, it is unlikely that the perturbation of the eigenvector of the noisy HBM is concentrated in just a few coordinates. We formalize this notion by deriving ℓ_∞ -norm deviation bounds between the eigenvectors of the HBM and the noisy HBM.
6. Final steps: we conclude that for sufficiently large n , every entry of the second eigenvectors (across all calls to Algorithm HS) correctly clusters the data.

Before diving into the proof, let us build some intuition with some simplified heuristic calculations, focusing on recovering just the first split. Let $W = A + R$ be the $n \times n$ noisy HBM. One can readily verify that the Laplacian of W , L_W , can be decomposed as $L_A + L_R$. Let $v^{(2)}, u^{(2)}$ be the second eigenvectors of L_A, L_W respectively.

We first show that the unperturbed eigenvector, $v^{(2)}$, clearly distinguishes the two outermost clusters. Specifically we show that $|v_i^{(2)}| = \Theta\left(\frac{1}{\sqrt{n}}\right)$ for all coordinates i and that its sign corresponds to the cluster identity of point i . We also establish that the eigengaps $\lambda_2 - \lambda_1$ and $\lambda_3 - \lambda_2$ are both $\Theta(n)$.

In step three of the proof, we show that $\|L_R\| \leq O(\sigma\sqrt{n\log n})$ with high probability. Equipped with the previous results, in step four of the proof, we apply the well-known Davis-Kahan perturbation theorem to show that:

$$\|v^{(2)} - u^{(2)}\|_2 = O\left(\sigma \frac{\sqrt{n\log n}}{\min(\lambda_2 - \lambda_1, \lambda_3 - \lambda_2)}\right) = O\left(\sigma \sqrt{\frac{\log n}{n}}\right).$$

At this point, we can already make a guarantee on the performance of spectral clustering. Since we argued that $|v_i^{(2)}| = \Theta\left(\sqrt{\frac{1}{n}}\right)$, if $\|v^{(2)} - u^{(2)}\|_\infty = o\left(\sqrt{\frac{1}{n}}\right)$ then we know that for n large enough, the spectral algorithm will correctly partition the data. Since the ℓ_∞ norm is bounded by the ℓ_2 norm we now know that if $\sigma = o\left(\sqrt{\frac{1}{\log n}}\right)$ then our algorithm will succeed.

The above argument is pessimistic in that it assumes the perturbation will be concentrated on a few entries of the eigenvector (this is when the ℓ_∞ norm is close to the ℓ_2 norm and the bound is tight). Consequently it leads to a poor performance guarantee. Instead, and in step five of our proof, we perform a much more careful analysis to show that all coordinates uniformly have low perturbation, obtaining a much tighter bound on $\|v^{(2)} - u^{(2)}\|_\infty$.

In what follows, we make the arguments across all levels of the hierarchy simultaneously. In step six, we put all of the pieces together and arrive at Theorem 4.3.1. We defer the proofs of all technical lemmas to Section 4.9.

Algorithm HS in the noiseless setting

We now show that in the absence of noise, Algorithm HS will perfectly cluster the data W .

Lemma 4.7.1. *Given an ideal noiseless Hierarchical Block Matrix W (i.e. $R = 0$) satisfying Assumption 1, HS will recover the true hierarchical clustering.*

The result shows that at all levels of the hierarchy, the sign pattern of the second eigenvectors correspond to the cluster memberships. The proof of this lemma can be found in Section 4.9.1.

Note that this lemma would not hold if Algorithm HS used either the normalized Laplacian or the similarity matrix directly. In fact, in Section 4.6, we show several examples that demonstrate the shortcomings of these approaches. In addition, note that we do not require Assumptions 2 and 3 for Lemma 4.7.1.

The fact that the second eigenvectors have the correct sign pattern does not ensure robustness to noise. To ensure robustness, we would like to verify that the coordinates of the second eigenvector are bounded away from 0. To apply results from perturbation theory, it is also essential to establish bounds on the first, second and third eigenvalues of the Laplacian.

Spectrum of HBMs

Step 2 of our proof outline requires us to characterize the spectrum of (noiseless) hierarchical block matrices. We do so in the following lemma.

Lemma 4.7.2. *(Spectrum of HBMs) Consider an $(n \times n)$ ideal Hierarchical Block Matrix*

$$A = \begin{pmatrix} A_L & A_S \\ A_S^\top & A'_L \end{pmatrix}$$

such that all values in off-diagonal blocks A_S are in $[\alpha_0, \beta_0]$ and all values in the diagonal blocks A_L, A'_L are in $[\alpha_1, \beta^]$ (here we take $\alpha_1 = \min\{\alpha_L, \alpha_R\}$).*

Suppose A satisfies Assumptions 1 and 2 with balance factor η . Suppose also that A satisfies Assumption 3. Then:

1. *Let $\lambda_1, \lambda_2, \lambda_3$ be the first, second and third smallest eigenvalue of L_A respectively, then the eigengap*

$$\delta \triangleq \min(|\lambda_2 - \lambda_1|, |\lambda_3 - \lambda_2|) \geq \min\left(n\alpha_0, \frac{n}{\eta+1}(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0)\right) = \Theta(n).$$

2. *Let $v^{(2)}$ be the second eigenvector of L_A , then every entry of $v^{(2)}$ satisfies*

$$\sqrt{\frac{1}{K_\eta n}} \leq |v^{(2)}(i)| \leq \sqrt{\frac{K_\eta}{n}}$$

where

$$K_\eta = \left(\frac{(\beta^* - \alpha_0)\beta_0 - \alpha_0 + \eta(\beta^* - \alpha_0)}{(\alpha_1 - \beta_0)\alpha_1 - \beta_0 - \eta(\beta_0 - \alpha_0)} \right)^2.$$

The proof of the lemma can be found in Section 4.9.2. In the proof, we first derive analogous spectral properties for a simplified family of matrices, that have $\alpha_s = \beta_s$ for all clusters. Then we use results from spectral graph theory to sandwich the eigenvalues of the HBM between the eigenvalues of two of these simpler matrices. Leveraging the eigenvector definitions we are able to similarly sandwich the entries of the eigenvectors.

Note that once we prove this Lemma, we can recursively apply it on sub-matrices that represent the similarity matrix of sub-clusters to characterize the eigenvectors and eigenvalues at every split of the hierarchical clustering. One complication with recursively applying Lemma 4.7.2 is that at different level i , we would get a different K_η . To succinctly present the final rates, we define K_η^* as the maximum over all K_η for all levels i :

$$K_\eta^* = \max_{s \in \mathcal{S}_m} \left(\frac{(\beta^* - \alpha_s)}{(\min\{\alpha_{s,L}, \alpha_{s,R}\} - \beta_s)} \frac{\beta_s - \alpha_s + \eta(\beta^* - \alpha_s)}{\min\{\alpha_{s,L}, \alpha_{s,R}\} - \beta_s - \eta(\beta_s - \alpha_s)} \right)^2$$

where β^* is the largest entry in the entire ideal HBM A . We must characterize the dependence of K_η^* on κ^* . Note in the expression for K_η^* that

$$\min\{\alpha_{s,L}, \alpha_{s,R}\} - \beta_s \geq \gamma^*$$

and that the terms in the numerator are all bounded by a constant depending on η and β^* which is the bound on the entries of the similarity matrix. Thus, we get

$$K_\eta^* \leq \frac{C_{\eta, \beta^*}}{\gamma^{*4}} \leq \frac{C_{\eta, \beta^*}}{\kappa^{*4}}.$$

Bounds on the noise

We now analyze the noise matrices. The main lemma that we will leverage repeatedly in our analysis bounds the spectral norm of the noise component of each Laplacian in the hierarchy.

Lemma 4.7.3. (*Hierarchical Laplacian Operator Norm Bound*) *Let R be the noise matrix associated with an $n \times n$ noisy Hierarchical Block Matrix satisfying Assumptions 1 and 2.*

Then with probability $1 - 4/n$, for all sub-clusters C_{li} , the corresponding noise Laplacian matrix $L_R^{C_{li}}$ will have operator norm bounded by

$$\left\| L_R^{C_{li}} \right\|_2 \leq C(\eta) \sigma \sqrt{m_{li} \log n}$$

for a constant C depending on η .

We stress that at this point, we have dealt with all of the randomness involved in recovering the clusters, across all levels. Specifically, we now know that with probability at least $1 - 4/n$, every noise Laplacian of size m_{li} will have spectral norm bounded by $O(\sigma \sqrt{m_{li} \log n})$.

Davis-Kahan for Laplacians and ℓ_2 Deviation Bounds

We now derive some results related to perturbation theory that will be useful in our final proof. The first is a variant of the Davis-Kahan theorem that bounds the eigenvector deviation in ℓ_2 -norm.

Lemma 4.7.4. (*Davis-Kahan*) *With probability at least $1 - \delta$,*

$$\|u^{(i)} - v^{(i)}\|_2 \leq \frac{\sqrt{2} \|L_R\|}{\xi_i}$$

where ξ_i denotes the eigengap for the i^{th} eigenvalue of L_A , i.e. $\xi_i = \min_{i \neq j} |\lambda_i - \lambda_j|$.

Before we proceed, we remark here that Lemma 4.7.4, combined with Lemma 4.7.3, immediately gives us an ℓ_2 deviation between the eigenvectors of the noisy HBM and the ideal HBM. Specifically, if we additionally use Lemma 4.7.2 to lower bound ξ_i , we see that for the cluster C_{li} :

$$\|u^{(2)} - v^{(2)}\|_2 = O\left(\sigma \sqrt{\frac{\log n}{m_{li}}}\right).$$

Using the uniform spectral bounds in Lemma 4.7.3, we arrive at this ℓ_2 -norm deviation bound for all clusters of size at least m_{li} with probability $1 - 4/n$.

Uniform bounds on $u^{(2)} - v^{(2)}$

Note that the above result is not sufficient to guarantee that spectral clustering will make no mistakes as $u^{(2)}$ could be spiked (and have flipped signs) even if it is close to $v^{(2)}$ in ℓ_2 . To make this guarantee, we perform a more careful analysis and show that $u^{(2)}$ is uniformly close to $v^{(2)}$ in every coordinate.

In this analysis, let us focus on a cluster C_s of size m_s . For ease of notation, we will denote the adjacency matrix of C_s by A and the perturbation of C_s by R . We will further use D_{Ai} and D_{Ri} to denote the sum of the i^{th} row of A and R respectively. Repeated application for all of the clusters, using the fact that all of the noise laplacians can be bounded, will guarantee the correctness of our algorithm across all clusters. Let $k = u^{(2)} - v^{(2)}$. The following lemma shows that with high probability, $k(i)$, the element-wise perturbation is uniformly low.

Lemma 4.7.5. *With the above definitions, we have:*

$$k(i) = \frac{1}{c_i} (v^{(2)}(i)(\lambda_2 - \mu_2) - A_i k + L_{Ri} v^{(2)} - R_i k)$$

where $c_i = \mu_2 - D_{Ai} - D_{Ri}$. Moreover, if $\sigma = o\left(\gamma \sqrt{\frac{m_s}{\log n}}\right)$, then with probability $\geq 1 - 6/n$:

$$\|k\|_\infty \leq \frac{2\sigma\sqrt{\log n}}{m_s \kappa^*} \left[s\sqrt{6K_\eta^*} + \frac{4\sqrt{3}\beta^*}{\kappa^*} + 4\sqrt{3K_\eta^*} + \frac{4C\sigma\sqrt{3}}{\kappa^*} \right].$$

The expression for $k(i)$ comes from algebraic manipulation of the eigenvector equations. The bound on $\|k\|_\infty$ involves analyzing each term in the expression for $k(i)$, using the properties we have previously derived.

Putting Everything Together To arrive at our final rate notice that if

$$\sigma = o\left(\min\left(\kappa^{*5} \sqrt{\frac{m_s}{\log n}}, \kappa^{*4} \sqrt{\frac{m_s}{\log n}}\right)\right)$$

then for large enough n we have $\|k\|_\infty \leq \sqrt{\frac{1}{K_\eta m_s}}$ and our algorithm makes no mistakes in resolving all clusters of size at least m_s .

4.7.2 Proof of Theorem 4.5.1

The proof of our lower bound is an application of the following form of Fano's Inequality from the book of Tsybakov [192]:

Lemma 4.7.6 (Theorem 2.5 of [192]). *Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \dots, \theta_M$ such that:*

1. $d(\theta_j, \theta_k) \geq 2s > 0, \forall 0 \leq j < k \leq M$.
2. $P_j \ll P_0, \forall j = 1, \dots, M$, and

$$\frac{1}{M} \sum_{j=1}^M K(P_j, P_0) \leq \alpha \log M$$

with $0 < \alpha < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \dots, M$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_{\hat{\theta}}(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

K denotes the KL-divergence and d is an arbitrary semi-metric.

We use this lemma with d being the Hamming distance. We choose the parameter family Θ as follows. First suppose that $n = 2^\alpha$ and $m = 2^\beta$ for integers $\alpha > \beta$. Let θ_0 be some hierarchical partitioning of $[n]$ into clusters of size m . Define θ_{jk}^s for $s \in [n/(2m)]$ that swaps that j^{th} element from the left child of cluster s with the k^{th} element from the right child of cluster s , where s can be any of the second level clusters in the hierarchy.

In total there are $n/(2m) \times m^2$ models as there are $n/(2m)$ different second level clusters, and for each there are m choices for the element in the left cluster and m choices for the element in the right cluster. Notice that each of these models are at a Hamming distance of *at least* one from the true partitioning θ_0 .

Since we are interested in the worst case we can make two further simplifications. The ideal (noiseless) matrix can be taken to be block-constant with gap γ , since the worst case is when the

diagonal blocks are at their lower bound and the off-diagonal blocks are at their upper bound. Further, we can consider matrices $W = A + R$, which are $(n \times n)$ matrices, with $R_{ij} \sim \mathcal{N}(0, \sigma^2)$. We also need to calculate the KL-divergence between the probability distributions induced by each of the parameters.

Each distribution is Gaussian and in comparison with θ_0 the mean of each distribution differs in at most $4m$ coordinates (inter- and intra- cluster similarities for the two objects that were swapped). In this coordinates, it differs by γ . Thus

$$K(P_{\theta_{jk}^s}, P_{\theta_0}) = \frac{2m\gamma^2}{\sigma^2}.$$

To arrive at Theorem 4.5.1 we simply apply these calculations in Lemma 4.7.6.

Notice that for the k -way problem we can use a similar construction. If there are k clusters, each of size n/k , then we define the family of models as follows. Starting with a base clustering θ_0 group the clusters into pairs, then define θ_{jk}^s as swapping the j^{th} element and the k^{th} element between the pair s of clusters. A similar computation shows that there are $k/2(n/k)^2$ such models. Moreover

$$K(P_{\theta_{jk}^s}, P_{\theta_0}) = \frac{2n\gamma^2}{k\sigma^2}.$$

Applying Lemma 4.7.6 yields the lower bound in Theorem 4.5.3.

4.7.3 Proof of Theorem 4.5.2

For the hierarchical upper bound, our algorithm recursively solves the balanced minimum cut problem. To analyze the algorithm described in Section 4.5

To analyze the procedure, it's useful to consider the random variable ζ^s defined as:

$$\zeta_{\hat{I}_s}^s = S(W^s, I_s) - S(W^s, \hat{I}_s).$$

Further, given a set \hat{I}_s , define the number of indices in which \hat{I}_s and I_s agree to be a_s .

It's not too hard to see that for a given a_s ,

$$\zeta_{\hat{I}_s}^s \sim \mathcal{N}\left(8a_s \left(\frac{|C_s|}{2} - a_s\right) \gamma, 16a_s \left(\frac{|C_s|}{2} - a_s\right) \sigma^2\right).$$

At any cluster C_s , the combinatorial procedure succeeds if $\zeta_{\hat{I}_s}^s > 0$ for all $\hat{I}_s \neq I_s$. A fairly simple application of the union bound shows that the probability of error of the entire combinatorial procedure is bounded by the probability of error across each of the clusters C_s . This probability (by application of the union bound) can be bounded as follows:

$$\begin{aligned} \mathbb{P}_{\text{error}} &\leq \sum_{i=1}^l 2^i \sum_{a=1}^{n/2^i-1} \binom{n/2^i}{a} \binom{n/2^i}{n/2^i-a} \mathbb{P}(\zeta_{n/2^i}^a \leq 0) \\ &\leq \sum_{i=1}^l 2^i \sum_{a=1}^{n/2^i-1} \binom{n/2^i}{a}^2 \exp\left\{\frac{-C_1 a(n/2^i-a)\gamma^2}{\sigma^2}\right\}. \end{aligned}$$

Working with just the inner summation, we break into two segments and get:

$$\begin{aligned}
& \sum_{a=1}^{n/2^{i+1}} \exp \left\{ C_2 a \log \frac{n}{2^i} - \frac{C_1 a (\frac{n}{2^i} - a) \gamma^2}{\sigma^2} \right\} + \sum_{a=n/2^{i+1}}^{n/2^i-1} \exp \left\{ C_3 (\frac{n}{2^i} - a) \log n/2^i - \frac{C_1 a (\frac{n}{2^i} - a) \gamma^2}{\sigma^2} \right\} \\
& \leq \max_{1 \leq a \leq n/2^{i+1}} \exp \left\{ C'_2 a \left(\log((n/2^i)^2) - \frac{C'_1 n/2^i \gamma^2}{\sigma^2} \right) \right\} \\
& + \max_{n/2^{i+1} \leq a \leq n/2^i} \exp \left\{ C'_3 (n/2^i - a) \left(\log((n/2^i)^2) - \frac{C'_1 n/2^i \gamma^2}{\sigma^2} \right) \right\}.
\end{aligned}$$

Pushing the 2^i term into the exponent, we see that if:

$$\frac{\gamma^2}{\sigma^2} \geq \frac{\log(n^2/2^i)}{C'_1 n/2^i} + \frac{\log(\log_2 n)/\delta}{C'_1 n/2^i}$$

then each term is smaller than $C \frac{\delta}{\log_2 n}$. Since there are at most $\log_2 n$ terms in the hierarchy, and consequently at most that many terms in the sum, we see that the probability of failure is upper bounded by δ . Of course to recover clusters of size m , we just need to work for all i such that $n/2^i \geq m$. Substituting this into the bound shows that it is sufficient for:

$$\frac{\gamma}{\sigma} \geq \sqrt{\frac{C_1 \log(nm)}{m} + \frac{C_2 \log \log_2 n / \delta}{m}}.$$

Rearranging this establishes the theorem.

4.8 Discussion and open problems

In this chapter we have presented a new analysis of spectral clustering in the presence of noise and established tight minimax upper and lower bounds. As our analysis of spectral clustering does not show that it is minimax-optimal it remains an open problem to further tighten, or establish the tightness of, our analysis, and to find a computationally efficient minimax procedure in the general case when similarities are not block constant. Our results apply only for binary hierarchical clusterings, yet k -way hierarchies are common in practice. In current work we are attempting to extend our results to k -way hierarchies.

Identifying conditions under which one can guarantee correctness for other forms of spectral clustering is another interesting direction for future work. For instance the recent work of Chaudhuri et al. [45], has shown that modifications of the normalized Laplacian is well suited for clustering non-homogenous degree graphs. Kumar and Kannan [120] have also recently shown that a particular spectral algorithm that uses the eigenvectors of the adjacency matrix followed by k -means succeeds at recovering clusters that satisfy a fairly general ‘‘proximity condition’’.

Finally, spectral clustering has close connections to density based clustering [88, 90]. A study of these connections was initiated in the work of Narayanan et al. [143] but these connections are still not well understood and are an interesting avenue for future work on spectral clustering.

4.9 Detailed proofs

4.9.1 Proof of Lemma 4.7.1

Our proof strategy is to first show that HS will correctly output the first split the hierarchical clustering in Lemma 4.9.1. Repeated application of this lemma concludes the proof. Recall that the ideal matrix has within cluster similarity greater than all between cluster similarities; this motivates the statement of Lemma 4.9.1.

Lemma 4.9.1. *Let W be a $(p + q) \times (p + q)$ matrix with the Large-Small block structure of*

$$\left(\begin{array}{c|c} W_L & W_S \\ \hline W_S^\top & W'_L \end{array} \right)$$

such that W_L is a $p \times p$ block, W'_L is a $q \times q$ block and

$$\min_{1 \leq i, j \leq p} (W_L)_{ij} > \max_{1 \leq i \leq p < j \leq p+q} (W_S)_{ij} > 0$$

$$\min_{p+1 \leq i, j \leq p+q} (W'_L)_{ij} > \max_{1 \leq i \leq p < j \leq p+q} (W_S)_{ij} > 0.$$

Let D be the diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. Let v be the smallest non-constant eigenvector of the graph-Laplacian $L = D - W$, then v has either the sign pattern of $\begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ where v_+ , the first p elements of v , are strictly positive and v_- , the other q elements of v , are strictly negative or the reverse sign pattern.

Proof. Step 1: First, we will show that if a $(p + q) \times (p + q)$ symmetric matrix B has the Positive-Negative block structure of

$$\left(\begin{array}{c|c} B_+ & B_- \\ \hline B_-^\top & B'_+ \end{array} \right)$$

where every *non-diagonal* element in the $p \times p$ block B_+ and the $q \times q$ block B'_+ is strictly positive and every element in the $p \times q$ block B_- is strictly negative, then the first eigenvector of B , call it v , either has the sign pattern of $\begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ where v_+ , the first p elements of v , are strictly positive and v_- , other q elements of v , are strictly negative or has the reverse sign pattern.

Let $v = \begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ be the largest eigenvector of B where v_+ are the first p elements and v_- are the other q elements. Let I_+, I_- be index sets of positive and negative elements in v_+ , and I the index of all elements in v_+ . Let J_+, J_- be index sets of positive and negative elements in v_- , and J the index of all elements in v_- . Then

$$v^\top B v = \underbrace{v_+^\top B_+ v_+}_{\text{term 1}} + \underbrace{v_+^\top B_- v_-}_{\text{term 2}} + \underbrace{v_-^\top B_-^\top v_+}_{\text{term 3}} + \underbrace{v_-^\top B'_+ v_-}_{\text{term 4}}.$$

Let us form a new vector w by changing the signs of all elements in I_- and all elements in J_+ . We now proceed to compare $w^\top B w$ with $v^\top B v$ term by term, noting that $\|w\|_2 = \|v\|_2 = 1$.

Term 1 is $v_+^\top B_+ v_+ = \sum_{i,j \in I} v_i B_{ij} v_j$. Since $B_{ij} > 0$, $w_i B_{ij} w_j \geq v_i B_{ij} v_j$ for all i, j , we notice that we have strictly increased term 1, provided that I_-, I_+ are non-empty. An analogous argument reveals that we do not decrease term 4 by changing v to w . Furthermore, we strictly increased term 4 if J_-, J_+ are non-empty.

Term 2 is $v_+^\top B_- v_-^\top = \sum_{i \in I, j \in J} v_i B_{ij} v_j$. Since $B_{ij} < 0$, we see that $w_i B_{ij} w_j = -|v_i| |B_{ij}| |v_j| \geq v_i B_{ij} v_j$ for all i, j with strict inequality whenever $i \in I_-, j \in J_-$ or $i \in I_+, j \in J_+$. Thus we have strictly increased term 2 (and 3 by analogous argument) provided that the index sets are non-empty.

We see then that unless I_-, J_+ are empty or I_+, J_- are empty, $w^\top B w > v^\top B v$. However, v is assumed to be largest eigenvector and hence maximize $v^\top B v$ among all unit-norm vectors. We reached a contradiction and thus, all of v_+ must have same sign and be opposite of v_- .

Now suppose $v_i = 0$, then $B_i^\top v = 0$ where B_i is the i -th row of B . However, since v cannot be all zero, we see then that $B_i^\top v > 0$. Thus, v_i cannot be zero for all i and v_+ is all positive and v_- is all negative.

Step 2: Now we prove the claim of the lemma. Let $\underline{1}$ be a vector of all ones. Since the W satisfy the Large-Small block structure there exist $c \in \mathbb{R}$ such that the matrix $B \triangleq c \underline{1} \underline{1}^\top - L = c \underline{1} \underline{1}^\top - D + W$ has the *Positive-Negative* block structure of

$$\left(\begin{array}{c|c} B_+ & B_- \\ \hline B_-^\top & B_+ \end{array} \right)$$

except on the diagonals.

Let $\{v^{(i)}\}$ be the eigenvectors of L with corresponding eigenvalue $\{\lambda_i\}$. Since we know that $\underline{1}$ is an un-normalized eigenvector of L with eigenvalue 0, let $v^{(1)} = \underline{1}$ and $\lambda_1 = 0$. All other eigenvectors of L must be orthogonal to $\underline{1}$ and hence, $\{v^{(i)}\}$ are also eigenvectors of B . Furthermore, for B , $\{v^{(i)}\}$ have the corresponding eigenvalues of $\{-\lambda_i\}$ except for $\{v^{(1)}\}$, which has the eigenvalue of $\{c\}$.

We know thus that the v , the largest eigenvector of B , is also the smallest non-constant eigenvector of L . By step 1, we know that v has the sign pattern of $v = (v_+ v_-)^\top$. \square

4.9.2 Proof of Lemma 4.7.2

We use several results from spectral graph theory to obtain these bounds in this subsection. To derive these bounds, we first must study a more structured matrix, which we call the Constant Block Matrix (CBM). The CBM has the same cluster structure as the HBM only it has constant off-block-diagonal similarities rather than ranges as with the HBM.

Definition 1. A similarity matrix A is a **Constant Block Matrix** if A is an ideal matrix with $\epsilon_s \triangleq \alpha_s = \beta_s$ for all clusters s .

Lemma 4.9.2. (*Spectrum of CBM*) Consider an $(n \times n)$ Constant-Block Matrix A characterized by an ϵ_s for each level s , with $\min\{\epsilon_{s-L}, \epsilon_{s-R}\} > \epsilon_s$ and with balance factor η . Then the laplacian L_A has the following eigenvalues $(\lambda_1 \leq \lambda_2, \leq \dots \leq \lambda_n)$ and eigenvectors (v_1, \dots, v_n) :

1. $v^{(1)} = \frac{1}{\sqrt{n}}\underline{1}$ with $\lambda_1 = 0$.
2. $\sqrt{\frac{1}{n\eta}} \leq |v^{(2)}(i)| \leq \sqrt{\frac{\eta}{n}}$ with $\lambda_2 = n\epsilon_0$.
3. $\frac{n}{1+\eta}(\eta\epsilon_0 + \min\{\epsilon_L, \epsilon_R\}) \leq \lambda_3 \leq \frac{n}{1+\eta}(\epsilon_0 + \eta \max\{\epsilon_L, \epsilon_R\})$.

Proof. (of Lemma 4.9.2) The first claim is true simply because L_A is a Laplacian Matrix.

We prove the remaining claims by induction on number of levels l in the A . As a base case, if A is a $n \times n$ constant matrix with $A_{ij} = \epsilon_0$ for all i, j , then it's easy to see that every vector orthogonal to $\underline{1}$ is an eigenvector of L_A with eigenvalue $n\epsilon_0$.

Suppose now that A is an $n \times n$ CBM with entries ϵ_s as in the lemma. Let C_L, C_R be the two first-level clusters. It's easy to check that the vector v with $v(i) = \sqrt{\frac{|C_R|}{n|C_L|}}, i \in C_R$ and $v(i) = \sqrt{\frac{|C_L|}{n|C_R|}}, i \in C_L$ is an eigenvector of L_A with eigenvalue $n\epsilon_0$. We want to show that $n\epsilon_0$ is the second smallest eigenvalue.

Since the diagonal blocks A_{LL} and A_{RR} are CBMs, the upper left block of the Laplacian is $(L_A)_{LL} = L_{(A_{LL})} + |C_R|\epsilon_0 I$ and the lower right block of the Laplacian is $(L_A)_{RR} = L_{(A_{RR})} + |C_L|\epsilon_0 I$. By induction, the second smallest eigenvalue of $L_{(A_{LL})}$ is $|C_L|\epsilon_L$. We can extend the corresponding eigenvector to an one for L_A by padding with zeros. This vector is associated with eigenvalue $|C_L|\epsilon_L + |C_R|\epsilon_0 > n\epsilon_0$.

Thus at least $|C_L| - 1$ eigenvalues of L_A are larger than $n\epsilon_0$. Applying the same argument to $L_{(A_{RR})}$ reveals that $n - 2$ eigenvalues are larger than $n\epsilon_0$. Since 0 is an eigenvalue of L_A , we conclude that $n\epsilon_0$ is the second smallest eigenvalue of L_A .

Since $\frac{1}{\eta} \leq \frac{|C_R|}{|C_L|} \leq \eta$, we have proved claim 2. Note that in proving claim 2, we have also shown that the third smallest eigenvalue of L_A is $\min(|C_L|\epsilon_L + |C_R|\epsilon_0, |C_R|\epsilon_R + |C_L|\epsilon_0)$. Apply the definition of η and we see that the third claim holds true as well. \square

Our proof of Lemma 4.7.2 will construct two ideal Constant-Block Matrices, show that eigenvalues and eigenvectors of the HBM A are constrained by the two CBMs, and then leverage Lemma 4.9.2 to get the final result. Before we proceed to the proof, we state two well-known results in Spectral Graph Theory that we will use:

Lemma 4.9.3. [176] If L_G and L_H are two graph Laplacians such that $L_G \succeq cL_H$, then $\lambda_k(G) \geq c\lambda_k(H)$. (where we say PSD matrices $A \succeq B$ if $A - B \succeq 0$)

Lemma 4.9.4. [176] Let $G = (V, E, w)$ and $H = (V, E, z)$ be two graphs that differ only in edge weights. Then $L_G \succeq \min_{e \in E} \frac{w(e)}{z(e)} L_H$.

Proof. (of Lemma 4.7.2): Let H_α be a two level ideal Constant-Block matrix with the same block structure as A . Let all entries of the diagonal blocks of H_α have value $\alpha_1 \triangleq \min\{\alpha_L, \alpha_R\}$

and let all entries of the off-diagonal blocks of H_α have value α_0 . Define another constant-block matrix H_β similarly, the diagonal blocks are β^* while the off-diagonal blocks are β^0 .

Lemma 4.9.2 characterizes the spectrum of H_α and H_β . Using this characterization, along with Lemmas 4.9.3 and 4.9.4, we have that $n\alpha_0 \leq \lambda_2(L_A) \leq n\beta_0$ and that $\frac{n}{1+\eta}(\eta\alpha_0 + \alpha_1) \leq \lambda_3(L_A) \leq \frac{n}{1+\eta}(\beta_0 + \eta\beta^*)$.

Combined with the fact that $\lambda_1 = 0$ for any Laplacian, we get that $\delta \geq \min(n\alpha_0, \frac{n}{\eta+1}(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0))$. Under Range Restriction Assumption 3, we see that $(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0) > 0$ and hence $\delta = \Theta(n)$.

To establish bounds on entries of $v^{(2)}$, we consider a single coordinate of $v^{(2)}$; using the definition of eigenvector we get that

$$v^{(2)}(i) = \frac{A_i v^{(2)}}{d_i - \lambda_2}$$

where A_i is the i -th row of A . From Lemma 4.7.1, we can assume without loss of generality that $v^{(2)}(i)$ is all strictly positive for one cluster and strictly negative for other. From the fact that $\underline{1}$ is an eigenvector of L_A , we get that $\sum_{i:v^{(2)}(i)>0} |v^{(2)}(i)| = \sum_{i:v^{(2)}(i)<0} |v^{(2)}(i)|$. Hence:

$$J(\alpha_1 - \beta_0) \leq A_i v^{(2)} \leq J(\beta^* - \alpha_0),$$

where $J = \frac{1}{2} \sum_i |v^{(2)}(i)|$. We can similarly derive an upper and lower bound for $d_i - \lambda_2$:

$$\begin{aligned} n \frac{1}{1+\eta} \alpha_1 + n \frac{\eta}{1+\eta} \alpha_0 - n\beta_0 \\ \leq d_i - \lambda_2 \leq n \frac{1}{1+\eta} \beta_0 + n \frac{\eta}{1+\eta} \beta^* - n\alpha_0. \end{aligned}$$

Note that with the Range Restriction, the lower bound of $d_i - \lambda_i$ is positive and is $\Theta(n)$. Combining these two results, we get

$$\begin{aligned} \frac{Jc_1}{n} &\leq |v^{(2)}(i)| \leq \frac{Jc_2}{n}, \\ c_1 &= \frac{(\alpha_1 - \beta_0)(\eta + 1)}{\beta_0 + \eta\beta^* - (1 + \eta)\alpha_0}, \\ c_2 &= \frac{(\beta^* - \alpha_0)(\eta + 1)}{\alpha_1 + \eta\alpha_0 - (1 + \eta)\beta_0}. \end{aligned}$$

Since $v^{(2)}$ must be a unit vector, we can bound J and get that

$$\frac{c_1}{c_2} \frac{1}{\sqrt{n}} \leq |v^{(2)}(i)| \leq \frac{c_2}{c_1} \frac{1}{\sqrt{n}}.$$

Set $K_\eta = \left(\frac{c_2}{c_1}\right)^2$ and we get the desired result. \square

4.9.3 Proof of Lemma 4.7.3

We begin with some preliminary lemmas concerning the behavior of sub-Gaussian random variables.

Lemma 4.9.5. (*Max of sub-Gaussians*) Let X_1, \dots, X_n be identically distributed sub-Gaussian random variables with scale σ . With probability $1 - \delta$

$$\max_{i=1, \dots, n} |X_i| \leq \sigma \sqrt{2 \log n + 2 \log \frac{2}{\delta}}.$$

Lemma 4.9.6. (*Sums of sub-Gaussians*) Suppose X_1, \dots, X_n are independent sub-Gaussian random variables, each with $\mathbb{E}(e^{tX_i}) \leq e^{\frac{\sigma_i^2 t^2}{2}}$. For any scalars a_1, \dots, a_n independent of X_1, \dots, X_n we have, $\sum_{i=1}^n a_i X_i$ is a sub-Gaussian random variable with $\mathbb{E}(e^{t \sum_{i=1}^n a_i X_i}) \leq e^{\frac{t^2 \sum_{i=1}^n a_i^2 \sigma_i^2}{2}}$.

Operator norm bounds on matrices of sub-Gaussians.

Lemma 4.9.7 (Proposition 2.4 of [164]). Consider a matrix R with independent sub-Gaussian entries with scale factor σ . The operator norm of R is $O(\sqrt{n})$ and satisfies

$$P(\|R\|_2 \geq A\sigma\sqrt{n}) \leq 2 \exp(-cA^2n)$$

for absolute constants c, C and for all $A \geq C$.

To obtain operator norm bounds on *symmetric* sub-Gaussian matrices, we just note that it suffices to consider the upper triangular entries and strictly lower triangular entries separately, and appeal to the above Lemma with the triangle inequality. By suitably adjusting constants we obtain the following:

Lemma 4.9.8. Consider a symmetric matrix R as described in Definition 2, with scale factor σ . The operator norm of R is $O(\sqrt{n})$ and satisfies

$$P(\|R\|_2 \geq A\sigma\sqrt{n}) \leq 2 \exp(-cA^2n)$$

for absolute constants c, C and for all $A \geq C$.

The matrix whose operator norm we will ultimately have to bound is L_R , we derive this bound next:

Lemma 4.9.9. (*Noise-Laplacian*) Let R be a perturbation matrix, let $L_R = D_R - R$. For all $n \geq n_0$, we have that with probability at least $1 - 4/n$,

$$\|L_R\|_2 \leq 4\sigma\sqrt{n \log n}$$

where n_0 is an absolute constant.

Proof.

$$\|L_R\|_2 = \|D_R - R\|_2 \leq \|D_R\|_2 + \|R\|_2.$$

D_R is diagonal and $\|D_R\|_2$ is the largest (in absolute value) diagonal element. Since every diagonal element of D_R is subgaussian with scale factor $\leq \sqrt{n}\sigma$, we can apply Lemma 4.9.5 and

get that $\|D_R\|_2 \leq \sigma\sqrt{n}\sqrt{2\log n + 2\log \frac{4}{\delta}}$ with probability at least $1 - \delta/2$. Setting $\delta = 4/n$ we have $\|D_R\|_2 \leq 2\sigma\sqrt{n\log n}$.

Using Lemma 4.9.8, we know that with probability $1 - 8/n$, for n large enough (depending on the absolute constants c and C), $\|R\|_2 = C\sigma\sqrt{n}$. Hence, for n large enough, $\|D_R\|_2 \geq \|R\|_2$ and $\|L_R\|_2 \leq 2\|D_R\|_2$ and we get the desired result. \square

In order to guarantee recovery of all clusters of size at least m , it is not sufficient to bound $\|L_R\|$ at just the top-most level of the hierarchy. We must ensure that the noise matrices for all of the subclusters we hope to recover have uniformly bounded spectral norm (where the specific bound could be different for different submatrices). The following lemmas establish the desired uniform bound.

Before we present the lemmas, we specify our notation. For each level $l \in \{0, \dots, n\}$ in the hierarchy, denote the set of clusters at level l by $\{C_{li} : i \in \{1 \dots, 2^l\}\}$ and let $m_{li} = |C_{li}|$. For any subcluster C_{li} we write the corresponding noise degree matrix as $D_R^{C_{li}}$ and the corresponding noise matrix as $R^{C_{li}}$.

Lemma 4.9.10. (*Hierarchical Noise Degree Bound*) *Let R be the noise matrix associated with a $n \times n$ noisy Hierarchical Block Matrix satisfying Assumptions 1 and 2. Then with probability $1 - 2/n$, for all sub-clusters C_{li} in the true hierarchical clustering, the corresponding noise degree matrix $D_R^{C_{li}}$ will have operator norm bounded by*

$$\|D_R^{C_{li}}\|_2 \leq \sigma\sqrt{6m_{li}\log n}$$

Proof. We first bound the number of levels in the tree. l is bounded by $\log n$ in the balanced binary case, but bounded by n in the worst case irrespective of η .

Now, at each level we bound at most n random draws from various sub-Gaussians. For instance, consider the first level. We need to bound the operator norm of a diagonal degree matrix, and each diagonal entry is a draw from a sub-Gaussian with scale factor at most $\sqrt{n}\sigma$, and there are at most n diagonal entries. On the second level we will have two matrices but still n degree random variables we will need to bound. Over l levels there are at most nl random variables to bound.

For a cluster C_{li} at level l of size m_{li} each diagonal entry of $D_R^{C_{li}}$ is a subgaussian with scale factor $\sigma\sqrt{m_{li}}$. To standardize we will look at $\|D_R^{C_{li}}\|_2/(\sigma\sqrt{m_{li}})$ so that all of the terms have scale factor 1. Now by the application of a union bound:

$$\begin{aligned} & \mathbb{P}[\exists C_{li} \|D_R^{C_{li}}\|_2/(\sigma\sqrt{2m_{li}}) \geq \epsilon] \leq \sum_{l=1}^n \sum_{i=1}^{2^l} \mathbb{P}[\|D_R^{C_{li}}\|_2/(\sigma\sqrt{2m_{li}}) \geq \epsilon] \\ & \leq \sum_{l=1}^n \sum_{i=1}^{2^l} \sum_{j \in C_{li}} \mathbb{P}[(D_R^{C_{li}})_{jj}/(\sigma\sqrt{2m_{li}}) \geq \epsilon] \leq \sum_{l=1}^n \sum_{i=1}^{2^l} \sum_{j \in C_{li}} 2 \exp\{-\epsilon^2\} \\ & \leq 2n^2 \exp\{-\epsilon^2\}. \end{aligned}$$

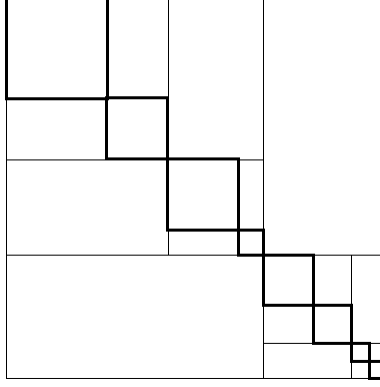


Figure 4.8: All sub-matrices corresponding to sub-clusters at level 3

Setting $\epsilon = \sqrt{3 \log n}$ bounds this probability by $2/n$. Wrapping up, we see that:

$$\frac{\|D_R^{C_{li}}\|_2}{\sigma \sqrt{2m_{li}}} \leq \sqrt{3 \log n}$$

for every cluster C_{li} with probability $\geq 1 - 2/n$. The lemma follows by algebraic manipulation. \square

Proof of Lemma 4.7.3. Let us now bound the number of levels in the tree. We will need to be more careful than in Lemma 4.9.10 where bounding l by n did not affect the rate. When the clusters are imbalanced with a balance factor η we have

$$l \leq \frac{1}{\log(\frac{1+\eta}{\eta})} \log n = C_\eta \log n,$$

with $C_\eta = \frac{1}{\log(\frac{1+\eta}{\eta})}$. To see this note that at each split the larger cluster is of size at most $\frac{\eta}{1+\eta}n$. After l levels the cluster size is at most 1, i.e.

$$\left(\frac{\eta}{1+\eta}\right)^l n = 1.$$

We can solve this to obtain that $l \leq C_\eta \log n$.

Returning to the proof, we note that we need to bound the norm of at most $2^{l+1} - 2 \leq e^{2l}$, sub-Gaussian matrices of varying sizes.

From Lemma 4.9.8 we know also that for each C_{li} , $\|R^{C_{li}}\|_2 \leq B_{li} \sigma \sqrt{m_{li}}$ holds with probability at least $\exp(-cB_{li}^2 m_{li})$, where $B_{li} \geq C$ for some absolute constant C .

By letting

$$B_{li} = \max \left(\sqrt{\frac{2C_\eta \log n + \log \frac{2}{\delta}}{cm_{li}}}, C \right),$$

we can take union bound over all $2^{l+1} - 2$ noise sub-matrices and get that with probability at least $1 - \delta$, for all sub-clusters C_{li} ,

$$\|R^{C_{li}}\|_2 \leq \max \left(\sigma \sqrt{\frac{2C_\eta \log n + \log \frac{1}{\delta}}{c}}, C\sigma\sqrt{m_{li}} \right).$$

Taking $\delta = 2/n$ we have

$$\|R^{C_{li}}\|_2 \leq \max \left(\sigma \sqrt{\frac{3C_\eta \log n}{c}}, C\sigma\sqrt{m_{li}} \right) \leq C(\eta)\sigma\sqrt{m_{li} \log n}$$

for a constant $C(\eta)$ depending on C, c and C_η . From Lemma 4.9.10, we know that with probability $1 - 2/n$, for every C_{li} , $\|D_R^{C_{li}}\|_2 \leq \sigma\sqrt{6m_{li} \log n}$.

Hence, with probability at least $1 - 4/n$, for every sub-cluster C_{li} , $\|L_R^{C_{li}}\|_2 \leq C(\eta)\sigma\sqrt{m_{li} \log n}$. \square

4.9.4 Davis Kahan

Proof. (of Lemma 4.7.4) Note that $L_R + L_A = L_W$.

From Davis-Kahan theorem, we know that

$$|\sin \theta_i| \leq \frac{\|L_R\|}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

where λ_i and λ_j are respectively the i -th and j -th smallest eigenvalue of L_A and θ_i is the angle between $v^{(i)}$ and $u^{(i)}$, i.e. $\cos \theta_i = v^{(i)\top} u^{(i)}$

Without loss of generality, we can orient vectors as desired and assume that $|\theta_i| \leq \frac{\pi}{2}$. Since $v^{(i)}$ and $u^{(i)}$ are unit vectors, we get that

$$\begin{aligned} \|u^{(i)} - v^{(i)}\|_2 &\leq \sqrt{\|v^{(i)}\|^2 + \|u^{(i)}\|^2 - 2v^{(i)\top} u^{(i)}} \\ &= |2 \sin \frac{\theta_i}{2}| \leq |2\sqrt{2} \sin \frac{\theta_i}{2} \cos \frac{\theta_i}{2}| = |\sqrt{2} \sin \theta_i|. \end{aligned}$$

The second inequality follow because $\sqrt{2} \cos \frac{\theta_i}{2} \geq 1$ under assumption that $|\theta_i| \leq \frac{\pi}{2}$. Combining this with Davis-Kahan gives us the desired result. \square

For ease of reference, we also state here a well-known result in perturbation theory that we use.

Lemma 4.9.11. (Weyl's Inequality) *Let L_W, L_A be $n \times n$ positive definite matrices and let $L_R = L_W - L_A$. Let $\lambda_1 \leq \dots \leq \lambda_n$ and $\mu_1 \leq \dots \leq \mu_n$ be the eigenvalues of L_A and L_W respectively. Then, for all i , $|\lambda_i - \mu_i| \leq \|L_R\|_2$.*

4.9.5 Proof of ℓ_∞ Deviations

Proof of Lemma 4.7.5. We start by manipulating the eigenvector equations for $u^{(2)}$ and $v^{(2)}$, which will give us an expression for $k(i)$:

$$\begin{aligned} k(i) &= \frac{(L_{Ai} + L_{Ri})u^{(2)}}{\mu_2} - v^{(2)}(i) = \frac{L_{Ai}(v^{(2)} + k)}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2} - v^{(2)}(i) \\ &= \frac{\lambda_2 v^{(2)}(i)}{\mu_2} + \frac{L_{Ai}k}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2} - v^{(2)}(i) = v^{(2)}(i) \frac{\lambda_2 - \mu_2}{\mu_2} + \frac{L_{Ai}k}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2} \\ &= \frac{1}{\mu_2} \left(v^{(2)}(i)(\lambda_2 - \mu_2) + D_{Ai}k(i) - A_i k + L_{Ri}v^{(2)} + D_{Ri}k(i) - R_i k \right), \end{aligned}$$

where L_{Ai} is the i -th row of L_A . Rearranging, we get

$$k(i) = \frac{1}{c_i} \left(v^{(2)}(i)(\lambda_2 - \mu_2) - A_i k + L_{Ri}v^{(2)} - R_i k \right)$$

where $c_i = \mu_2 - D_{Ai} - D_{Ri}$. We are interested in the absolute difference and by the triangle inequality, we have:

$$|k(i)| \leq \frac{\overbrace{|v^{(2)}(i)(\lambda_2 - \mu_2)|}^{T_1} + \overbrace{|A_i k|}^{T_2} + \overbrace{|L_{Ri}v^{(2)}|}^{T_3} + \overbrace{|R_i k|}^{T_4}}{\underbrace{|c_i|}_{T_5}}.$$

Call the numerator terms T_1, T_2, T_3 and T_4 and the denominator T_5 . We will bound each separately.

Bound on T_1 : Using Lemma 4.7.2, Weyl's inequality (Lemma 4.9.11), and our spectral norm bound (Lemma 4.7.3), we see that with probability at least $1 - 4/n$:

$$T_1 = |v^{(2)}(i)(\lambda_2 - \mu_2)| = |v^{(2)}(i)| |\lambda_2 - \mu_2| \leq \sqrt{\frac{K_\eta^*}{m_s}} \|L_R\|_2 \leq 2\sigma \sqrt{6K_\eta^* \log n}.$$

Bound on T_2 : Remember that, $\kappa^* = \min(\alpha_0, \frac{\gamma_s^*}{1+\eta})$.

$$T_2 = |A_i k| \leq \|A_i\|_2 \|k\|_2 \leq \frac{\sqrt{m_s} \beta^* \sqrt{2} \|L_R\|_2}{\xi_2} \leq \frac{4\sigma \beta^*}{\kappa^*} \sqrt{3 \log n},$$

where ξ_2 is the eigengap corresponding to the second eigenvector. The first inequality is Cauchy-Schwarz while the second follows from Lemma 4.7.4. The third inequality uses Lemma 4.7.2 to bound the eigengap ξ_2 which is at least $m_s \kappa^*$, and Lemma 4.7.3 to bound $\|L_R\|_2$. This inequality holds under the same $1 - 4/n$ probability event used in the T_1 bound.

Bound on T_3 : The terms T_3 and T_4 are the main “noise” terms.

$$T_3 = |L_{Ri}v^{(2)}| = |D_{Ri}v^{(2)}(i) - R_i v^{(2)}|.$$

Since each entry R_{ij} is subgaussian with scale factor σ , R_{ij} and $R_{ij'}$ are independent for all $j \neq j'$, and $\|v^{(2)}\|_2 = 1$, we conclude that $R_i v^{(2)}$ is distributed as a subgaussian with scale factor σ . Moreover, $D_{Ri}v^{(2)}(i)$ is a subgaussian random variable with scale factor $\leq \sqrt{K_\eta^*} \sigma$ since $v^{(2)}(i) \leq \sqrt{\frac{K_\eta^*}{m_s}}$ and each entry D_{Ri} is subgaussian with scale factor $\sqrt{m_s} \sigma$. Since $\sigma^2 \leq K_\eta^* \sigma^2$, T_3 is a draw from a subgaussian with scale factor $\leq \sqrt{2K_\eta^*} \sigma$.

To ensure that T_3 is uniformly low for all i , we take a union bound and use Lemma 4.9.5. Note that this union bound is across all levels of the hierarchy, so there are $nl \leq n^2$ subgaussians that we must bound. We get that with probability at least $1 - 2/n$,

$$T_3 \leq 4\sigma \sqrt{K_\eta^* 3 \log n}.$$

Bound on T_4 :

$$T_4 = |R_i k| \leq \|R_i\|_2 \|k\|_2 \leq \|R\|_2 \|k\|_2.$$

From the proof of Lemma 4.7.3, we see that for $m_s = \omega(\log n)$, and for n large enough, under the $1 - 4/n$ probability event described in T_1 ,

$$\|R\|_2 \leq C\sigma \sqrt{m_s}$$

for some absolute constant C . So we have,

$$T_4 \leq C\sigma \sqrt{m_s} \frac{\sqrt{2} \|L_R\|_2}{\xi_2} \leq \frac{4C\sigma^2 \sqrt{3 \log n}}{\kappa^*}.$$

Bound on T_5 : The term T_5 appears in the denominator and here, we establish a lower bound on it.

$$T_5 = |\mu_2 - D_{Ai} - D_{Ri}| = |D_{Ai} + D_{Ri} - \mu_2|.$$

Note that $D_{Ai} \geq \frac{m_s}{1+\eta}(\eta\alpha_s + \alpha_{s+})$ where $\alpha_{s+} \triangleq \min\{\alpha_{s0L}, \alpha_{s0R}\}$ and that $\mu_2 \leq \lambda_2 + \|L_R\|_2 \leq m_s \beta_s + \|L_R\|_2$. Hence:

$$\begin{aligned} T_5 &\geq \left| \frac{m_s}{1+\eta}(\eta\alpha_s + \alpha_{s+}) + D_{Ri} - m_s \beta_s - \|L_R\|_2 \right| \\ &\geq \frac{m_s}{1+\eta} \left| \alpha_{s+} + \eta\alpha_s - (1+\eta)\beta_s - \frac{1+\eta}{m_s}(2\|D_R\|_2 + \|R\|_2) \right|. \end{aligned}$$

The inequalities only hold provided that the term inside the absolute value is ≥ 0 . Note that $\alpha_{s+} + \eta\alpha_s - (1+\eta)\beta_s$ is just γ . We will show that for large enough n , this is indeed true. Under the $1 - 4/n$ probability event described in the T_1 bound, we have:

$$2\|D_R\|_2 + \|R\|_2 \leq 3\sigma \sqrt{6m_s \log n}.$$

Now, provided that $\sigma = o(\gamma \sqrt{\frac{m_s}{\log n}})$, and using the definition of γ , we have that then $\frac{1+\eta}{m_s}(2\|D_R\|_2 + \|R\|_2) = o(\gamma)$, and for large enough n , we can conclude that $\gamma - \frac{1+\eta}{m_s}(2\|D_R\|_2 + \|R\|_2) \geq \frac{\gamma}{2} \geq \frac{\gamma_S^*}{2}$. From the statement of the theorem we have $\sigma = o(\min(\kappa^{*5} \sqrt{\frac{m_s}{\log n}}, \kappa^{*4} \sqrt[4]{\frac{m_s}{\log n}})) = o(\gamma \sqrt{\frac{m_s}{\log n}})$. Therefore:

$$T_5 \geq \frac{m_s}{1+\eta} \frac{\gamma_S^*}{2} \geq \frac{m_s \kappa^*}{2}.$$

Combining all of the bounds yields the statement of the Lemma. \square

4.10 Proof of Theorem 4.4.1

The proof will be very similar to that of Theorem 4.3.1.

The difficulty here is that the spectral embedding of each point is not just a single number, but rather a k -dimensional vector. To make matters worse, because L_A has a k -dimensional eigenspace associated with eigenvalue 0 (in other words, eigenvalue 0 has geometric multiplicity k), there are many different possible spectral embeddings of each point—one for each set of basis of the eigenspace.

Let $u^{(1)}, \dots, u^{(k)}$ be perturbed eigenvectors of L_W . The set of $u^{(j)}$'s cannot be close to all sets of lowest k eigenvectors of L_A because there are infinite number of sets of lowest k eigenvectors of L_A due to geometric multiplicity. Thus, the best we can say is that there exist at least one set of lowest k eigenvectors of L_A that is close to $u^{(1)}, \dots, u^{(k)}$. Lemma 4.10.1, 4.10.2 formalize these concepts.

The following Lemmas extend Davis-Kahan theorem to describe perturbation of subspaces:

Lemma 4.10.1. *Let W be a matrix with eigenvalues $\mu_1 \leq \mu_2, \dots \leq \mu_n$ (possibly with multiplicity) and corresponding eigenvectors u_1, u_2, \dots, u_n . Let A be a matrix with eigenvalues $\lambda_1 \leq \lambda_2, \dots \leq \lambda_n$ (possibly with multiplicity) and corresponding eigenvectors v_1, v_2, \dots, v_n . Let $R \equiv W - A$.*

Let $U = \text{span}\{u_i\}_{i \in I}$ where I is some index set. Let $V = \text{span}\{v_i\}_{i \in I}$. Then we have, for all unit-normed $u \in U$:

$$\|P_{V^\perp} u\|_2 \leq \frac{2\|R\|_2}{\delta} \sqrt{k}.$$

where $k \equiv \dim U = \dim V$, P_{V^\perp} is the orthogonal projection onto V^\perp , $\delta \equiv \min_{i \in I} \delta_i$ and $\delta_i \equiv \min_{j \notin I} |\lambda_i - \lambda_j|$.

Intuitively, U , an eigen-subspace of W must be close to V , the corresponding eigen-subspace of A . We simply quantified “close” as the projection of U onto V^\perp .

Proof. Let U, V be eigen-subspaces of W, A as defined in theorem. Fix $i \in I$ and let μ_i be an eigen-value that correspond to $u_i \in U$. Define $\bar{A} = A - \lambda_i I$ and $\bar{W} = W - \lambda_i I$.

Recall that u_i is the eigenvector of W that correspond to μ_i ; we can expand u_i in the eigenbasis of A and get $u_i = \sum_j c_j v_j$.

$$\begin{aligned} \|\bar{A}u_i\|_2^2 &= \|\bar{A} \sum_j c_j v_j\|_2^2 \\ &= \sum_j c_j^2 (\lambda_j - \lambda_i)^2 \\ &\geq \sum_{j \notin I} c_j^2 (\lambda_j - \lambda_i)^2 \\ &\geq \delta_i^2 \sum_{j \notin I} c_j^2 \\ &= \delta_i^2 \|P_{V^\perp} u_i\|_2^2. \end{aligned}$$

By using Weyl's Inequality, we can upper bound $\|\bar{A}u_i\|$ as such:

$$\|\bar{A}u_i\|_2 \leq \|\bar{W}u_i\|_2 + \|R\|_2 \leq |\mu_i - \lambda_i| + \|R\|_2 \leq 2\|R\|_2.$$

Combine the two results, we get:

$$\|P_{V^\perp} u_i\|_2 \leq \frac{2\|R\|_2}{\delta_i}.$$

Let $u \in U$ and let $\|u\|_2 = 1$, then $u = \sum_{j \in I} c_j u_j$. We will now upper bound $\|P_{V^\perp} u\|_2$.

$$\begin{aligned} \|P_{V^\perp} u\|_2^2 &= \left\| \sum_{j \in I} c_j P_{V^\perp} u_j \right\|_2^2 \\ &= \sum_{j \in I} c_j^2 \|P_{V^\perp} u_j\|_2^2 + \sum_{j \neq i, i \in I} c_j c_i \langle P_{V^\perp} u_j, P_{V^\perp} u_i \rangle. \end{aligned}$$

We already have that $\|P_{V^\perp} u_i\|_2^2 \leq \frac{4\|R\|_2^2}{\delta_i^2}$. Define $\delta = \min_i \delta_i$, then we have $\|P_{V^\perp} u_i\|_2^2 \leq \frac{4\|R\|_2^2}{\delta^2}$.

By Cauchy-Schwartz, we get $\langle P_{V^\perp} u_j, P_{V^\perp} u_i \rangle \leq \|P_{V^\perp} u_j\|_2 \|P_{V^\perp} u_i\|_2 \leq \frac{4\|R\|_2^2}{\delta^2}$.

Combine the two above bounds, we can now continue upper bounding $\|P_{V^\perp} u\|$:

$$\begin{aligned} \|P_{V^\perp} u\|_2^2 &\leq \frac{4\|R\|_2^2}{\delta^2} \left(\sum_{j \in I} c_j^2 + \sum_{j \neq i, i \in I} |c_i| |c_j| \right) \\ &\leq \frac{4\|R\|_2^2}{\delta^2} \left(\sum_{j \in I} |c_j| \right)^2 \\ &\leq \frac{4\|R\|_2^2}{\delta^2} k \sum_{j \in I} |c_j|^2 \\ &\leq \frac{4\|R\|_2^2}{\delta^2} k. \end{aligned}$$

Thus, we get $\|P_{V^\perp}u\|_2 \leq \frac{2\|R\|_2}{\delta}\sqrt{k}$ as desired. \square

Lemma 4.10.2. (*Eigenspace-Perturbation*) Let $U = \text{span}\{u_i\}_{i \in I}$ and $V = \text{span}\{v_i\}_{i \in I}$ be eigen-subspaces of matrices W, A respectively.

Assume $\frac{2\|R\|_2}{\delta}\sqrt{k} \leq 1/2$, then there exist a V -invariant isometry (orthonormal matrix) Θ such that for all i

$$\|\Theta v_i - u_i\|_2 \leq \frac{6\|R\|_2}{\delta}\sqrt{k}.$$

We say that Θ is V -invariant if for all $v \in V$, $\Theta v \in V$.

The difficulty of proving Lemma 4.10.2 comes from the fact that $P_V u_i$ and $P_V u_j$ need not be orthogonal even if u_i and u_j are orthogonal. We use the next PSD Deviation Lemma to address this difficulty.

Lemma 4.10.3. (*PSD Deviation*) Let K be a positive definite matrix with some eigenvectors that span V . Let $0 \leq \theta < 1$ and let all eigenvalues of K be between $1 + \theta$ and $1 - \theta$.

Then $\|Kv - v\|_2 \leq \theta\|v\|_2$ for all $v \in V$.

Proof. (of Lemma 4.10.3) Let w_1, \dots, w_k be the eigenvectors of K that span V with corresponding eigenvalues $\lambda_1, \dots, \lambda_k$.

Then $u = \sum_k c_k w_k$ and we get:

$$\begin{aligned} \|Kv - v\|_2 &= \left\| \sum_k c_k K w_k - \sum_k c_k w_k \right\|_2 \\ &= \left\| \sum_k c_k \lambda_k w_k - \sum_k c_k w_k \right\|_2 \\ &= \left\| \sum_k c_k (\lambda_k - 1) w_k \right\|_2 \\ &\leq (\max_i |\lambda_i - 1|) \left\| \sum_k c_k w_k \right\|_2 \\ &= \theta \|v\|_2. \end{aligned}$$

\square

Now we can prove the Eigenspace Perturbation lemma:

Proof. (of Lemma 4.10.2)

Define $v'_i = P_V u_i$ for $i \in I$. The collection of vectors $\{v'_i\}_{i \in I}$ need not be orthogonal, but we claim they are independent. To see this, suppose that there exist coefficients c_i such that $\sum_{i \in I} c_i v'_i = 0$. Then

$$\sum_{i \in I} c_i v'_i = \sum_{i \in I} c_i P_V(u_i) = P_V\left(\sum_{i \in I} c_i u_i\right) = 0.$$

The vector $\sum_{i \in I} c_i u_i$ is in U and non-zero. Hence, by Lemma 4.10.1 and the assumption that $\frac{2\|R\|_2}{\delta} \sqrt{k} \leq 1/2$, $\|P_V(\sum_{i \in I} c_i u_i)\|_2 \geq \frac{1}{2} \|\sum_{i \in I} c_i u_i\|_2 > 0$. This is a contradiction.

Because v'_i 's are independent, there exist a basis-transform linear operator K such that $Kv'_i = v_i$ for all i and $Kw = w$ for all $w \notin V$. Note that K is V -invariant since $\{v'_i\}_{i \in I}$ spans V .

Let $K = \Psi K^*$ be the V -invariant polar decomposition of K , that is, Ψ is an isometry, K^* is positive semidefinite, and K^* and Ψ are both V -invariant. Since Ψ is an isometry and hence preserves inner product, we get that the collection of vectors $\{K^*v'_i\}_{i \in I}$ must be orthogonal.

Also, since Ψ is an isometry and hence preserves norm, we get that $\|K^*v'_i\|_2 = 1$ for all $i \in I$ and $K^* \circ P_V$ is an isometry when restricted to subspace U . Since the eigenvalues of P_V restricted to U are bounded between 1 and $1 - \frac{2\|R\|_2}{\delta} \sqrt{k}$, we get that the eigenvalues of K^* restricted to $\text{range}(P_V) = V$ must be bounded between 1 and $1 / \left(1 - \frac{2\|R\|_2}{\delta} \sqrt{k}\right)$.

By assumption from theorem, we can bound, by using the fact that $\frac{1}{1-a} \leq 1+2a$ for $0 \leq a \leq 1/2$, the eigenvalues of K^* between 1 and $1 + 4\frac{\|R\|_2}{\delta} \sqrt{k}$. Hence, by Lemma 4.10.3, we get that for all $v \in V$, $\|K^*v - v\|_2 \leq 4\frac{\|R\|_2}{\delta} \sqrt{k} \|v\|_2$. Thus, we get:

$$\begin{aligned} \|u_i - K^*P_V u_i\|_2 &\leq \|u_i - P_V u_i\|_2 + \|K^*P_V u_i - P_V u_i\|_2 \\ &\leq 2\frac{\|R\|_2}{\delta} + 4\frac{\|R\|_2}{\delta} \sqrt{k} \|P_V u_i\|_2 \\ &\leq 6\frac{\|R\|_2}{\delta} \sqrt{k}. \end{aligned}$$

We used the fact that $\|u_i - P_V u_i\|_2 = \|P_{V^\perp} u_i\|_2$, and Lemma 4.10.1 for the second inequality and the trivial upper bound that $\|P_V u_i\|_2 \leq 1$ for the third inequality.

Since $v_i = KP_V u_i = \Psi K^* P_V u_i$, $\Psi^{-1}v_i = K^* P_V u_i$. Hence, we have proven the theorem with Ψ^{-1} as the isometry. \square

The next lemma describes the spectrum of the Laplacian of a k -Block Diagonal similarity matrix in a manner similar to Lemma 4.9.2 and Lemma 4.7.2.

Lemma 4.10.4. *Let A be a k -Block Diagonal Matrix with blocks $A^{(1)}, \dots, A^{(k)}$ such that all entries in $A^{(1)}, \dots, A^{(k)}$ are between β_1 and β_0 where $0 < \beta_0 \leq \beta^*$ and all remaining entries of A are 0, i.e.*

$$W = \begin{bmatrix} A^{(1)} & & \dots & 0 \\ & A^{(2)} & & \\ \dots & & \dots & \\ 0 & & & A^{(k)} \end{bmatrix}.$$

Let $0 < \nu < 1$ be such that νn is the size of the largest cluster. Then:

1. $\lambda_1, \dots, \lambda_k$, the lowest k eigenvalues of L_A , are 0 with corresponding eigenvectors

$$\begin{aligned} v^{(1)} &= \frac{1}{\sqrt{|C_1|}} (\mathbf{1}_{C_1}, \mathbf{0}_{C_2}, \dots, \mathbf{0}_{C_k}) \\ v^{(2)} &= \frac{1}{\sqrt{|C_2|}} (\mathbf{0}_{C_1}, \mathbf{1}_{C_2}, \dots, \mathbf{0}_{C_k}) \\ &\dots \\ v^{(k)} &= \frac{1}{\sqrt{|C_k|}} (\mathbf{0}_{C_1}, \mathbf{0}_{C_2}, \dots, \mathbf{1}_{C_k}) \end{aligned}$$

where $\mathbf{0}_{C_1}$ is an all-zero vector of length $|C_1|$.

2. $\lambda_{k+1} \geq \frac{\nu n}{\eta} \beta_0$ (note that $\frac{\nu n}{\eta}$ lower bounds size of the smallest cluster).

Proof. The first claim follows because L_A is also block-diagonal and the diagonal blocks $(L_A)^{(i)} = L_{A^{(i)}}$.

To prove the second claim, we construct a block-diagonal matrix S with the same block structure as A and furthermore, the diagonals $S^{(1)}, \dots, S^{(k)}$ all have constant value of β_0 . The claim then follows by Lemma 4.9.3 and Lemma 4.9.4. \square

Now we proceed to the proof of Theorem 4.4.1:

Proof. (of Theorem 4.4.1)

Let $j \in \{1, \dots, k\}$, define $v^{(j)} = \Theta v^{(j)}$. Since Θ is V -invariant, we know that $L_A v^{(j)} = 0$.

Let let $h^{(j)} = u^{(j)} - v^{(j)}$.

$$\begin{aligned} h^{(j)}(i) &= u^{(j)}(i) - v^{(j)}(i) \\ &= \frac{(L_{A_i} + L_{R_i})u^{(j)}}{\mu_j} - v^{(j)}(i) \\ &= \frac{L_{A_i}(v^{(j)} + h^{(j)})}{\mu_j} + \frac{L_{R_i}u^{(j)}}{\mu_j} - v^{(j)}(i) \\ &= \frac{L_{A_i}h^{(j)}}{\mu_j} + \frac{L_{R_i}u^{(j)}}{\mu_j} - v^{(j)}(i) \\ &= \frac{D_{A_i}h^{(j)}(i) - A_i h^{(j)}}{\mu_j} + \\ &\quad \left(\frac{L_{R_i}v^{(j)}}{\mu_j} + \frac{D_{R_i}h^{(j)}(i)}{\mu_j} - \frac{R_i h^{(j)}}{\mu_j} \right) - v^{(j)}(i). \end{aligned}$$

We will collect the terms containing $h(i)$ and get

$$\begin{aligned} & \mu_j h^{(j)}(i) - D_{A_i} h^{(j)}(i) - D_{R_i} h^{(j)}(i) \\ &= -A_i h^{(j)} + L_{R_i} v^{(j)} - R_i h^{(j)} - v^{(j)}(i) \mu_j \end{aligned}$$

and hence

$$h^{(j)}(i) = \frac{1}{\underbrace{|\mu_j - D_{A_i} - D_{R_i}|}_{T_5}} \left(\underbrace{|v^{(j)}(i)|}_{T_1} + \underbrace{|A_i h^{(j)}|}_{T_2} + \underbrace{|L_{R_i} v^{(j)}|}_{T_3} + \underbrace{|R_i h^{(j)}|}_{T_4} \right).$$

Call the numerator terms T_1, T_2, T_3, T_4 and call the denominator term T_5 . We will bound each of these terms uniformly across all clusters $j = 1, \dots, k$ and across all elements $h^{(j)}(i), i = 1, \dots, n$.

Bound for T_1 : Since Θ is V -invariant, we know that $v^{(j)} = \sum_{t=1}^k \alpha_t v^{(t)}$ and hence, $v^{(j)}$ has vector-structure of $(\frac{1}{\sqrt{|C_1|}} \alpha_1, \frac{1}{\sqrt{|C_2|}} \alpha_2, \frac{1}{\sqrt{|C_3|}} \alpha_3, \dots)$ where $\underline{\alpha}_1$ is sub-vector of length $|C_1|$ etc.

Because $\alpha_t \leq 1$ for all j , we know that $|v^{(j)}(i)| \leq \sqrt{\frac{\eta}{\nu n}}$.

We can bound $|\mu_j| \leq \|L_R\|_2 + |\lambda_j| = \|L_R\|_2$ by Weyl's Inequality. By Lemma 4.9.9, $\|L_R\|_2 \leq 4\sigma\sqrt{n \log n}$ with probability at least $1 - \frac{4}{n}$. Hence, T_1 is upper bounded by $4\sigma\sqrt{\frac{\eta}{\nu}}\sqrt{\log n}$.

Bound for T_2 : $|A_i h^{(j)}| \leq \|A_i\|_2 \|h^{(j)}\|_2 \leq \sqrt{\nu n} \beta^* \frac{6\sqrt{k} \|L_R\|_2}{\xi}$ where the bound on $\|h^{(j)}\|_2$ comes from Lemma 4.10.2.

Also, by Lemma 4.10.4, $\xi \equiv \lambda_{k+1} - \lambda_k = \lambda_{k+1} \geq \frac{\nu n}{\eta} \beta_0$. Thus, $|A_i h^{(j)}| \leq 6 \frac{\beta^*}{\beta_0} \eta \sqrt{\frac{k}{\nu n}} \|L_R\|_2$.

In the $1 - \frac{4}{n}$ probability event described in Lemma 4.9.9, we get that

$$|A_i h^{(j)}| \leq 6 \frac{\beta^*}{\beta_0} \eta \sqrt{\frac{k}{\nu}} 4\sigma \sqrt{\log n}.$$

Note that in order to invoke Lemma 4.10.2, we need to satisfy the condition that $\frac{6\sqrt{k} \|L_R\|_2}{\xi} \leq \frac{1}{2}$. Since

$$\begin{aligned} \frac{6\sqrt{k} \|L_R\|_2}{\xi} &\leq \frac{6\sigma\eta\sqrt{k}4\sqrt{\log n}}{\nu\beta_0\sqrt{n}} \\ &\leq \frac{6\sigma\eta}{\nu\beta_0} 4\sqrt{\frac{k \log n}{n}} \end{aligned}$$

and since $\sigma = o\left(\frac{\beta_0}{k} \left(\frac{n}{k \log n}\right)^{1/4}\right)$ under assumption of the theorem, for large enough n , the condition of Lemma 4.10.2 will be satisfied.

Bound for T_3 :

$$|L_{R_i} v^{(j)}| \leq |D_{R_i} v^{(j)}(i)| + |R_i v^{(j)}|$$

We see that $|D_{R_i} v^{(j)}(i)| \leq |D_{R_i}| |v^{(j)}(i)|$. We know that in the same $1 - \frac{4}{n}$ probability event described in $T1$ bound, $|D_{R_i}| \leq 4\sigma\sqrt{n \log n}$. Hence,

$$|D_{R_i}| |v^{(j)}(i)| \leq 4\sigma\sqrt{\frac{\eta}{\nu} \log n}.$$

The second term $|R_i v^{(j)}|$ is trickier to bound. We first expand $v^{(j)}$ in terms of $v^{(1)}, \dots, v^{(k)}$.

$$\begin{aligned} |R_i v^{(j)}| &\leq \left| \sum_{t=1}^k \alpha_t R_i v^{(t)} \right| \\ &\leq \left(\sum_{t=1}^k |\alpha_t| \right) \max_{t=1, \dots, k} |R_i v^{(t)}| \\ &\leq \sqrt{k} \max_{t=1, \dots, k} |R_i v^{(t)}|. \end{aligned}$$

We know

$$R_i v^{(t)} = \frac{1}{\sqrt{|C_t|}} \sum_{l=1}^{|C_t|} R_{il}.$$

By Lemma 4.9.6, we get that $R_i v^{(t)}$ is subgaussian with scale factor σ . Hence, with probability at least $1 - \frac{2}{n}$, uniform across $i = 1, \dots, n$, $\max_{t=1, \dots, k} |R_i v^{(t)}| \leq \sigma\sqrt{6 \log n}$.

Hence, T_3 can be bounded as

$$|D_{R_i} v^{(j)}| + |R_i v^{(j)}| \leq 4\sigma\sqrt{\frac{\eta}{\nu} \log n} + \sigma\sqrt{6k \log n}.$$

Bound for $T4$:

$$\begin{aligned} |R_i h^{(j)}| &\leq \|R_i\|_2 \|h^{(j)}\|_2 \\ &\leq C\sigma\sqrt{n} \frac{6\sqrt{k} \|L_R\|_2}{\xi} \\ &\leq (C\sigma\sqrt{n}) \frac{12\sqrt{kn}\eta\sqrt{4 \log n}}{\nu n \beta_0} \\ &\leq 12C\sigma^2 \frac{\sqrt{k} \eta}{\beta_0 \nu} \sqrt{4 \log n} \end{aligned}$$

where we will assume the $1 - \frac{4}{n}$ probability event described in $T1$ bound.

Bound for $T5$: Recall that since $T5$ appears in the denominator, we need a lower bound for it as opposed to an upper bound.

$$\begin{aligned}
|\mu_j - D_{A_i} - D_{R_i}| &= |D_{A_i} + D_{R_i} - \mu_j| \\
&\geq \left| \frac{\nu n}{\eta} \beta_0 + D_{R_i} - \|L_R\|_2 \right| \\
&\geq \left| \frac{\nu n}{\eta} \beta_0 - 3 \|L_R\|_2 \right| \\
&\geq \frac{\nu n}{\eta} \left| \beta_0 - \underbrace{\sigma \frac{\eta}{\nu n} 4 \sqrt{n \log n}}_{\text{decaying term}} \right|,
\end{aligned}$$

where the third inequality occurs under the same $1 - \frac{4}{n}$ probability event described in $T1$ bound.

Recall that we assume $\sigma = o\left(\frac{\beta_0}{k} \left(\frac{n}{k \log n}\right)^{1/4}\right)$ in the statement of the theorem and under this condition, for large enough n , the decaying term will be less than $\frac{\beta_0}{2}$.

$$|\mu_j - D_{A_i} - D_{R_i}| \geq \frac{\nu n \beta_0}{\eta} \frac{1}{2}.$$

Suppose that both the event described in $T1$ and the event described in $T3$ hold, which happens with probability $1 - \frac{8}{n}$ by union bound, the following bounds hold simultaneously for all $j = 1, \dots, k$.

$$\begin{aligned}
T_1 &\leq 4\sigma \sqrt{\frac{\eta}{\nu}} \sqrt{\log n}, \\
T_2 &\leq 4\sigma \frac{\beta_1}{\beta_0} \eta \sqrt{\frac{k}{\nu}} \sqrt{\log n}, \\
T_3 &\leq 4\sigma \sqrt{\frac{\eta}{\nu} \log n} + \sigma \sqrt{6k \log n}, \\
T_4 &\leq 12C\sigma^2 \frac{\sqrt{k} \eta}{\beta_0 \nu} \sqrt{4 \log n}, \\
T_5 &\geq \frac{\nu n \beta_0}{\eta} \frac{1}{2}.
\end{aligned}$$

Combining everything together, we conclude that, uniformly across all $j = 1, \dots, k$:

$$\|h^{(j)}\|_\infty \leq 12\sigma \sqrt{4 \log n} \frac{2\eta}{\nu n \beta_0} \left[\sqrt{\frac{\eta}{\nu}} + \frac{\beta_1 \eta}{\beta_0} \sqrt{\frac{k}{\nu}} + \sqrt{k} + C\sigma \frac{\sqrt{k} \eta}{\beta_0 \nu} \right].$$

Since we hold β_1 and η to be a constant and $\nu \leq 1$, we see that the last term of the sum dominates the entire sum. We also note that $\nu \geq \frac{1}{k}$ and thus $\frac{1}{\nu} \leq k$.

It is then straightforward to check that under the assumption that $\sigma^2 = o\left(\sqrt{\frac{n}{\log n} \frac{\beta_0^2}{k^{5/2}}}\right)$, then for large enough n , $\|h^{(j)}\|_\infty \leq \sqrt{\frac{1}{8\nu nk}}$.

Embedding of each point onto basis $\{v^{(1)}, \dots, v^{(k)}\}$ is k -dimensional vector with exactly one non-zero coordinate. By the above definition, we can see that if points $p_1, p_2 \in \mathbb{R}^k$ are in the different clusters, then $\|p_1 - p_2\| \geq \sqrt{\frac{2}{\nu n}}$.

Let $v'^{(j)} = \Theta v^{(j)}$ be the transformed orthonormal basis, we will show that the embeddings of points onto the transformed basis maintain the same pair-wise distance. We know that $v'^{(j)} = \sum_j \alpha_{jt} v^{(t)}$ and hence, $v'^{(j)}$ has vector-structure of $(\frac{1}{\sqrt{|C_1|}}\alpha_{j1}, \frac{1}{\sqrt{|C_2|}}\alpha_{j2}, \frac{1}{\sqrt{|C_3|}}\alpha_{j3}, \dots)$ where $\underline{\alpha}_{j1}$ is sub-vector of length $|C_1|$ whose every entry is α_{j1} .

Let $p_1, p_2 \in \mathbb{R}^k$ be two points in the transformed-basis-embedding. Let p_1 be in cluster a and p_2 be in cluster b , then $\|p_1 - p_2\|_2 = \|\frac{1}{\sqrt{|C_a|}}(\alpha_{1a}, \dots, \alpha_{ka}) - \frac{1}{\sqrt{|C_b|}}(\alpha_{1b}, \dots, \alpha_{kb})\|$. Thus, if p_1, p_2 are in the same cluster, $\|p_1 - p_2\| = 0$.

Let $\alpha^a \triangleq (\alpha_{1a}, \dots, \alpha_{ka})$ and $\alpha^b \triangleq (\alpha_{1b}, \dots, \alpha_{kb})$. Then

$$\begin{aligned} & \left\| \frac{1}{\sqrt{|C_a|}}\alpha^a - \frac{1}{\sqrt{|C_b|}}\alpha^b \right\|^2 \\ &= \frac{1}{|C_a|} \|\alpha^a\|^2 - \frac{1}{\sqrt{|C_a||C_b|}} 2\langle \alpha^a, \alpha^b \rangle + \frac{1}{|C_b|} \|\alpha^b\|^2. \end{aligned}$$

Define $k \times k$ matrix M such that $M_{jt} = \alpha_{jt}$. Hence, row j of M contains the linear coefficients of $v'^{(j)}$ in term of basis $\{v^{(1)}, \dots, v^{(k)}\}$. Since $v'^{(j)}$'s are orthonormal, it must be that rows of M are orthonormal and therefore, M must be an isometry and its columns are also orthonormal.

Thus, we get that $\|\alpha^a\| = \|\alpha^b\| = 1$ and $\langle \alpha^a, \alpha^b \rangle = 0$ and that $\|p_1 - p_2\|^2 = \frac{1}{|C_a|} + \frac{1}{|C_b|} \geq \frac{2}{\nu n}$ and that if p_1, p_2 are in different clusters, then $\|p_1 - p_2\| \geq \sqrt{\frac{2}{\nu n}}$.

Let q_1, q_2 be perturbed version of p_1, p_2 , that is, the same points embedded in $(u^{(1)}, \dots, u^{(k)})$ -basis. Since each coordinate of the perturbed vector $u^{(j)}$ can change by at most $\sqrt{\frac{1}{8\nu nk}}$ from $v'^{(j)}$, we get that $\|p_1 - q_1\|_2 \leq \sqrt{\frac{1}{8\nu n}}$ and likewise for $\|p_2 - q_2\|_2$.

If q_1, q_2 are in the same cluster, $\|q_1 - q_2\|_2 \leq \sqrt{\frac{1}{2\nu n}}$ and if q_1, q_2 are in different clusters, $\|q_1 - q_2\|_2 \geq \sqrt{\frac{2}{\nu n}} - \sqrt{\frac{1}{2\nu n}} \geq \sqrt{\frac{1}{2\nu n}}$.

Since the maximum distance between two points in the same cluster is less than minimum distance between two points in different clusters, in our modified k -means procedure, the k chosen cluster centers will be in different clusters and the remaining points will be assigned to the correct clusters. \square

4.11 Proofs of Information Theoretic Limits

4.11.1 Lower Bounds

The proof of our lower bound is an application of Fano's Inequality from the book of Tsybakov [192]:

Theorem 4.11.1. *Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \dots, \theta_M$ such that:*

1. $d(\theta_j, \theta_k) \geq 2s > 0, \forall 0 \leq j < k \leq M.$
2. $P_j \ll P_0, \forall j = 1, \dots, M,$ and

$$\frac{1}{M} \sum_{j=1}^M K(P_j, P_0) \leq \alpha \log M$$

with $0 < \alpha < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \dots, M.$ Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_{\theta}(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right) > 0.$$

We choose the family of parameters Θ as follows. First suppose that $n = 2^\alpha$ and $m = 2^\beta$ for integers $\alpha > \beta.$ Let θ_0 be some hierarchical partitioning of $[n]$ into clusters of size $m.$ Define θ_{jk}^s for $s \in [n/(2m)]$ that swaps that j th element from the left child of cluster s with the k th element from the right child of cluster $s,$ where s can be any of the second level clusters in the hierarchy.

In total there are $n/(2m) \times m^2$ models as there are $n/(2m)$ different second level clusters, and for each there are m choices for the element in the left cluster and m choices for the element in the right cluster. We also need to calculate the Hamming distance between the estimate $\hat{\theta}$ and the true parameter $\theta_0.$ It is clear that $d(\theta_{jk}^s, \theta_{j'k'}^{s'}) \geq 1,$ as long as one of $s \neq s', j \neq j'$ or $k \neq k'$ holds.

Finally we need the KL-divergence between the probability distributions induced by the parameters. Each distribution is Gaussian and in comparison with θ_0 the mean of each distribution differs in at most $4m$ coordinates (inter- and intra- cluster similarities for the two objects that were swapped). In this coordinates, it differs by $\gamma.$ Thus $K(P_{\theta_{jk}^s}, P_{\theta_0}) = \frac{2m\gamma^2}{\sigma^2}.$ With these calculations, we apply Theorem 4.11.1 and arrive at the bound.

For the k -way problem we apply a similar analysis. If there are k clusters, each of size n/k , then we define the family of models as follows. Starting with a base clustering θ_0 group the clusters into pairs, then define θ_{jk}^s as swapping the j th element and the k th element between the pair s of clusters. A similar computation shows that there are $k/2(n/k)^2$ such models. Moreover the $K(P_{\theta_{jk}^s}, P_{\theta_0}) = \frac{2n\gamma^2}{k\sigma^2}$. Applying Theorem 4.11.1 yields the lower bound in Theorem 4.5.3.

4.11.2 Upper Bounds

For the hierarchical upper bound, our algorithm recursively solves the balanced minimum cut problem. When run on n data points, it outputs a set of $n/2$ coordinates (denoted by \widehat{I}) so as to maximize the contrast between the two diagonal blocks (the $\widehat{I}\widehat{I}$ and $\widehat{I}^C\widehat{I}^C$ blocks) and the two off-diagonal blocks (the $\widehat{I}\widehat{I}^C$ and $\widehat{I}^C\widehat{I}$ blocks). For any subcluster (equivalently set of items) C_s , denote the true clusters by I_s and I_s^C and the submatrix formed by these items as W^s .

Define:

$$S(W^s, I_s) = \sum_{i \in I, j \in I} W_{ij}^s + \sum_{i \in I^c, j \in I^c} W_{ij}^s - \sum_{i \in I, j \in I^c} W_{ij}^s - \sum_{i \in I^c, j \in I} W_{ij}^s.$$

At each subcluster C_s , our algorithm exactly minimizes $S(W^s, I_s)$ subject to $|I_s| = |C_s|/2$. To analyze the procedure, it's useful to consider the random variable ζ^s defined as:

$$\zeta_{\widehat{I}}^s = S(W^s, I_s) - S(W^s, \widehat{I}_s).$$

Further, given a set \widehat{I}_s , define the number of indices in which \widehat{I}_s and I_s agree to be a_s .

It's not too hard to see that for a given a_s ,

$$\zeta_{a_s}^s \sim \mathcal{N}\left(8a_s \left(\frac{|C_s|}{2} - a_s\right) \gamma, 16a_s \left(\frac{|C_s|}{2} - a_s\right) \sigma^2\right)$$

At any cluster C_s , the combinatorial procedure succeeds if $\zeta_{\widehat{I}_s}^s \geq 0$ for all $\widehat{I}_s \neq I_s$. A fairly simple application of the union bound shows that the probability of error of the entire combinatorial procedure is bounded by the probability of error across each of the clusters C_s . This probability (by application of the union bound) can be bounded as follows:

$$\begin{aligned} \mathbb{P}_{\text{error}} &\leq \sum_{i=1}^l 2^i \sum_{a=1}^{n/2^i-1} \binom{n/2^i}{a} \binom{n/2^i}{n/2^i-a} \mathbb{P}(\zeta_{n/2^i}^a \leq 0) \\ &\leq \sum_{i=1}^l 2^i \sum_{a=1}^{n/2^i-1} \binom{n/2^i}{a}^2 \exp\left\{\frac{-C_1 a(n/2^i-a)\gamma^2}{\sigma^2}\right\}. \end{aligned}$$

Working with just the inner summation, we break into two segments and get:

$$\begin{aligned}
& \sum_{a=1}^{n/2^{i+1}} \exp \left\{ C_2 a \log n/2^i - \frac{C_1 a (n/2^i - a) \gamma^2}{\sigma^2} \right\} \\
& + \sum_{a=n/2^{i+1}}^{n/2^i-1} \exp \left\{ C_3 (n/2^i - a) \log n/2^i - \frac{C_1 a (n/2^i - a) \gamma^2}{\sigma^2} \right\} \\
\leq & \max_{1 \leq a \leq n/2^{i+1}} \exp \left\{ C'_2 a \left(\log((n/2^i)^2) - \frac{C'_1 n/2^i \gamma^2}{\sigma^2} \right) \right\} \\
& + \max_{n/2^{i+1} \leq a \leq n/2^i} \exp \left\{ C'_3 (n/2^i - a) \left(\log((n/2^i)^2) - \frac{C'_1 n/2^i \gamma^2}{\sigma^2} \right) \right\}.
\end{aligned}$$

Pushing the 2^i term into the exponent, we see that if:

$$\frac{\gamma^2}{\sigma^2} \geq \frac{\log(n^2/2^i)}{C'_1 n/2^i} + \frac{\log(\log_2 n)/\delta}{C'_1 n/2^i}.$$

Each term is smaller than $C \frac{\delta}{\log_2 n}$. Since there are at most $\log_2 n$ terms in the hierarchy, and consequently at most that many terms in the sum, we see that the probability of failure is upper bounded by δ . Of course to recover clusters of size m , we just need to work for all i such that $n/2^i \geq m$. Substituting this into the bound shows that it is sufficient for:

$$\frac{\gamma}{\sigma} \geq \sqrt{\frac{C_1 \log(nm)}{m} + \frac{C_2 \log \log_2 n / \delta}{m}}.$$

The k -way combinatorial algorithm and analysis are similar in spirit to the hierarchical ones. The algorithm will find a set of $m \triangleq n/k$ objects that maximizes the difference between within-cluster and between cluster similarity:

$$\hat{I} = \operatorname{argmax}_{I \subset [n], |I|=m} S(W, I) = \operatorname{argmax}_{I \subset [n], |I|=m} \sum_{i,j \in I} W_{ij} + \sum_{i,j \notin I} W_{ij} - \sum_{i \in I, j \notin I} W_{ij} - \sum_{i \notin I, j \in I} W_{ij}$$

and we would like to show that $\hat{I} = C$ for some cluster $C \in \mathcal{C}^*$ the true clustering. As before it is convenient to analyze the difference between a candidate solution \hat{I} and the true solution:

$$\zeta_{\hat{I}} = \min_{C \in \mathcal{C}^*} S(W, C) - S(W, \hat{I}).$$

If $\zeta_{\hat{I}} > 0$ for all $\hat{I} \notin \mathcal{C}^*$ then we know that the algorithm will certainly pick out one of the true clusters. If we remove those points and repeatedly apply of the algorithm, we will be able to identify all of the clusters.

We need to compute the mean and variance of $\zeta_{\hat{I}}$. If \hat{I} has s_i elements from each cluster and i^* is the index of the ‘‘closest’’ cluster to \hat{I} , then $\mathbb{E}[\zeta_{\hat{I}}] = c_s \gamma$ and $\operatorname{Var}(\zeta_{\hat{I}}) = c_s \sigma^2$ where:

$$c_s = \left(\sum_{i=1}^k 4(m - s_i) s_i + \sum_{i \neq i^*} \sum_{j \neq i, i^*} 2 s_i s_j \right).$$

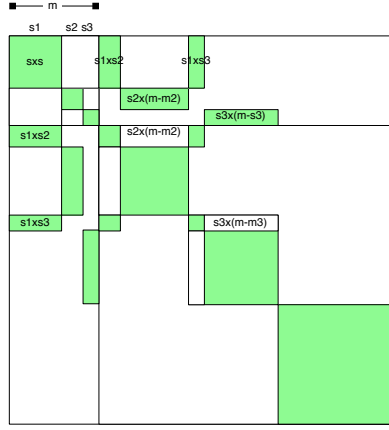


Figure 4.9: Understanding the combinatorial k -way algorithm union bound

See Figure 4.9 for a heuristic explanation of this bound.

We can now proceed to union bound:

$$\mathbb{P}_{\text{error}} \leq \sum_{i^*=1}^k \sum_{s_{i^*}=1}^{m-1} \sum_{s_{-i^*} | \sum_{i \neq i^*} s_i = m - s_{i^*}} \left(\prod_{i=1}^k \binom{m}{s_i} \right) \mathbb{P}[\zeta_s \leq 0].$$

We upper bound the number of ζ_s terms in the following way: First identify a cluster i^* with which to compare (this is the cluster with maximal similarity to \widehat{I} , but we will ignore this constraint). Then select the overlap between C_{i^*} and \widehat{I} , which is certainly no less than 1 but also cannot be more than $m - 1$. Then select the remaining elements of \widehat{I} , ensuring that $|\widehat{I}| = m$. We do this by first assigning the s_i for $i \neq i^*$ and then counting the number of ways to select the elements. We now apply Gaussian tail bounds, and approximate binomial coefficients:

$$\begin{aligned} &\leq \sum_{i^*=1}^k \sum_{s_{i^*}=1}^{m-1} \binom{m}{s_{i^*}} \binom{m - s_{i^*} + k - 2}{m - s_{i^*}} \max_{s_{-i^*} | \sum_{i \neq i^*} s_i = m - s_{i^*}} \exp \left\{ \frac{-c_s \gamma^2}{\sigma^2} + \left(\sum_{i \neq i^*} s_i \right) \log m \right\} \\ &\leq \sum_{i^*=1}^k \sum_{s_{i^*}=1}^{m-1} \binom{m}{s_{i^*}} \max_{s_{-i^*} | \sum_{i \neq i^*} s_i = m - s_{i^*}} \exp \left\{ \frac{-c_s \gamma^2}{\sigma^2} + \left(\sum_{i \neq i^*} s_i \right) (\log m + \log(m - s_{i^*} + k - 2)) \right\}. \end{aligned}$$

Now we break the second sum into two parts, where $s_{i^*} \leq m/2$ and $s_{i^*} > m/2$. In the first case

the whole expression is bounded by:

$$\begin{aligned}
&\leq \frac{km}{2} \max_{s_1, \dots, s_k | \sum_i s_i = m} \exp \left\{ \frac{-c_s \gamma^2}{\sigma^2} + 2 \left(\sum_{i=1}^k s_i \right) (\log(n+k)) \right\} \\
&\leq \frac{km}{2} \exp \left\{ \sum_{i=1}^k s_i \left(\frac{-2m\gamma^2}{\sigma^2} + 2 \log(n+k) \right) \right\} \\
&\leq \frac{km}{2} \exp \left\{ m \left(\frac{-2m\gamma^2}{\sigma^2} + 2 \log(n+k) \right) \right\}.
\end{aligned}$$

Here to arrive at the second line, we substituted in for c_s and noticed that $(m - s_i) \geq m/2$ for all i here. This expression is smaller than $\delta/2$ when:

$$\frac{\gamma}{\sigma} \geq \sqrt{\frac{\log(n+k)}{m} + \frac{\log(km/4\delta)}{m^2}}.$$

For the second case, the whole expression is bounded by:

$$\begin{aligned}
&\leq k \sum_{s_{i^*} = m/2}^{m-1} \max_{s_{-i^*} | \sum_{i \neq i^*} s_i = m - s_{i^*}} \exp \left\{ \frac{-c_s \gamma^2}{\sigma^2} + (m - s_{i^*}) \log m + \sum_{i \neq i^*} s_i (2 \log(n+k)) \right\} \\
&\leq k \sum_{s_{i^*} = m/2}^{m-1} \exp \left\{ (m - s_{i^*}) \left(\frac{-2m\gamma^2}{\sigma^2} + \log m \right) + \sum_{i \neq i^*} s_i \left(\frac{-2m\gamma^2}{\sigma^2} + 2 \log(n+k) \right) \right\} \\
&\leq k \sum_{s_{i^*} = m/2}^{m-1} \exp \left\{ 2(m - s_{i^*}) \left(\frac{-2m\gamma^2}{\sigma^2} + 2 \log(n+k) \right) \right\}.
\end{aligned}$$

In the second line we introduced c_s noting that $s_{i^*} \geq m/2$ and $(m - s_i) \geq m/2$ for $i \neq i^*$. Now if the term in the exponential is negative, it is maximized when $s_{i^*} = m - 1$ in this case we have:

$$\leq \frac{km}{2} \exp \left\{ 2 \left(\frac{-2m\gamma^2}{\sigma^2} + 2 \log(n+k) \right) \right\}.$$

To make this smaller than $\delta/2$ we require:

$$\frac{\gamma}{\sigma} \geq \sqrt{\frac{\log(n+k)}{m} + \frac{\log(km/4\delta)}{2m}}.$$

With both of these bounds, the total probability is smaller than δ . When $m = n/k$ the bound on γ/σ is met when:

$$\sigma = o \left(\gamma \sqrt{\frac{n}{k \log(n/\delta)}} \right).$$

4.11.3 McSherry's Algorithm

We obtain Theorem 4.5.4 via a slightly modified version of Theorem 12 of McSherry [138].

Theorem 4.11.2. *With probability at least $1 - \delta$, we have for all u :*

$$\|A_u - \widehat{W}_u\|_2 \leq \gamma_1 + \gamma_2$$

where

$$\gamma_1 \leq C_1 \sigma \sqrt{nk/s}, \gamma_2 \leq C_2 \sigma \sqrt{k \log(n/\delta)}$$

where s is a lower bound on the cluster size.

First we consider the implications of the theorem and then present its proof. When we have gap γ , for any two points u, v in different clusters we have:

$$\|A_u - A_v\|_2 \geq \sqrt{2s}\gamma.$$

In particular, if:

$$\sqrt{2s}\gamma \geq 4C_1 \sigma \sqrt{nk/s} + 4C_2 \sigma \sqrt{k \log(n/\delta)}$$

then the algorithm succeeds in recovering the clusters, since it is straightforward to see that every column in \widehat{W} is closer to every other column in its own cluster than *any* column in any other cluster. Taking $s = \Theta(n/k)$ we get that if:

$$\gamma \geq C_1 \frac{\sigma k \sqrt{nk}}{n} + C_2 \sigma k \sqrt{\frac{\log(n/\delta)}{n}}$$

for slightly modified constants, we succeed in recovering the clusters. This establishes Theorem 4.5.4

Proof of Lemma 4.11.2. The proof will show for any u ,

$$\|P_{W_1} W_{2u} - A_{2u}\|_2 \leq \gamma_1 \text{ and } \|P_{W_1} (A_{2u} - W_{2u})\|_2 \leq \gamma_2$$

where the subscript u denotes the u^{th} column of the matrix. Combining, these two with the identical proof for the other partition, and using triangle inequality we will arrive at the final theorem. Consider,

$$\|(I - P_{W_1})A_2\|_2 = \|(I - P_{W_1})A_1\|_2 = \|(I - P_{W_1})W_1 - (I - P_{W_1})(W_1 - A_1)\|_2 \leq 2\|W_1 - A_1\|_2.$$

The first equality follows because by our exact bisection assumption A_1 and A_2 can be taken to be identical. The inequality follows from two observations.

$$\|(I - P_{W_1})W_1\|_2 = \|W_1 - P_{W_1}W_1\|_2 \leq \|W_1 - A_1\|_2,$$

which holds since the left side of the inequality is the $k + 1^{\text{th}}$ eigenvalue of W_1 and A_1 is a rank- k matrix. The second observation is that

$$\|(I - P_{W_1})(W_1 - A_1)\|_2 \leq \|I - P_{W_1}\|_2 \|W_1 - A_1\|_2 \leq \|W_1 - A_1\|_2$$

since P_{W_1} is a projection matrix all of its eigenvalues are positive and bounded by 1.

Now, note that $(I - P_{W_1})A_2$ is of rank at most $2k$ and for any column u there are at least $s/2$ identical columns in $(I - P_{W_1})A_2$. From this we get that for any u ,

$$\|A_{2u} - P_{W_1}A_{2u}\| \leq \frac{\|(I - P_{W_1})A_2\|_F}{\sqrt{s/2}} \leq 4\sqrt{\frac{k}{s}}\|W_1 - A_1\|_2 \leq C_1\sigma\sqrt{\frac{nk}{s}} \equiv \gamma_1$$

with probability at least $\delta/2$ using the operator norm bound on $W_1 - A_1$. Now,

$$\|P_{W_1}(A_{2u} - W_{2u})\|_2 = \sqrt{\sum_{j=1}^k ((A_{2u} - W_{2u})^T P_{W_{1j}})^2}.$$

Noting that $P_{W_{1j}}$ is a unit vector independent of $(A_{2u} - W_{2u})$, each term in this sum is sub-Gaussian with scale factor at most σ . To make a guarantee for any u we will also combine this with a union bound.

From this, a calculation shows that with probability at least $1 - \delta$ we have,

$$\|P_{W_1}(A_{2u} - W_{2u})\|_2 \leq \sqrt{kt}$$

where $t = C_2\sigma\sqrt{\log(n/\delta)}$. This is just γ_2 . □

Chapter 5

Minimax Localization of Bi-Clusters in Large Noisy Matrices

In this chapter we consider the problem of identifying a sparse set of relevant columns and rows in a large data matrix with highly corrupted entries. This problem of identifying groups from a collection of bipartite variables such as proteins and drugs, biological species and gene sequences, malware and signatures, etc is commonly referred to as biclustering or co-clustering. Despite its great practical relevance, and although several ad-hoc methods are available for bi-clustering, theoretical analysis of the problem is largely non-existent.

We consider bi-clustering in a framework that is also closely related to structured multiple hypothesis testing [2, 11, 12], an area of statistics that has recently witnessed a flurry of activity.

In this chapter we make the following contributions

1. We prove lower bounds on the minimum signal strength needed for successful recovery of a bi-cluster as a function of the noise variance, size of the matrix and bi-cluster of interest.
2. We show that a combinatorial procedure based on the scan statistic achieves this optimal limit.
3. We characterize the SNR required by several computationally tractable procedures for bi-clustering including element-wise thresholding, column/row average thresholding and a convex relaxation approach to sparse singular vector decomposition.

5.1 Introduction

Bi-clustering is the problem of identifying a (typically) sparse set of relevant columns and rows in a large, noisy data matrix. This problem along with the first algorithm to solve it were proposed by Hartigan [89] as a way to directly cluster data matrices to produce clusters with greater

interpretability. Bi-clustering routinely arises in several applications such as discovering groups of proteins and drugs that interact with each other [131], learning phylogenetic relationships between different species based on alignments of snippets of their gene sequences [201], identifying malware that have similar signatures [25] and identifying groups of users with similar tastes for commercial products [194]. In these applications, the data matrix is often indexed by (object, feature) pairs and the goal is to identify clusters in this set of bipartite variables.

In standard clustering problems, the goal is only to identify meaningful groups of objects and the methods typically use the entire feature vector to define a notion of similarity between the objects. Bi-clustering can be thought of as high-dimensional clustering where only a subset of the features are relevant for identifying similar objects, and the goal is to identify not only groups of objects that are similar, but also which features are relevant to the clustering task. Consider, for instance gene expression data where the objects correspond to genes, and the features correspond to their expression levels under a variety of experimental conditions. Our present understanding of biological systems leads us to expect that subsets of genes will be co-expressed only under a small number of experimental conditions. Although, pairs of genes are not expected to be similar under *all* experimental conditions it is critical to be able to discover local expression patterns, which can for instance correspond to joint participation in a particular biological pathway or process. Thus, while clustering aims to identify *global* structure in the data, bi-clustering take a more *local* approach by jointly clustering *both* objects and features.

Prevalent techniques for finding biclusters are typically heuristic procedures with little or no theoretical underpinning. In order to study, understand and compare bi-clustering algorithms we consider a simple theoretical model of bi-clustering [122, 123, 182]. This model is akin to the spiked covariance model of Johnstone [102] widely used in the study of PCA in high-dimensions.

We will focus on the following simple observation model for the matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$:

$$\mathbf{A} = \beta \mathbf{u} \mathbf{v}' + \mathbf{\Delta} \tag{5.1}$$

where $\mathbf{\Delta} = \{\Delta_{ij}\}_{i \in [n_1], j \in [n_2]}$ is a random matrix whose entries are i.i.d. $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$ known, $\mathbf{u} = \{u_i : i \in [n_1]\}$ and $\mathbf{v} = \{v_i : i \in [n_2]\}$ are unknown deterministic unit vectors in \mathbb{R}^{n_1} and \mathbb{R}^{n_2} , respectively, and $\beta > 0$ is a constant. To simplify the presentation, we assume that $\mathbf{u} \propto \{0, 1\}^{n_1}$ and $\mathbf{v} \propto \{0, 1\}^{n_2}$. Let $K_1 = \{i : u_i \neq 0\}$ and $K_2 = \{i : v_i \neq 0\}$ be the sets indexing the non-zero components of the vectors \mathbf{u} and \mathbf{v} , respectively. We assume that \mathbf{u} and \mathbf{v} are sparse, that is, $k_1 := |K_1| \ll n_1$ and $k_2 := |K_2| \ll n_2$. While the sets (K_1, K_2) are unknown, we assume that their cardinalities are known. Notice that the magnitude of the signal for all the coordinates in the bicluster $K_1 \times K_2$ is $\frac{\beta}{\sqrt{k_1 k_2}}$. The parameter β measures the strength of the signal, and is the key quantity we will be studying.

We focus on the case of a single bicluster that appears as an elevated sub-matrix of size $k_1 \times k_2$ with signal strength β embedded in a large $n_1 \times n_2$ data matrix with entries corrupted by additive Gaussian noise with variance σ^2 . Under this model, the bi-clustering problem is formulated as the problem of estimating the sets K_1 and K_2 , based on a single noisy observation \mathbf{A} of the unknown signal matrix $\beta \mathbf{u} \mathbf{v}'$. Bi-clustering is most subtle when the matrix is large with several

irrelevant variables, the entries are highly noisy, and the bicluster is small as defined by a sparse set of rows/columns. We provide a sharp characterization of tuples of $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which it is possible to recover the bicluster and study several common methods and establish the regimes under which they succeed.

We establish minimax lower and upper bounds for the following class of models. Let

$$\Theta(\beta_0, k_1, k_2) := \{(\beta, K_1, K_2) : \beta \geq \beta_0, |K_1| = k_1, K_1 \subset [n_1], |K_2| = k_2, K_2 \subset [n_2]\} \quad (5.2)$$

be a set of parameters. For a parameter $\theta \in \Theta$, let \mathbb{P}_θ denote the joint distribution of the entries of $\mathbf{A} = \{a_{ij}\}_{i \in [n_1], j \in [n_2]}$, whose density with respect to the Lebesgue measure is

$$\prod_{ij} \mathcal{N}(a_{ij}; \beta(k_1 k_2)^{-1/2} \mathbb{I}\{i \in K_1, j \in K_2\}, \sigma^2), \quad (5.3)$$

where the notation $\mathcal{N}(z; \mu, \sigma^2)$ denotes the distribution $p(z) \sim \mathcal{N}(\mu, \sigma^2)$ of a Gaussian random variable with mean μ and variance σ^2 , and \mathbb{I} denotes the indicator function.

We derive a lower bound that identifies tuples of $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which we can recover the true bi-clustering from a noisy high dimensional matrix. We show that a combinatorial procedure based on the scan statistic achieves the minimax optimal limits, however it is impractical as it requires enumerating all possible sub-matrices of a given size in a large matrix. We analyze the scalings (i.e. the relation between β and $(n_1, n_2, k_1, k_2, \sigma^2)$) under which some computationally tractable procedures for bi-clustering including element-wise thresholding, column/row average thresholding and sparse singular vector decomposition (SSVD) succeed with high probability.

We consider the detection of both small and large biclusters of weak activation, and show that at the minimax scaling the problem is surprisingly subtle (e.g., even detecting big clusters is quite hard).

In Table 5.1, we describe our main findings and compare the scalings under which the various algorithms succeed.

Algorithm	Combinatorial	Thresholding	Row/Column Averaging	Sparse SVD
SNR scaling	Minimax	Weak	Intermediate	Weak
Bicluster size	Any Theorem 5.3.1	Any Theorem 5.4.1	$(n_1^{1/2+\alpha} \times n_2^{1/2+\alpha}), \alpha \in (0, 1/2)$ Theorem 5.4.2	Any Theorem 5.4.3

Table 5.1: Bi-clustering

Where the scalings are,

1. **Minimax:** $\beta \sim \sigma \max \left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)} \right)$.
2. **Weak:** $\beta \sim \sigma \max \left(\sqrt{k_1 k_2 \log(n_1 - k_1)}, \sqrt{k_1 k_2 \log(n_2 - k_2)} \right)$.

3. Intermediate (for large clusters): $\beta \sim \sigma \max \left(\frac{\sqrt{k_1 k_2 \log(n_1 - k_1)}}{n_2^\alpha}, \frac{\sqrt{k_1 k_2 \log(n_2 - k_2)}}{n_1^\alpha} \right)$.

Element-wise thresholding does not take advantage of any structure in the data matrix and hence does not achieve the minimax scaling for any bicluster size. If the clusters are big enough row/column averaging performs better than element-wise thresholding since it can take advantage of structure. We also study a convex relaxation for sparse SVD, based on the DSPCA algorithm proposed by d’Aspremont et al. [55] that encourages the singular vectors of the matrix to be supported over a sparse set of variables. However, despite the increasing popularity of this method, we show that it is only guaranteed to yield a sparse set of singular vectors when the SNR is quite high, equivalent to element-wise thresholding, and fails for stronger scalings of the SNR.

5.1.1 Related work

Due to its practical importance and difficulty bi-clustering has attracted considerable attention (for some recent surveys see the papers [36, 134, 150, 184]). Broadly algorithms for bi-clustering can be categorized as either score-based searches, or spectral algorithms. Many of the proposed algorithms for identifying relevant clusters are based on heuristic searches whose goal is to identify large average sub-matrices or sub-matrices that are well fit by a two-way ANOVA model. Sun and Nobel [182] provide some statistical backing for these exhaustive search procedures. In particular, they show how to construct a test via exhaustive search to distinguish when there is a small sub-matrix of weak activation from the “null” case when there is no bicluster.

The premise behind the spectral algorithms is that if there was a sub-matrix embedded in a large matrix, then this sub-matrix could be identified from the left and right singular vectors of \mathbf{A} . In the case when exactly one of \mathbf{u} and \mathbf{v} is random, the model Eq. 5.1 can be related to the spiked covariance model of Johnstone [102]. In the case when \mathbf{v} is random, the matrix \mathbf{A} has independent columns and dependent rows. Therefore, $\mathbf{A}'\mathbf{A}$ is a spiked covariance matrix and it is possible to use the existing theoretical results on the first eigenvalue to characterize the left singular vector of \mathbf{A} . A lot of recent work has dealt with estimation of sparse eigenvectors of $\mathbf{A}'\mathbf{A}$, see for example the papers [6, 103, 169, 204, 212]. For bi-clustering applications, the assumption that exactly one \mathbf{u} or \mathbf{v} is random, is not justifiable, therefore, theoretical results for the spiked covariance model do not translate directly. Singular vectors of the model Eq. 5.1 have been analyzed by Onatski [148], improving on earlier results of Bai [15]. These results however are asymptotic and do not consider the case when \mathbf{u} and \mathbf{v} are sparse.

Our setup for the bi-clustering problem also falls in the framework of structured normal means multiple hypothesis testing problems, where for each entry in the matrix the hypotheses are that the entry has mean 0 versus an elevated mean. The presence of a bicluster (sub-matrix) however imposes structure on which elements are elevated concurrently. Recently, several papers have investigated the structured normal means setting for ordered domains. For example, Arias-Castro et al. [13] consider the detection of elevated intervals and other parametric structures along an ordered line or grid, Arias-Castro et al. [12] consider detection of elevated connected paths in

tree and lattice topologies, Arias-Castro et al. [11] consider nonparametric cluster structures in a regular grid. In addition, Addario-Berry et al. [2] consider testing of different elevated structures in a general but known graph topology. Our setup for the bi-clustering problem requires identification of an elevated submatrix in an *unordered* matrix. At a high level, all these results suggest that it is possible to leverage the structure to improve the SNR threshold at which the hypothesis testing problem is feasible. However, computationally efficient procedures that achieve the minimax SNR thresholds are only known for a few of these problems. Our results for bi-clustering have a similar flavor, in that the minimax threshold requires a combinatorial procedure whereas the computationally efficient procedures we investigate are often sub-optimal.

The rest of this chapter is organized as follows. In Section 5.2, we provide a lower bound on the minimum signal strength needed for successfully identifying the bicluster. Section 5.3 presents a combinatorial procedure which achieves the lower bound and hence is minimax optimal. We investigate some computationally efficient procedures in Section 5.4. Simulation results are presented in Section 5.5 and we conclude in Section 5.6. All proofs are deferred to Section 5.7.

5.2 Lower bound

In this section, we derive a lower bound for the problem of identifying the correct bicluster, indexed by K_1 and K_2 , in model Eq. 5.1. In particular, we derive conditions on $(\beta, n_1, n_2, k_1, k_2, \sigma^2)$ under which any method is going to make an error when estimating the correct cluster. Intuitively, if either the signal-to-noise ratio β/σ or the cluster size is small, the minimum signal strength needed will be high since it is harder to distinguish the bicluster from the noise.

Theorem 5.2.1. *Let $\alpha \in (0, \frac{1}{8})$ and*

$$\begin{aligned} \beta_{\min} &= \beta_{\min}(n_1, n_2, k_1, k_2, \sigma) \\ &= \sigma\sqrt{\alpha} \max \left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)}, \sqrt{\frac{k_1 k_2 \log(n_1 - k_1)(n_2 - k_1)}{k_1 + k_2 - 1}} \right). \end{aligned} \quad (5.4)$$

Then for all $\beta_0 \leq \beta_{\min}$,

$$\inf_{\Phi} \sup_{\theta \in \Theta(\beta_0, k_1, k_2)} \mathbb{P}_{\theta}[\Phi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))] \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \frac{2\alpha}{\log M} \right) \xrightarrow{n_1, n_2 \rightarrow \infty} 1 - 2\alpha, \quad (5.5)$$

where $M = \min(n_1 - k_1, n_2 - k_2)$, $\Theta(\beta_0, k_1, k_2)$ is given in Eq. 5.2 and the infimum is over all measurable maps $\Phi : \mathbb{R}^{n_1 \times n_2} \mapsto 2^{[n_1]} \times 2^{[n_2]}$.

The result can be interpreted in the following way: for any biclustering procedure Φ , if $\beta_0 \leq \beta_{\min}$, then there exists some element in the model class $\Theta(\beta_0, k_1, k_2)$ such that the probability of incorrectly identifying the sets K_1 and K_2 is bounded away from zero.

The proof is based on a standard technique described in Chapter 2.6 of the book [192]. We start by identifying a subset of parameter tuples that are hard to distinguish. Once a suitable

finite set is identified, tools for establishing lower bounds on the error in multiple-hypothesis testing can be directly applied. These tools only require computing the Kullback-Leibler (KL) divergence between two distributions \mathbb{P}_{θ_1} and \mathbb{P}_{θ_2} , which in the case of model Eq. 5.1 are two multivariate normal distributions. These constructions and calculations are described in detail in the Section 5.7.

5.3 Minimax optimal combinatorial procedure

We now investigate a combinatorial procedure achieving the lower bound of Theorem 5.2.1, in the sense that, for any $\theta \in \Theta(\beta_{\min}, k_1, k_2)$, the probability of recovering the true bicluster (K_1, K_2) tends to one as n_1 and n_2 grow unbounded. This scan procedure consists in enumerating all possible pairs of subsets of the row and column indexes of size k_1 and k_2 , respectively, and choosing the one whose corresponding submatrix has the largest overall sum. In detail, for an observed matrix \mathbf{A} and two candidate subsets $\tilde{K}_1 \subset [n_1]$ and $\tilde{K}_2 \subset [n_2]$, we define the associated score $\mathcal{S}(\tilde{K}_1, \tilde{K}_2) := \sum_{i \in \tilde{K}_1} \sum_{j \in \tilde{K}_2} a_{ij}$. The estimated bicluster is the pair of subsets of sizes k_1 and k_2 achieving the highest score:

$$\Psi(\mathbf{A}) := \underset{(\tilde{K}_1, \tilde{K}_2)}{\operatorname{argmax}} \mathcal{S}(\tilde{K}_1, \tilde{K}_2) \quad \text{subject to} \quad |\tilde{K}_1| = k_1, |\tilde{K}_2| = k_2. \quad (5.6)$$

The following theorem determines the signal strength β needed for the decoder Ψ to find the true bicluster.

Theorem 5.3.1. *Let $\mathbf{A} \sim \mathbb{P}_\theta$ with $\theta \in \Theta(\beta, k_1, k_2)$ and assume that $k_1 \leq n_1/2$ and $k_2 \leq n_2/2$. If*

$$\beta \geq 4\sigma \max \left(\sqrt{k_1 \log(n_1 - k_1)}, \sqrt{k_2 \log(n_2 - k_2)}, \sqrt{\frac{2k_1 k_2 \log(n_1 - k_1)(n_2 - k_2)}{k_1 + k_2}} \right) \quad (5.7)$$

then $\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq 2[(n_1 - k_1)^{-1} + (n_2 - k_2)^{-1}]$ where Ψ is the decoder defined in Eq. 5.6.

Comparing to the lower bound in Theorem 5.2.1, we observe that the combinatorial procedure using the decoder Ψ that looks for all possible clusters and chooses the one with largest score achieves the lower bound up to constants. Unfortunately, this procedure is not practical for data sets commonly encountered in practice, as it requires enumerating all $\binom{n_1}{k_1} \binom{n_2}{k_2}$ possible submatrices of size $k_1 \times k_2$. The combinatorial procedure requires the signal to be positive, but not necessarily constant throughout the bicluster. In fact it is easy to see that provided the average signal in the bicluster is larger than that stipulated by the theorem this procedure succeeds with high probability irrespective of how the signal is distributed across the bicluster. Finally, we remark that the estimation of the cluster is done under the assumption that k_1 and k_2 are known. Establishing minimax lower bounds and a procedure that adapts to unknown k_1 and k_2 is an open problem.

5.4 Computationally efficient biclustering procedures

In this section we investigate the performance of various procedures for biclustering, that, unlike the optimal scan statistic procedure studied in the previous section, are computationally tractable. For each of these procedures however, computational ease comes at the cost of suboptimal performance: recovery of the true bicluster is only possible if the β is much larger than the minimax signal strength of Theorem 5.2.1.

5.4.1 Element-wise thresholding

The simplest procedure that we analyze is based on element-wise thresholding. The bicluster is estimated as

$$\Psi_{\text{thr}}(\mathbf{A}, \tau) := \{(i, j) \in [n_1] \times [n_2] : |a_{ij}| \geq \tau\} \quad (5.8)$$

where $\tau > 0$ is a parameter. The following theorem characterizes the signal strength β required for the element-wise thresholding to succeed in recovering the bicluster.

Theorem 5.4.1. *Let $\mathbf{A} \sim \mathbb{P}_\theta$ with $\theta \in \Theta(\beta, k_1, k_2)$ and fix $\delta > 0$. Set the threshold τ as*

$$\tau = \sigma \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}.$$

If

$$\beta \geq \sqrt{k_1 k_2} \sigma \left(\sqrt{2 \log \frac{k_1 k_2}{\delta}} + \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}} \right)$$

then $\mathbb{P}[\Psi_{\text{thr}}(\mathbf{A}, \tau) \neq K_1 \times K_2] = o(\delta/(k_1 k_2))$.

Comparing Theorem 5.4.1 with the lower bound in Theorem 5.2.1, we observe that the signal strength β needs to be $\mathcal{O}(\max(\sqrt{k_1}, \sqrt{k_2}))$ larger than the lower bound. This is not surprising, since the element-wise thresholding is not exploiting the structure of the problem, but is assuming that the large elements of the matrix \mathbf{A} are positioned randomly. From the proof it is not hard to see that this upper bound is tight up to constants, i.e. if

$$\beta \leq c \sqrt{k_1 k_2} \sigma \left(\sqrt{2 \log \frac{k_1 k_2}{\delta}} + \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}} \right)$$

for a small enough constant c then thresholding will no longer recover the bi-cluster with probability at least $1 - \delta$. It is also worth noting that thresholding neither requires the signal in the bi-cluster to be constant nor positive provided it is larger in magnitude, at every entry, than the threshold specified in the theorem.

5.4.2 Row/Column averaging

Next, we analyze another a procedure based on column and row averaging. When the bicluster is large this procedure exploits the structure of the problem and outperforms the simple element-wise thresholding and the sparse SVD, which is discussed in the following section. The averaging procedure works only well if the bicluster is “large”, as specified below, since otherwise the row or column average is dominated by the noise.

More precisely, the averaging procedure computes the average of each row and column of \mathbf{A} and outputs the k_1 rows and k_2 columns with the largest average. Let $\{r_{r,i}\}_{i \in [n_1]}$ and $\{r_{c,j}\}_{j \in [n_2]}$ denote the positions of rows and columns when they are ordered according to row and column averages in descending order. The bicluster is estimated then as

$$\Psi_{\text{avg}}(\mathbf{A}) := \{i \in [n_1] : r_{r,i} \leq k_1\} \times \{j \in [n_2] : r_{c,j} \leq k_2\}. \quad (5.9)$$

The following theorem characterizes the signal strength β required for the averaging procedure to succeed in recovering the bicluster.

Theorem 5.4.2. *Let $\mathbf{A} \sim \mathbb{P}_\theta$ with $\theta \in \Theta(\beta, k_1, k_2)$. If $k_1 = \Omega(n_1^{1/2+\alpha})$ and $k_2 = \Omega(n_2^{1/2+\alpha})$, where $\alpha \in (0, 1/2)$ is a constant and,*

$$\beta \geq 4\sigma \max \left(\frac{\sqrt{k_1 k_2 \log(n_1 - k_1)}}{n_2^\alpha}, \frac{\sqrt{k_1 k_2 \log(n_2 - k_2)}}{n_1^\alpha} \right)$$

then $\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq [n_1^{-1} + n_2^{-1}]$.

Comparing to Theorem 5.4.1, we observe that the averaging requires lower signal strength than the element-wise thresholding when the bicluster is large, that is, $k_1 = \Omega(\sqrt{n_1})$ and $k_2 = \Omega(\sqrt{n_2})$. Unless both $k_1 = \mathcal{O}(n_1)$ and $k_2 = \mathcal{O}(n_2)$, the procedure does not achieve the lower bound of Theorem 5.2.1, however, the procedure is simple and computationally efficient. It is also not hard to show that this theorem is sharp in its characterization of the averaging procedure. Further, unlike thresholding, averaging requires the signal to be positive in the bicluster.

It is interesting to note that a large bicluster can also be identified without assuming the normality of the noise matrix Δ . This non-parametric extension is based on a simple sign-test, and the details are provided in Section 5.7.

5.4.3 Sparse singular value decomposition (SSVD)

An alternate way to estimate K_1 and K_2 would be based on the singular value decomposition (SVD), i.e. finding $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ that maximize $\langle \tilde{\mathbf{u}}, \mathbf{A} \tilde{\mathbf{v}} \rangle$, and then threshold the elements of $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$. Unfortunately, such a method would perform poorly when the signal β is weak and the dimensionality is high, since, due to the accumulation of noise, $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are poor estimates of \mathbf{u} and \mathbf{v} and do not exploit the fact that \mathbf{u} and \mathbf{v} are sparse.

In fact, it is now well understood (see for example the paper of Benaych-Georges and Rao Nadakuditi [26]) that SVD is strongly inconsistent when the signal strength is weak, i.e. $\angle(\tilde{\mathbf{u}}, \mathbf{u}) \rightarrow \pi/2$ (and similarly for \mathbf{v}) almost surely. See the paper of Sun and Nobel [182] for a clear exposition and discussion of this inconsistency in the SVD setting.

To properly exploit the sparsity in the singular vectors, it seems natural to impose a cardinality constraint to obtain a sparse singular vector decomposition (SSVD):

$$\max_{\mathbf{u} \in \mathbf{S}^{n_1-1}, \mathbf{v} \in \mathbf{S}^{n_2-1}} \langle \mathbf{u}, \mathbf{A}\mathbf{v} \rangle \quad \text{subject to} \quad \|\mathbf{u}\|_0 \leq k_1, \|\mathbf{v}\|_0 \leq k_2,$$

which can be further rewritten as

$$\max_{\mathbf{Z} \in \mathbb{R}^{n_2 \times n_1}} \text{tr} \mathbf{A}\mathbf{Z} \quad \text{subject to} \quad \mathbf{Z} = \mathbf{v}\mathbf{u}', \|\mathbf{u}\|_2 = 1, \|\mathbf{v}\|_2 = 1, \|\mathbf{u}\|_0 \leq k_1, \|\mathbf{v}\|_0 \leq k_2. \quad (5.10)$$

The above problem is non-convex and computationally intractable.

Inspired by the convex relaxation methods for sparse principal component analysis proposed by d'Aspremont et al. [55], we consider the following relaxation the SSVD:

$$\max_{\mathbf{X} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}} \text{tr} \mathbf{A}\mathbf{X}^{21} - \lambda \mathbf{1}'|\mathbf{X}^{21}| \mathbf{1} \quad \text{subject to} \quad \mathbf{X} \succeq \mathbf{0}, \text{tr} \mathbf{X}^{11} = 1, \text{tr} \mathbf{X}^{22} = 1, \quad (5.11)$$

where \mathbf{X} is the block matrix

$$\begin{bmatrix} \mathbf{X}^{11} & \mathbf{X}^{12} \\ \mathbf{X}^{21} & \mathbf{X}^{22} \end{bmatrix}$$

with the block \mathbf{X}^{21} corresponding to \mathbf{Z} in Eq. 5.10. If the optimal solution $\hat{\mathbf{X}}$ is of rank 1, then, necessarily, $\hat{\mathbf{X}} = \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix} (\hat{\mathbf{u}}' \hat{\mathbf{v}}')$. Based on the sparse singular vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$, we estimate the bicluster as

$$\hat{K}_1 = \{j \in [n_1] : \hat{u}_j \neq 0\} \quad \text{and} \quad \hat{K}_2 = \{j \in [n_2] : \hat{v}_j \neq 0\}. \quad (5.12)$$

The user defined parameter λ controls the sparsity of the solution $\hat{\mathbf{X}}^{21}$, and, therefore, provided the solution is of rank one, it also controls the sparsity of the vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ and of the estimated bicluster.

The following theorem provides *sufficient* conditions for the solution $\hat{\mathbf{X}}$ to be rank one and to recover the bicluster.

Theorem 5.4.3. *Consider the model in Eq. 5.1. Assume $k_1 \asymp k_2$ and $k_1 \leq n_1/2$ and $k_2 \leq n_2/2$. If*

$$\beta \geq 2\sigma \sqrt{k_1 k_2 \log(n_1 - k_1)(n_2 - k_2)} \quad (5.13)$$

then the solution $\hat{\mathbf{X}}$ of the optimization problem in Eq. 5.11 with $\lambda = \frac{\beta}{2\sqrt{k_1 k_2}}$ is of rank 1 with probability $1 - \mathcal{O}(k_1^{-1})$. Furthermore, we have that $(\hat{K}_1, \hat{K}_2) = (K_1, K_2)$ with probability $1 - \mathcal{O}(k_1^{-1})$.

It is worth noting that SSVD correctly recovers *signed* vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ under this signal strength. In particular, the procedure works even if the u and v in Equation 5.1 are signed.

The following theorem establishes *necessary* conditions for the SSVD to have a rank 1 solution that correctly identifies the bicluster.

Theorem 5.4.4. Consider the model in Eq. 5.1. Fix $c \in (0, 1/2)$. Assume that $k_1 \asymp k_2$ and $k_1 = o(n^{1/2-c})$ and $k_2 = o(n_2^{1/2-c})$. If

$$\beta \leq 2\sigma \sqrt{ck_1 k_2 \log \max(n_1 - k_1, n_2 - k_2)}, \quad (5.14)$$

with $\lambda = \frac{\beta}{2\sqrt{k_1 k_2}}$ then the optimization problem Eq. 5.11 does not have a rank 1 solution that correctly recovers the sparsity pattern with probability at least $1 - \mathcal{O}(\exp(-(\sqrt{k_1} + \sqrt{k_2})^2))$ for sufficiently large n_1 and n_2 .

From Theorem 5.4.4 observe that the sufficient conditions of Theorem 5.4.3 are sharp. In particular, the two theorems establish that the SSVD does not establish the lower bound given in Theorem 5.2.1. The signal strength needs to be of the same order as for the element-wise thresholding, which is somewhat surprising since from the formulation of the SSVD optimization problem it seems that the procedure uses the structure of the problem. From numerical simulations in Section 5.5 we observe that although SSVD requires the same scaling as thresholding, it consistently performs slightly better at a fixed signal strength.

5.5 Simulation results

We test the performance of the three computationally efficient procedures on synthetic data: thresholding, averaging and sparse SVD. For sparse SVD we use an implementation posted online by d’Aspremont et al. [55]. We generate data from Eq. 5.1 with $n = n_1 = n_2$, $k = k_1 = k_2$, $\sigma^2 = 1$ and $\mathbf{u} = \mathbf{v} \propto (\mathbf{1}'_k, \mathbf{0}'_{n-k})'$. For each algorithm we plot the Hamming fraction (i.e. the Hamming distance between $\hat{\mathbf{s}}_u$ and \mathbf{s}_u rescaled to be between 0 and 1) against the rescaled sample size. In each case we average the results over 50 runs.

For thresholding and sparse SVD the rescaled scaling (x-axis) is $\frac{\beta}{k\sqrt{\log(n-k)}}$ and for averaging the rescaled scaling (x-axis) is $\frac{\beta n^\alpha}{k\sqrt{\log(n-k)}}$. We observe that there is a sharp threshold between success and failure of the algorithms, and the curves show good agreement with our theory.

The vertical line shows the point after which successful recovery happens for all values of n . We can make a direct comparison between thresholding and sparse SVD (since the curves are identically rescaled) to see that at least empirically sparse SVD succeeds at a smaller scaling constant than thresholding even though their asymptotic rates are identical.

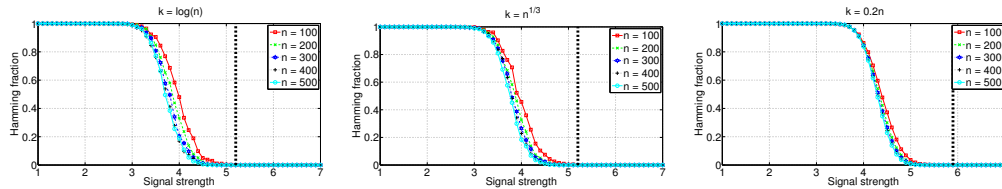


Figure 5.1: Thresholding: Hamming fraction versus rescaled signal strength.

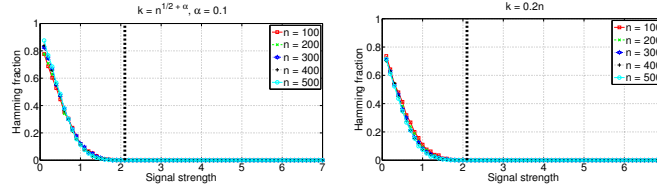


Figure 5.2: Averaging: Hamming fraction versus rescaled signal strength.

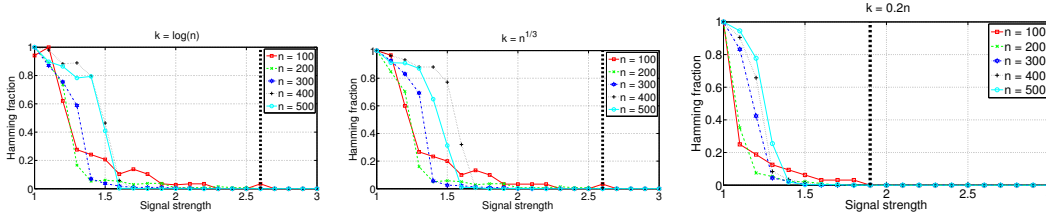


Figure 5.3: Sparse SVD: Hamming fraction versus rescaled signal strength.

5.6 Discussion

In this chapter, we analyze bi-clustering using a simple statistical model Eq. 5.1, where a sparse rank one matrix is perturbed with noise. Using this model, we have characterized the minimal signal strength below which no procedure can succeed in recovering the bi-cluster. This lower bound can be matched using an exhaustive search technique. However, it is still an open problem to find a computationally efficient procedure that is minimax optimal.

Amini and Wainwright [6] analyze the convex relaxation procedure proposed in d’Aspremont et al. [55] for high-dimensional sparse PCA. Under the minimax scaling for this problem they show that provided a rank-1 solution exists it has the desired sparsity pattern (they were however not able to show that a rank-1 solution exists with high probability). Somewhat surprisingly, we show that in the SVD case a rank-1 solution with the desired sparsity pattern *does not* exist with high probability. The two settings however are not identical since the noise in the spiked covariance model is Wishart rather than Gaussian, and has correlated entries. It would be interesting to analyze whether our negative result has similar implications for the sparse PCA setting.

The focus of this chapter has been on a model with one cluster, which although simple, provides several interesting theoretical insights. In practice, data often contains multiple clusters which need to be estimated. Many existing algorithms (see e.g. the papers [122] and [123]) try to estimate multiple clusters and it would be useful to analyze these theoretically.

5.7 Technical proofs

This section collects proofs of the main results stated in Sections 5.2, 5.3 and 5.4, as well as some additional results.

5.7.1 Proof of Theorem 5.2.1

We use a standard technique based on multiple hypothesis testing to obtain a lower bound on the minimal signal strength (see Section 2.6. in the book [192]). Without loss of generality, we assume $\sigma = 1$. Set $K_1 = [k_1]$ and $K_2 = [k_2]$, and let $\tau_0 = \beta(k_1 k_2)^{-1/2}$, so that the joint density of \mathbf{A} is

$$\prod_{ij} \mathcal{N}(a_{ij}; \tau_0 \mathbb{I}\{i \in K_1, j \in K_2\}, 1).$$

To lower bound the probability of error, we use the following relationship

$$\inf_{\Psi} \sup_{\theta \in \Theta} \mathbb{P}_{\theta}(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))) \geq \inf_{\Psi} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} \mathbb{P}_{\theta}(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta)))$$

where $\{\theta_0, \theta_1, \dots, \theta_M\}$ is a carefully chosen subset of Θ . Specifically, we select $\theta_0 = (\beta, K_1, K_2)$ and we choose the remaining points $\{\theta_1, \dots, \theta_M\}$, with $M = n_2 - k_2$, so that

$$\theta_{j-k_2} = (\beta, K_1, K_2^{(j)}), \quad j = k_2 + 1, \dots, n_2,$$

where $K_2^{(j)} := [k_2 - 1] \cup \{j\}$. For a $\theta \in \Theta$, below we denote with $(K_1(\theta), K_2(\theta))$ the associated bi-cluster.

Let $\phi(u)$ denote the density function of $\mathcal{N}(0, 1)$ with respect to the Lebesgue measure. With this, we can compute the Kullback-Leibler divergence between \mathbb{P}_{θ_0} and \mathbb{P}_{θ_j} :

$$\begin{aligned} D(\mathbb{P}_{\theta_0} | \mathbb{P}_{\theta_j}) &= \int \log \frac{d\mathbb{P}_{\theta_0}}{d\mathbb{P}_{\theta_j}} d\mathbb{P}_{\theta_0} \\ &= \sum_{i \in K_1} \int \log \frac{\phi(u_{ik_2} - \tau_0)}{\phi(u_{ik_2})} \phi(u_{ik_2} - \tau_0) du_{ik_2} \\ &\quad + \sum_{i \in K_1} \int \log \frac{\phi(u_{ij})}{\phi(u_{ij} - \tau_0)} \phi(u_{ij}) du_{ij} \\ &= \sum_{i \in K_1} \int (u_{ik_2} \tau_0 - \frac{\tau_0^2}{2}) \phi(u_{ik_2} - \tau_0) du_{ik_2} \\ &\quad + \sum_{i \in K_1} \int (\frac{\tau_0^2}{2} - u_{ij} \tau_0) \phi(u_{ij}) du_{ij} \\ &= \sum_{i \in K_1} \int u_{ik_2} \tau_0 \phi(u_{ik_2} - \tau_0) du_{ik_2} \\ &= k_1 \tau_0^2. \end{aligned} \tag{5.15}$$

Now it follows from Theorem 2.5 in the book [192] that, if

$$\tau_0 \leq \sqrt{\frac{\alpha \log(n_2 - k_2)}{k_1}},$$

then

$$\inf_{\Psi} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} \mathbb{P}_{\theta}(\Psi(\mathbf{A}) \neq (K_1(\theta), K_2(\theta))) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \frac{2\alpha}{\log M}\right) \xrightarrow{n_1, n_2 \rightarrow \infty} 1 - 2\alpha.$$

We chose the subset $\{\theta_1, \dots, \theta_M\}$ by fixing the set K_1 and alternating the last element of the set K_2 . Alternatively, we can fix K_2 and change the last element of the set K_1 or alternate both K_1 and K_2 . Repeating the argument above for these cases, we have that the probability of making an error is bounded away from zero if

$$\tau_0 \leq \max \left(\sqrt{\frac{\alpha \log(n_2 - k_2)}{k_1}}, \sqrt{\frac{\alpha \log(n_1 - k_1)}{k_2}}, \sqrt{\frac{\alpha \log(n_1 - k_1)(n_2 - k_1)}{k_1 + k_2 - 1}} \right), \quad (5.16)$$

which completes the proof.

5.7.2 Proof of Theorem 5.3.1

Without loss of generality, we assume that the noise variance $\sigma = 1$ and the true unknown sets $K_1 = [k_1]$ and $K_2 = [k_2]$. Define

$$F(\tilde{K}_1, \tilde{K}_2) := \sum_{i \in K_1} \sum_{j \in K_2} \mathbf{A}_{ij} - \sum_{i \in \tilde{K}_1} \sum_{j \in \tilde{K}_2} \mathbf{A}_{ij} \quad (5.17)$$

and note that an error is made if $F(\tilde{K}_1, \tilde{K}_2) < 0$, so that

$$\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] = \mathbb{P}[\cup_{\tilde{K}_1, \tilde{K}_2} \{F(\tilde{K}_1, \tilde{K}_2) < 0\}].$$

Observe that $F(\tilde{K}_1, \tilde{K}_2)$ depends only on the amount of overlap between $K_1 \times K_2$ and $\tilde{K}_1 \times \tilde{K}_2$. In particular, we have that

$$F(\tilde{K}_1, \tilde{K}_2) = F(d) \stackrel{d}{=} \mathcal{N}(d\beta(k_1 k_2)^{-1/2}, 2d\sigma^2) \quad (5.18)$$

where $d = k_1 k_2 - |K_1 \cap \tilde{K}_1| |K_2 \cap \tilde{K}_2|$. Therefore, using the union bound, we have that

$$\mathbb{P}[\Psi(\mathbf{A}) \neq (K_1, K_2)] \leq \sum_{i=0}^{k_1} C_{k_1}^i C_{n_1 - k_1}^{k_1 - i} \sum_{j=0}^{k_2} C_{k_2}^j C_{n_2 - k_2}^{k_2 - j} \mathbb{P}[F(k_1 k_2 - ij) < 0],$$

where, for readability, we have adopted the notation $C_n^i = \binom{n}{i}$.

Let $\tau_0 = \beta(k_1 k_2)^{-1/2}$. Using Eq. 5.18,

$$\begin{aligned} \mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2)) &\leq \sum_{i=0}^{k_1} C_{k_1}^i C_{n_1-k_1}^{k_1-i} \sum_{j=0}^{k_2} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \mathbb{P}(F(k_1 k_2 - ij) < 0) \\ &= \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} p_{ij} - p_{k_1 k_2} \end{aligned}$$

with

$$p_{ij} = C_{k_1}^i C_{n_1-k_1}^{k_1-i} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2})$$

and $\bar{\Phi}(\cdot)$ is the survival function of $\mathcal{N}(0, 1)$. Therefore, $\mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2))$ can be bounded by

$$\underbrace{(k_1 - 1)(k_2 - 1) \max_{\substack{i=0, \dots, k_1-1 \\ j=0, \dots, k_2-1}} p_{ij}}_{T_1} + \underbrace{(k_1 - 1) \max_{i=0, \dots, k_1-1} p_{ik_2}}_{T_2} + \underbrace{(k_2 - 1) \max_{j=0, \dots, k_2-1} p_{k_1 j}}_{T_3}.$$

We'll show how to handle T_1 , while T_2 and T_3 can be handled in an similar way.

$$\begin{aligned} T_1 &= (k_1 - 1)(k_2 - 1) \max_{\substack{i=0, \dots, k_1-1 \\ j=0, \dots, k_2-1}} C_{k_1}^i C_{n_1-k_1}^{k_1-i} C_{k_2}^j C_{n_2-k_2}^{k_2-j} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\ &\leq (k_1 - 1)(k_2 - 1) \max_{\substack{i=0, \dots, k_1-1 \\ j=0, \dots, k_2-1}} (n_1 - k_1)^{2(k_1-i)} (n_2 - k_2)^{2(k_2-j)} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\ &\leq \max_{\substack{i=0, \dots, k_1-1 \\ j=0, \dots, k_2-1}} (n_1 - k_1)^{3(k_1-i)} (n_2 - k_2)^{3(k_2-j)} \bar{\Phi}(\tau_0 \sqrt{(k_1 k_2 - ij)/2}) \\ &\leq \max_{\substack{i=0, \dots, k_1-1 \\ j=0, \dots, k_2-1}} (n_1 - k_1)^{3(k_1-i)} (n_2 - k_2)^{3(k_2-j)} \exp \left\{ -\frac{\tau_0^2}{4} \left(k_1 k_2 - \frac{ik_2}{2} - \frac{jk_1}{2} \right) \right\}. \end{aligned}$$

It is easy to see that the maximum is achieved at $i = k_1 - 1$ and $j = k_2 - 1$, which gives

$$T_1 \leq (n_1 - k_1)^3 (n_2 - k_2)^3 \exp \left(-\frac{\tau_0^2 (k_1 + k_2)}{8} \right).$$

Using the same reasoning

$$T_2 \leq (n_1 - k_1)^3 \exp \left(-\frac{\tau_0^2 k_2}{4} \right) \quad \text{and} \quad T_3 \leq (n_2 - k_2)^3 \exp \left(-\frac{\tau_0^2 k_1}{4} \right).$$

Probability of making an error can be bounded as $\mathbb{P}(\Psi(\mathbf{A}) \neq (K_1, K_2)) \leq T_1 + T_2 + T_3$, which concludes the proof.

5.7.3 Proof of Theorem 5.4.1

The proof follows from an applications of the union bound and the tail bound for the standard normal random variable given in Eq. 5.25. We have that

$$\min_{(i,j) \in K_1 \times K_2} |a_{ij}| \geq (k_1 k_2)^{-1/2} \beta - \max_{(i,j) \in K_1 \times K_2} |\Delta_{ij}| \geq (k_1 k_2)^{-1/2} \beta - \sigma \sqrt{2 \log \frac{k_1 k_2}{\delta}}$$

with probability $1 - 2\delta_1/(\sqrt{4\pi \log(1/\delta_1)})$ where $\delta_1 = \delta/(k_1 k_2)$. Similarly,

$$\max_{(i,j) \notin K_1 \times K_2} |a_{ij}| = \max_{(i,j) \notin K_1 \times K_2} |\Delta_{ij}| \leq \sigma \sqrt{2 \log \frac{(n_1 - k_1)(n_2 - k_2) + k_1(n_2 - k_2) + k_2(n_1 - k_1)}{\delta}}$$

with probability $1 - 2\delta_2/\sqrt{4\pi \log(1/\delta_2)}$ where $\delta_2 = \delta/|\{(i,j) \notin K_1 \times K_2\}|$. Combining the last two displays, the theorem follows.

5.7.4 Proof of Theorem 5.4.2

First consider identifying the rows. The sum of the elements of each row without activation is a draw from $\mathcal{N}(0, n_2 \sigma^2)$ and there are $(n_1 - k_1)$ of these, while the sum of the elements of each row with activation is a draw from $\mathcal{N}(4\sigma \max(\sqrt{n_2 \log(n_1)}, \sqrt{n_2 \log(n_2)} \left(\frac{n_2}{n_1}\right)^\alpha), n_2 \sigma^2)$, and there are k_1 of these.

Consider the probability that all the rows without activation have sum strictly less than $2\sigma \sqrt{n_2 \log(n_1)}$, and those with activation have sum strictly greater than the same quantity. If this condition is satisfied then selecting the k_1 rows with highest sum produces no errors. It is also easy to see that to upper bound the probability of error it suffices to show that the probability of error is small if the activation rows were drawn from $\mathcal{N}(4\sigma \sqrt{n_2 \log(n_1)}, n_2 \sigma^2)$.

The result follows from applying a standard Gaussian tail bound, followed by a union bound, i.e.

$$\mathbb{P}(X - \mu > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

therefore, noting the symmetry we can bound

$$\mathbb{P}(\text{error}) \leq n_1 \exp\left(-\frac{4\sigma^2 n_2 \log(n_1)}{2n_2 \sigma^2}\right) = n_1 (n_1)^{-2} = \delta_1.$$

A similar argument shows that we can bound δ_2 , the probability of making an error in identifying the columns. The result follows.

5.7.5 Proof of Theorem 5.4.3

We prove the theorem using a constructive procedure. Our arguments are adapted from the arguments used in the proof of Theorem 2 in Amiri and Wainwright [6]. We construct a rank one solution $\widehat{\mathbf{X}}$ that is a global solution of the problem in Eq. 5.11. Using Theorem 5.7.13, which states the first order conditions for a global optimum, we have that

$$-\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}}' & \mathbf{0} \end{pmatrix} + (\widehat{\pi}_1 - \widehat{\pi}_2) \mathbf{I}_{n_1+n_2} = \widehat{\mathbf{K}}, \quad (5.19)$$

where $\widehat{\mathbf{S}} \in \partial \|\widehat{\mathbf{X}}\|_1$ is an element of the subgradient of the element-wise ℓ_1 norm evaluated at $\widehat{\mathbf{X}}$, $\widehat{\pi}_1$ and $\widehat{\pi}_2$ are Lagrange multipliers associated with the constraint $\text{tr } \widehat{\mathbf{X}} = 2$, and $\widehat{\mathbf{K}}$ is an element of the normal cone to \mathcal{S}_+^n evaluated at $\widehat{\mathbf{X}}$. For $\widehat{\mathbf{S}}$, we have that $\max_{ij} |\widehat{S}_{ij}| \leq 1$ and $\text{tr } \widehat{\mathbf{S}}' \mathbf{X}^{12} = \mathbf{1}' |\mathbf{X}| \mathbf{1}$. From Eq. 5.30, we have that $\widehat{\mathbf{K}} = -\widehat{\mathbf{Z}}^\perp \mathbf{B} \widehat{\mathbf{Z}}^\perp$ where columns of $\widehat{\mathbf{Z}}^\perp$ form orthonormal basis for the null space of $\widehat{\mathbf{X}}$ and $\mathbf{B} \in \mathcal{S}_+^n$. See §5.7.10 for more details.

Suppose that the matrix $\widehat{\mathbf{X}}$ is rank one and that the sparsity pattern of $\widehat{\mathbf{X}}^{12}$ correctly recovers K_1 and K_2 . Then we have that $\widehat{\mathbf{S}}_{K_1 K_2} = \mathbf{s}_{\widehat{\mathbf{u}}} \mathbf{s}'_{\widehat{\mathbf{v}}}$ where $\mathbf{s}_{\widehat{\mathbf{u}}} = \text{sign}(\widehat{\mathbf{u}}_{K_1})$ and $\mathbf{s}_{\widehat{\mathbf{v}}} = \text{sign}(\widehat{\mathbf{v}}_{K_2})$. Furthermore, $\widehat{\mathbf{X}}_{K_1 K_2}^{12} = \widehat{\mathbf{u}}_{K_1} \widehat{\mathbf{v}}'_{K_2}$ where $\widehat{\mathbf{u}}_{K_1}$ is a left singular vector and $\widehat{\mathbf{v}}_{K_2}$ is a right singular vector of $\mathbf{A}_{K_1 K_2} - \lambda \widehat{\mathbf{S}}_{K_1 K_2}$ associated with the largest singular vector. In fact, the following Lemma will show that $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are left and right singular vectors of $\mathbf{A}_{K_1 K_2} - \lambda \mathbf{s}_{\widehat{\mathbf{u}}} \mathbf{s}'_{\widehat{\mathbf{v}}}$ where $\mathbf{s}_{\mathbf{u}} = \text{sign}(\mathbf{u}_{K_1})$ and $\mathbf{s}_{\mathbf{v}} = \text{sign}(\mathbf{v}_{K_2})$. That is, $\mathbf{s}_{\widehat{\mathbf{u}}}$ and $\mathbf{s}_{\widehat{\mathbf{v}}}$ recover signs of $\mathbf{s}_{\mathbf{u}}$ and $\mathbf{s}_{\mathbf{v}}$. Note that singular vectors are uniquely defined only up to a rotation, therefore, we use a convention that the first non-zero coordinate of a left singular vector is positive.

Let $\mathbf{M} = \mathbf{A}_{K_1 K_2} - \lambda \text{sign}(\mathbf{u}_{K_1}) \text{sign}(\mathbf{v}_{K_2})'$ and let $\alpha = \beta/2$. Since $\lambda = \frac{\beta}{2\sqrt{k_1 k_2}}$, we have that $\mathbf{M} = \alpha \mathbf{u}_{K_1} \mathbf{v}'_{K_2} + \Delta_{K_1 K_2}$. Let $\widehat{\alpha} = \sigma_1(\mathbf{M})$ be the largest singular value of \mathbf{M} .

Lemma 5.7.1. *Under the conditions of Theorem 5.4.3, we have that*

$$\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = \mathcal{O} \left(\sqrt{\frac{\log k_1}{k_1 k_2 \log(n_1 - k_2)(n_2 - k_2)}} \right)$$

and

$$\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty = \mathcal{O} \left(\sqrt{\frac{\log k_2}{k_1 k_2 \log(n_1 - k_2)(n_2 - k_2)}} \right)$$

with probability $1 - \mathcal{O}(k_1^{-1})$.

Under the assumptions of Theorem 5.4.3 $\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = o(1/\sqrt{k_1})$ and $\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty = o(1/\sqrt{k_2})$ as $n_1, n_2 \rightarrow \infty$, which shows that $\mathbf{s}_{\widehat{\mathbf{u}}}$ and $\mathbf{s}_{\widehat{\mathbf{v}}}$ recover signs of $\mathbf{s}_{\mathbf{u}}$ and $\mathbf{s}_{\mathbf{v}}$.

Next, we set elements of $\widehat{\mathbf{S}}_{K_1^c K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2^c}$ such that $(\widehat{\mathbf{u}}'_{K_1}, \mathbf{0}')$ and $(\widehat{\mathbf{v}}'_{K_2}, \mathbf{0}')$ are singular vectors of $\mathbf{A} - \lambda \widehat{\mathbf{S}}$. Note that for these two singular vectors the choice of $\widehat{\mathbf{S}}_{K_1^c K_2^c}$ is irrelevant. Let $\widehat{\mathbf{S}}_{K_1^c K_2} = \lambda^{-1} \Delta_{K_1^c K_2}$ and $\widehat{\mathbf{S}}_{K_1 K_2^c} = \lambda^{-1} \Delta_{K_1 K_2^c}$. Using a normal tail bound Eq. 5.25 and the union bound

$$\|\widehat{\mathbf{S}}_{K_1^c K_2}\|_\infty \leq \frac{4\sigma \sqrt{k_1 k_2 \log[(n_1 - k_1)k_2]}}{\beta} \quad \text{and} \quad \|\widehat{\mathbf{S}}_{K_1 K_2^c}\|_\infty \leq \frac{4\sigma \sqrt{k_1 k_2 \log[(n_2 - k_2)k_1]}}{\beta}$$

with probability $1 - \mathcal{O}[(n_1 - k_1)^{-1} k_2^{-1}]$. Under the assumptions of the theorem we have that $\|\widehat{\mathbf{S}}_{K_1 K_2^c}\|_\infty < 1$ and $\|\widehat{\mathbf{S}}_{K_1^c K_2}\|_\infty < 1$.

Let $\widehat{\mathbf{x}} = (\widehat{\mathbf{u}}'_{K_1}, \mathbf{0}', \widehat{\mathbf{v}}'_{K_2}, \mathbf{0}')$, so that $\widehat{\mathbf{X}} = \widehat{\mathbf{x}} \widehat{\mathbf{x}}'$. We have established so far that $\widehat{\mathbf{x}}$ is an eigenvector of

$$- \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{S}} \\ \widehat{\mathbf{S}}' & \mathbf{0} \end{pmatrix}.$$

Therefore, multiplying Eq. 5.19 by $\widehat{\mathbf{x}}$ from right and taking a dot product with $\widehat{\mathbf{x}}$ we have that $\widehat{\alpha} = \widehat{\pi}_1 - \widehat{\pi}_2$. Finally, we need to set $\widehat{\mathbf{S}}_{K_1^c K_2^c}$ such that Eq. 5.19 holds. Set $\widehat{\mathbf{K}}$ to the left hand side of Eq. 5.19, then we need to show that $\widehat{\mathbf{K}} \succeq \mathbf{0}$. By construction of $\widehat{\mathbf{X}}$, we have that $\widehat{\mathbf{K}}_{(K_1 K_2)(K_1 K_2)} \succeq \mathbf{0}$. Therefore, we only need to show that $\widehat{\mathbf{K}}_{(K_1^c K_2^c)(K_1^c K_2^c)} \succeq \widehat{\mathbf{K}}_{(K_1^c K_2^c)(K_1 K_2)} (\widehat{\mathbf{K}}_{(K_1 K_2)(K_1 K_2)})^\dagger \widehat{\mathbf{K}}_{(K_1 K_2)(K_1^c K_2^c)}$. With the current choice of $\widehat{\mathbf{S}}_{K_1^c K_2^c}$ and $\widehat{\mathbf{S}}_{K_1 K_2^c}$, we can choose $\widehat{\mathbf{S}}_{K_1^c K_2^c} = \lambda^{-1} \Delta_{K_1^c K_2^c}$ to satisfy Eq. 5.19. From Eq. 5.25 and the union bound

$$\|\widehat{\mathbf{S}}_{K_1^c K_2^c}\|_\infty \leq \frac{4\sigma \sqrt{k_1 k_2 \log[(n_1 - k_1)(n_2 - k_2)]}}{\beta}$$

with probability $1 - \mathcal{O}((n_1 - k_1)^{-1}(n_2 - k_2)^{-1})$. Under the assumptions of the theorem we have that $\|\widehat{\mathbf{S}}_{K_1^c K_2^c}\|_\infty < 1$. This concludes the proof of the theorem.

5.7.6 Proof of Theorem 5.4.4

Without loss of generality assume $\sigma = 1$. From the proof of Theorem 5.4.3, it is sufficient to show that $\widehat{\mathbf{S}}_{K_1^c K_2^c}$ cannot be chosen so that $\widehat{\mathbf{K}}_{(K_1^c K_2^c)(K_1^c K_2^c)} \succeq \mathbf{0}$. This is equivalent to showing that

$$\min_{\|\mathbf{S}_{K_1^c K_2^c}\|_\infty \leq 1} \max_{\|\mathbf{x}\|_2=1} \mathbf{x}' \left[\begin{pmatrix} \mathbf{0} & \mathbf{A}_{K_1^c K_2^c} \\ \mathbf{A}'_{K_1^c K_2^c} & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \mathbf{S}_{K_1^c K_2^c} \\ \mathbf{S}'_{K_1^c K_2^c} & \mathbf{0} \end{pmatrix} \right] \mathbf{x} > \widehat{\alpha} \quad (5.20)$$

with probability tending to 1. The left hand side of Eq. 5.20 is lower bounded by

$$\frac{2\|\Delta_{K_1^c K_2^c} + \lambda \mathbf{S}_{K_1^c K_2^c}\|_F}{\min(\sqrt{n_1 - k_1}, \sqrt{n_2 - k_2})}.$$

Entries of $\mathbf{A}_{K_1^c K_2^c}$ are soft-thresholded towards zero by $\mathbf{S}_{K_1^c K_2^c}$ to minimize the Frobenious norm. Using Eq. 5.25,

$$\mathbb{P}[|\mathcal{N}(0, 1)| > 2\lambda] \geq \frac{4\lambda}{\sqrt{2\pi(4\lambda^2+1)}} \exp(-2\lambda^2) =: c_\lambda.$$

Using the assumption that $\lambda = \sqrt{c \log \max(n_1 - k_1, n_2 - k_2)}$, we get that $c_\lambda = (\max(n_1 - k_1, n_2 - k_2))^{-2c} L_n$, where $L_n = \mathcal{O}(\text{polylog}(\max(n_1 - k_1, n_2 - k_2)))$.

Let $Z \sim \text{Bin}(N, c_\lambda)$ with $N = (n_2 - k_2)(n_1 - k_1)$. From Lemma 5.7.11, $Z > Nc_\lambda/2$ with probability $1 - 2\exp(-Nc_\lambda/8)$. Conditioning on the event $\{Z > Nc_\lambda/2\}$, the left hand side of 5.20 is lower bounded by

$$\frac{2\lambda\sqrt{2Nc_\lambda}}{\min(\sqrt{n_2 - k_2}, \sqrt{n_1 - k_1})} = 2\lambda\sqrt{2c_\lambda} \max(\sqrt{n_1 - k_1}, \sqrt{n_2 - k_2}).$$

Plugging in the expression for c_λ found above, we see that the left hand side of 5.20 is lower bounded by $(\max(n_1 - k_1, n_2 - k_2))^{1/2-c} L_n$.

Lemma 5.7.9 provides an upper bound for the right hand side of 5.20 of the form $\lambda\sqrt{k_1k_2} + 2(\sqrt{k_1} + \sqrt{k_2})$ with probability $1 - 2\exp(-(\sqrt{k_1} + \sqrt{k_2})^2/2)$. We can conclude that 5.20 holds with probability tending to one, since

$$(\max(n_1 - k_1, n_2 - k_2))^{1/2-c} L_n \geq \sqrt{ck_1k_2 \log \max(n_1 - k_1, n_2 - k_2)} + 2(\sqrt{k_1} + \sqrt{k_2})$$

for sufficiently large n_1 and n_2 as $k_1 = o(n^{1/2-c})$ and $k_2 = o(n^{1/2-c})$ under assumptions.

The theorem follows since $\beta = 2\lambda\sqrt{k_1k_2}$. The constant c can be chosen so that $c < 1/2$.

5.7.7 Proof of Lemma 5.7.1

It follows directly from Weyl's theorem (see for example the book [179]) that

$$|\alpha - \hat{\alpha}| \leq \sigma_1(\Delta_{K_1K_2}). \quad (5.21)$$

Denote $\hat{\mathbf{u}}_{K_1}$ and $\hat{\mathbf{v}}_{K_2}$ the singular vectors of \mathbf{M} associated with $\hat{\alpha}$, that is,

$$\begin{aligned} \mathbf{M}\hat{\mathbf{v}}_{K_2} &= \hat{\alpha}\hat{\mathbf{u}}_{K_1}, \quad \text{and} \\ \mathbf{M}'\hat{\mathbf{u}}_{K_1} &= \hat{\alpha}\hat{\mathbf{v}}_{K_2}. \end{aligned} \quad (5.22)$$

Let $\mathbf{u}_{K_1}^\perp \in \{\mathbf{a} \in \mathbb{R}^{k_1} : \mathbf{a} \perp \mathbf{u}_{K_1}, \|\mathbf{a}\| = 1\}$ and $\mathbf{v}_{K_2}^\perp \in \{\mathbf{a} \in \mathbb{R}^{k_2} : \mathbf{a} \perp \mathbf{v}_{K_2}, \|\mathbf{a}\| = 1\}$. With this we write $\hat{\mathbf{v}}_{K_2} = c_1^v \mathbf{v}_{K_2} + c_0^v \mathbf{v}_{K_2}^\perp$ and $\hat{\mathbf{u}}_{K_1} = c_1^u \mathbf{u}_{K_1} + c_0^u \mathbf{u}_{K_1}^\perp$ where $(c_1^v)^2 + (c_0^v)^2 = 1$ and $(c_1^u)^2 + (c_0^u)^2 = 1$. Lemma 5.7.2 gives a lower bound on c_1^u and c_1^v and is proven below.

From Eq. 5.22 we have

$$\alpha c_1^v \mathbf{u}_{K_1} + \Delta_{K_1K_2} \hat{\mathbf{v}}_{K_2} = \hat{\alpha} \hat{\mathbf{u}}_{K_1}$$

which further decomposes into

$$\alpha c_1^v \mathbf{u}_{K_1} + \Delta_{K_1K_2} (c_1^v \mathbf{v}_{K_2} + c_0^v \mathbf{v}_{K_2}^\perp) = \hat{\alpha} (\hat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}) + \hat{\alpha} \mathbf{u}_{K_1}.$$

Using Taylor series expansion $\hat{\alpha}^{-1} \lesssim \alpha^{-1} + \sigma_1(\Delta_{K_1K_2})\alpha^{-2}$. Now

$$\begin{aligned} & \|\hat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty \\ & \leq |\hat{\alpha}^{-1} \alpha c_1^v - 1| \|\mathbf{u}_{K_1}\|_\infty + \hat{\alpha}^{-1} |c_1^v| \|\Delta_{K_1K_2} \mathbf{v}_{K_2}\|_\infty + \hat{\alpha}^{-1} |c_0^v| \|\Delta_{K_1K_2} \mathbf{v}_{K_2}^\perp\|_\infty + o(1) \\ & \leq 2\alpha^{-1} \sigma_1(\Delta_{K_1K_2}) \|\mathbf{u}_{K_1}\|_\infty + \alpha^{-1} \|\Delta_{K_1K_2} \mathbf{v}_{K_2}\|_\infty + 2\alpha^{-2} \sigma_1(\Delta_{K_1K_2}) \|\Delta_{K_1K_2}\|_{\infty,2} + o(1) \end{aligned}$$

using Eq. 5.21 and Lemma 5.7.2. The three terms in the display above can be bounded using Lemma 5.7.9, Lemma 5.7.6 and Lemma 5.7.7. Then

$$\|\hat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = \alpha^{-1} \mathcal{O} \left(\sqrt{k_1} \|\mathbf{u}_{K_1}\|_\infty + \sqrt{\log k_1} + \alpha^{-1} k_2 \right) = \alpha^{-1} \mathcal{O}(\sqrt{\log k_1})$$

with probability $1 - \mathcal{O}(k_1^{-1})$. A similar calculation gives a bound on $\|\hat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty$. This completes the proof of Lemma 5.7.1.

The following Lemma establishes a lower bound on $\hat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1}$ and $\hat{\mathbf{v}}'_{K_2} \mathbf{v}_{K_2}$ under our sign convention.

Lemma 5.7.2. *We have that $c_1^u \geq 1 - 2\alpha^{-1}\sigma_1(\Delta_{K_1K_2})$ and $c_1^v \geq 1 - 2\alpha^{-1}\sigma_1(\Delta_{K_1K_2})$.*

Proof of Lemma 5.7.2. From Eq. 5.22 we have

$$\alpha \widehat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1} \mathbf{v}'_{K_2} \widehat{\mathbf{v}}_{K_2} + \widehat{\mathbf{u}}'_{K_1} \Delta_{K_1K_2} \widehat{\mathbf{v}}_{K_2} = \widehat{\alpha}.$$

Using the triangle inequality

$$\begin{aligned} |\alpha - \alpha \widehat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1} \mathbf{v}'_{K_2} \widehat{\mathbf{v}}_{K_2}| &\leq |\alpha - \widehat{\alpha}| + |\widehat{\alpha} - \alpha \widehat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1} \mathbf{v}'_{K_2} \widehat{\mathbf{v}}_{K_2}| \\ &\leq 2\sigma_1(\Delta_{K_1K_2}), \end{aligned}$$

since $|\widehat{\mathbf{u}}'_{K_1} \Delta_{K_1K_2} \widehat{\mathbf{v}}_{K_2}| \leq \sigma_1(\Delta_{K_1K_2})$. Under our sign convention, this implies that

$$1 - \widehat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1} \mathbf{v}'_{K_2} \widehat{\mathbf{v}}_{K_2} \leq 2\alpha^{-1}\sigma_1(\Delta_{K_1K_2}).$$

We conclude that

$$\begin{aligned} \widehat{\mathbf{u}}'_{K_1} \mathbf{u}_{K_1} &\geq 1 - 2\alpha^{-1}\sigma_1(\Delta_{K_1K_2}), \quad \text{and} \\ \widehat{\mathbf{v}}'_{K_1} \mathbf{v}_{K_1} &\geq 1 - 2\alpha^{-1}\sigma_1(\Delta_{K_1K_2}). \end{aligned}$$

□

5.7.8 Identifying Large Biclusters Without Normality Assumption

We now consider a computationally feasible nonparametric procedure for biclustering that makes minimal assumptions on the distribution of the noise and on the form of the signal. When the clusters are large in a sense specified by the theorem below, the procedure recovers the true bicluster with large probability.

Let F be any distribution with median zero and positive, continuous density. As before, we let Δ be a $n_1 \times n_2$ error matrix filled with iid draws from F . We now assume that

$$\mathbf{A} = \mathbf{B} + \Delta$$

where $\mathbf{B} = \{B_{ij}\}_{i \in [n_1], j \in [n_2]}$ is such that $B_{ij} = 0$ for $(i, j) \in K_1 \times K_2$ and

$$\beta \equiv \min_{i \in K_1, j \in K_2} B_{ij} > 0.$$

Let C_j denote the number of positive entries in the j^{th} column of A and let R_i denote the number of positive entries in the i^{th} row of A . Define $\Psi(A)$ to consist of all rows such that $R_i > r \equiv (n_2/2) + \sqrt{n_2 \log n_2}$ and all columns such that $C_j > c \equiv (n_1/2) + \sqrt{n_1 \log n_1}$.

Let $Z \sim F$ and define $\pi = \mathbb{P}(Z + \beta > 0) = 1 - F(\beta)$. Finally, we measure the signal strength by the quantities

$$\psi_1 = k_1 \left[\frac{1}{2} - F(-\beta) \right], \quad \psi_2 = k_2 \left[\frac{1}{2} - F(-\beta) \right].$$

Theorem 5.7.3. *Suppose that the following conditions hold:*

$$\psi_1 > \sqrt{4 \log(k_2 n_1)} \quad (5.23)$$

$$\begin{aligned} \psi_2 &> \sqrt{4 \log(k_1 n_2)} \\ \psi_1 &\geq \sqrt{n_1 \log n_1} \\ \psi_2 &\geq \sqrt{n_2 \log n_2}. \end{aligned} \quad (5.24)$$

Then

$$\mathbb{P}(\Psi(A) \neq (K_1, K_2)) \leq 4 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Proof. Consider a null column that does not intersect the cluster. Then $C_j \sim \text{Binomial}(n_1, 1/2)$. By Hoeffding's inequality, $\mathbb{P}(C_j > c) \leq 1/n_1^2$. Similarly for a null row, $\mathbb{P}(R_j > r) \leq 1/n_2^2$. By the union bound, the probability of including any null row or column is at most $n_1/n_1^2 + n_2/n_2^2 = (1/n_1) + (1/n_2)$.

Now consider a non-null column. For simplicity assume that all nonzero β_{ij} are equal to the minimum value β . The extension to the general case is straightforward. Then $C_j = U + V$ where $U \sim \text{Binomial}(n_1 - k_1, 1/2)$ and $V \sim \text{Binomial}(k_1, \pi)$ where $\pi = \mathbb{P}(Z + \beta > 0) = 1 - F(-\beta)$. Here, $Z \sim F$. The probability of excluding column j is $\mathbb{P}(U + V < c)$. Now $U + V$ is the sum of independent but not identically distributed Bernoulli random variables. Applying Hoeffding's inequality for non identically distributed variables we have $\mathbb{P}(U + V < c) \leq e^{-2(\mu - c)^2/n_1}$ where

$$\mu = \mathbb{E}(U + V) = \frac{n_1 - k_1}{2} + k_1 \pi.$$

Substituting for μ and c and using the fact that $\pi - 1/2 = 1/2 - F(-\beta)$,

$$\begin{aligned} \mathbb{P}(U + V < c) &\leq e^{-2(\mu - c)^2/n_1} \\ &= \exp \left(\frac{k_1}{\sqrt{n_1}} (\pi - 1/2) - \frac{1}{2} \sqrt{\log n_1} \right)^2 \\ &\leq \exp \left(-\frac{k_1^2 (\pi - 1/2)^2}{4n_1} \right) \end{aligned}$$

where we used (5.24). By (5.23), the last quantity is less than $1/(k_2 n_1)$. Taking the union bound over all the k_2 columns in the cluster, the probability of missing a relevant column is at most $1/n_1$. A similar bound applies to the rows. \square

5.7.9 Concentration inequalities

We now collect useful results on tail bounds of various random quantities used throughout the chapter. We start by stating a lower and upper bound on the survival function of the standard

normal random variable. Let $Z \sim \mathcal{N}(0, 1)$ be a standard normal random variable. Then for $t > 0$

$$\frac{1}{\sqrt{2\pi}} \frac{t}{t^2 + 1} \exp(-t^2/2) \leq \mathbb{P}(Z > t) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2). \quad (5.25)$$

We will use the above inequality to bound some quantities involving norms of random matrices with independent standard normal entries. We provide a few more definitions.

Definition 6. Let ϵ be a positive number. A set X is an ϵ -net of a set Y if for any $y \in Y$, there exists $x \in X$ such that $\|y - x\| \leq \epsilon$.

The following result is the standard ϵ -net argument.

Lemma 5.7.4. Let $\mathcal{N} \subset \mathcal{S}^{n_2-1}$ be an ϵ -net \mathcal{N} of \mathcal{S}^{n_2-1} and let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a linear map. Then there is a vector $y \in \mathcal{N}$ such that

$$\|\mathbf{A}y\| \geq (1 - \epsilon) \max_{x \in \mathcal{S}^{n_2-1}} \|\mathbf{A}x\|.$$

The minimum size of the ϵ -net is well-known.

Lemma 5.7.5. There is an ϵ -net of a unit sphere in d dimensions of size at most $(\frac{3}{\epsilon})^d$.

Lemma 5.7.6. Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Then for any fixed $x \in \mathcal{S}^{n_2-1}$,

$$\mathbb{P}[\|\mathbf{A}x\|_\infty \geq t] \leq \frac{2n_1}{\sqrt{2\pi}t} \exp(-t^2/2).$$

Proof of Lemma 5.7.6. Observe that $\mathbf{A}x \sim N(0, \mathbf{I}_{n_1})$. The result follows from an application of a standard Gaussian tail bound and the union bound. \square

The following two results bound operator norms $\|\mathbf{A}\|_{\infty, 2}$ and $\|\mathbf{A}\|_{\infty, \infty}$.

Lemma 5.7.7. Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Fix $\delta > 0$. Then

$$\|\mathbf{A}\|_{\infty, 2} \leq \sqrt{8 \left(\log n_1 + n_2 \log 6 + \log \frac{2}{\sqrt{2\pi}\delta} \right)} =: K_{\delta, n_1, n_2} \quad (5.26)$$

with probability $1 - \delta/K_{\delta, n_1, n_2}$.

Proof of Lemma 5.7.7. By definition, we have that

$$\|\mathbf{A}\|_{\infty, 2} = \max_{\|x\|_2 \leq 1} \|\mathbf{A}x\|_\infty.$$

Let $\mathcal{N} \subset \mathcal{S}^{n_2}$ be an ϵ -net of \mathcal{S}^{n_2-1} . Using Lemma 5.7.4 we have that

$$\mathbb{P}[\|\mathbf{A}\|_{\infty, 2} \geq t] \leq \mathbb{P}[(1 - \epsilon)^{-1} \max_{y \in \mathcal{N}} \|\mathbf{A}y\|_\infty \geq t].$$

Setting $\epsilon = \frac{1}{2}$, applying Lemma 5.7.5, Lemma 5.7.6 and using the union bound, we have that

$$\mathbb{P}[\|\mathbf{A}\|_{\infty, 2} \geq t] \leq \frac{2n_1}{\sqrt{2\pi}t} 6^{n_2} \exp(-t^2/8).$$

We can conclude the proof by setting $t = K_{\delta, n_1, n_2}$. \square

Lemma 5.7.8. Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. Fix $\delta > 0$. Then there exists a sufficiently large constant C such that

$$\|\mathbf{A}\|_{\infty, \infty} \leq \sqrt{8 \left(n_2 \log n_1 + n_2^2 \log 6 + n_2 \log \frac{2}{\sqrt{2\pi}\delta} \right)} =: \sqrt{n_2} K_{\delta, n_1, n_2} \quad (5.27)$$

with probability $1 - \delta/K_{\delta, n_1, n_2}$ where K_{δ, n_1, n_2} is defined in Eq. 5.26.

Proof of Lemma 5.7.8. For any $\mathbf{x} \in \mathbb{R}^{n_2}$, $\|\mathbf{x}\|_2 \leq \sqrt{k} \|\mathbf{x}\|_\infty$. Now

$$\|\mathbf{A}\|_{\infty, \infty} = \max_{\|\mathbf{x}\|_\infty \leq 1} \|\mathbf{A}\mathbf{x}\|_\infty \leq \max_{\|\mathbf{x}\|_2 \leq \sqrt{n_2}} \|\mathbf{A}\mathbf{x}\|_\infty = \sqrt{n_2} \|\mathbf{A}\|_{\infty, 2}.$$

The result follows from Lemma 5.7.7. □

Lemma 5.7.9 ([57]). Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a random matrix whose elements are independent standard normal random variables. We have that

$$\mathbb{P}[\sigma_1(\mathbf{A}) \geq \sqrt{n_1} + \sqrt{n_2} + t] \leq 2 \exp(-t^2/2). \quad (5.28)$$

Lemma 5.7.10. If $z_k \sim \text{Bin}(k, \pi_k)$, then for all $k \geq 1$ and all $\pi_k \in (0, 1)$ it holds that

$$\mathbb{P}[z_k = 0] \leq \exp(-k\pi_k).$$

Proof. $\mathbb{P}[z_k = 0] = (1 - \pi_k)^k = \exp(-k \log(\frac{1}{1-\pi_k})) = \exp(-k(\pi_k + \mathcal{O}(\pi_k^2))) \leq \exp(-k\pi_k)$. □

Lemma 5.7.11. If $z_k \sim \text{Bin}(k, \pi_k)$, then

$$\mathbb{P}[z_k \leq k\pi_k - t] \leq \exp(-t^2/(2k\pi_k))$$

and

$$\mathbb{P}[z_k \geq k\pi_k + t] \leq \exp(-t^2/(2(k\pi_k + t/3))).$$

5.7.10 Convex analysis

The following results are standard. We use them to derive the KKT condition for the optimization problem in Eq. 5.11.

Definition 7. Let \mathcal{C} be a convex set. The function $\delta(x|\mathcal{C})$ defined as

$$\delta(x|\mathcal{C}) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{if } x \notin \mathcal{C} \end{cases}$$

is called the indicator function of the convex set \mathcal{C} .

Definition 8. Let $\partial\delta(x|\mathcal{C})$ denote the normal cone to \mathcal{C} at x defined as

$$\partial\delta(a|\mathcal{C}) = \{y : \langle x - a, y \rangle \leq 0, \forall x \in \mathcal{C}\}.$$

The normal cone be equivalently defined as

$$\partial\delta(a|\mathcal{C}) = \{y : \sup_{x \in \mathcal{C}} \langle x, y \rangle = \langle a, y \rangle\}.$$

If a is interior to \mathcal{C} then $\partial\delta(a|\mathcal{C}) = \{0\}$, and if a is exterior to \mathcal{C} then $\partial\delta(a|\mathcal{C}) = \emptyset$

Let \mathcal{S}_+^n be the cone of positive semi-definite symmetric matrices in $\mathbb{R}^{n \times n}$.

Theorem 5.7.12 ([75]). *The normal cone to \mathcal{S}_+^n is defined as*

$$\partial\delta(\mathbf{A}|\mathcal{S}_+^n) = \begin{cases} \emptyset & \text{if } \mathbf{A} \notin \mathcal{S}_+^n \\ \{\mathbf{B} : -\mathbf{B} \in \mathcal{S}_+^n, \text{tr } \mathbf{A}\mathbf{B} = 0\} & \text{if } \mathbf{A} \in \mathcal{S}_+^n. \end{cases} \quad (5.29)$$

Alternatively for $\mathbf{A} \in \mathcal{S}_+^n$, equation Eq. 5.29 becomes

$$\partial\delta(\mathbf{A}|\mathcal{S}_+^n) = \{\mathbf{B} = -\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}' : \mathbf{\Lambda} \in \mathcal{S}_+^n\} \quad (5.30)$$

where columns of \mathbf{Z} form orthonormal basis for the null space of \mathbf{A} .

Theorem 5.7.13 ([162], Chapter 5). *If $\widehat{\mathbf{A}}$ solves the problem*

$$\begin{aligned} \min & f(\mathbf{A}) \\ \text{subject to} & \mathbf{A} \in \mathcal{S}_+^n, \quad g(\mathbf{A}) \leq 0, \end{aligned}$$

then $\widehat{\mathbf{A}}$ is feasible and there exist matrices $\widehat{\mathbf{G}} \in \partial f(\widehat{\mathbf{A}})$, $\widehat{\mathbf{B}} \in \partial\delta(\widehat{\mathbf{A}}|\mathcal{S}_+^n)$, $\widehat{\mathbf{C}} \in \partial g(\widehat{\mathbf{A}})$ and a multiplier $\widehat{\pi} \geq 0$, $\widehat{\pi}g(\widehat{\mathbf{A}}) = 0$ such that

$$\widehat{\mathbf{G}} + \widehat{\mathbf{B}} + \widehat{\pi}\widehat{\mathbf{C}} = 0.$$

5.7.11 Nuclear norm and ℓ_1 norm penalty

Under the model Eq. 5.1, the problem of biclustering can be thought of recovering a matrix that is both low rank and sparse. As pointed out by a reviewer, from this point of view a natural combination of the nuclear norm and the ℓ_1 norm leads to the following optimization problem

$$\min_{\mathbf{X} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}} \frac{1}{2} \|\mathbf{A} - \mathbf{X}\|_F^2 + \lambda_1 \|\mathbf{X}\|_* + \lambda_2 \mathbf{1}'\mathbf{X}\mathbf{1}. \quad (5.31)$$

The norm $\|\mathbf{X}\|_*$ is the nuclear norm defined as the sum of the singular values of \mathbf{X} , that is, if $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ is the singular value decomposition of \mathbf{X} , then $\|\mathbf{X}\|_* = \sum_i D_{ii}$. The tuning parameter λ_1 control the rank of the solution and λ_2 controls the sparsity of the solution. Compared to the optimization procedure in Eq. 5.11, there is an additional tuning parameter that needs to be selected in practice. Combination of the nuclear norm and the ℓ_1 norm was shown useful in robust PCA [43]. For the problem of biclustering, the formulation in Eq. 5.31 does not lead to improvement over Eq. 5.11 as we show below.

We analyze the problem Eq. 5.31 in a similar way to the proof of Theorem 5.4.3. That is, we construct a rank one solution $\widehat{\mathbf{X}}$ that is a global solution of the objective Eq. 5.31. The following Lemma gives a subgradient of the nuclear norm used in stating the first order conditions for a global optimum.

Lemma 5.7.14. *If $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ is the singular value decomposition of \mathbf{X} then the subdifferential of $\|\cdot\|_*$ is equal to*

$$\partial\|\mathbf{X}\|_* = \{\mathbf{U}\mathbf{V}' + \mathbf{Z} : \sigma_1(\mathbf{Z}) \leq 1, \mathbf{U}'\mathbf{Z} = 0 \text{ and } \mathbf{Z}\mathbf{V} = 0\}. \quad (5.32)$$

Now, the first order condition for a global optimum of Eq. 5.31 is

$$\widehat{\mathbf{X}} - \mathbf{A} + \lambda_1 \widehat{\mathbf{K}} + \lambda_2 \widehat{\mathbf{S}} = \mathbf{0} \quad (5.33)$$

where $\widehat{\mathbf{S}} \in \partial\|\widehat{\mathbf{X}}\|_1$ and $\widehat{\mathbf{K}} \in \partial\|\widehat{\mathbf{X}}\|_*$.

Suppose that the matrix $\widehat{\mathbf{X}}$ is rank one and that the sparsity pattern of $\widehat{\mathbf{X}}$ correctly recovers K_1 and K_2 . Denote $\widehat{\mathbf{X}} = \widehat{\alpha}\widehat{\mathbf{u}}\widehat{\mathbf{v}}'$. Then we have that $\widehat{\mathbf{S}}_{K_1K_2} = \mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}}$ where $\mathbf{s}_{\widehat{\mathbf{u}}} = \text{sign}(\widehat{\mathbf{u}}_{K_1})$ and $\mathbf{s}_{\widehat{\mathbf{v}}} = \text{sign}(\widehat{\mathbf{v}}_{K_2})$. Furthermore, from Lemma 5.7.14, we know that $\widehat{\mathbf{K}} = \widehat{\mathbf{u}}\widehat{\mathbf{v}}' + \widehat{\mathbf{Z}}$ with $\sigma_1(\mathbf{Z}) \leq 1$, $\widehat{\mathbf{u}}'\mathbf{Z} = 0$ and $\mathbf{Z}\widehat{\mathbf{v}} = 0$.

Observe that the problem Eq. 5.31 can be rewritten as

$$\max_{\mathbf{X} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}} \text{tr } \mathbf{A}'\mathbf{X} - \frac{1}{2} \text{tr } \mathbf{X}'\mathbf{X} - \lambda_1 \|\mathbf{X}\|_* - \lambda_2 \mathbf{1}'\mathbf{X}\mathbf{1}.$$

Under the assumption that $\widehat{\mathbf{X}} = \widehat{\alpha}\widehat{\mathbf{u}}\widehat{\mathbf{v}}'$ with $\widehat{\mathbf{u}} = (\widehat{\mathbf{u}}'_{K_1}, \mathbf{0}')'$ and $\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}'_{K_2}, \mathbf{0}')'$, the above equation becomes

$$\max_{\widehat{\alpha}, \widehat{\mathbf{u}}_{K_1}, \widehat{\mathbf{v}}_{K_2}} \widehat{\alpha}\widehat{\mathbf{u}}'_{K_1} \mathbf{A}_{K_1K_2} \widehat{\mathbf{v}}'_{K_2} - \widehat{\alpha}^2 - \frac{1}{2} \lambda_1 \widehat{\alpha} - \lambda_2 \widehat{\alpha} \widehat{\mathbf{u}}'_{K_1} \mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}} \widehat{\mathbf{v}}_{K_2} \quad \text{subject to } \|\widehat{\mathbf{u}}_{K_1}\|_2 = 1, \|\widehat{\mathbf{v}}_{K_2}\|_2 = 1. \quad (5.34)$$

The objective Eq. 5.31 is strongly convex, which implies that $\widehat{\alpha}$, $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are unique if the global solution is of rank one. This in turn implies that $\widehat{\mathbf{u}}_{K_1}$ and $\widehat{\mathbf{v}}_{K_2}$ are left and right singular vectors of $\mathbf{A}_{K_1K_2} - \lambda_2 \mathbf{s}_{\widehat{\mathbf{u}}}\mathbf{s}'_{\widehat{\mathbf{v}}}$. Setting $\lambda_2 = \frac{\beta}{2\sqrt{k_1k_2}}$ and $\alpha = \beta/2$, we observe that the results of Lemma 5.7.1 hold here. That is, under the conditions of Theorem 5.4.3, it holds that

$$\|\widehat{\mathbf{u}}_{K_1} - \mathbf{u}_{K_1}\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_1}{k_1k_2 \log(n_1 - k_2)(n_2 - k_2)}}\right)$$

and

$$\|\widehat{\mathbf{v}}_{K_2} - \mathbf{v}_{K_2}\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log k_2}{k_1k_2 \log(n_1 - k_2)(n_2 - k_2)}}\right)$$

with probability $1 - \mathcal{O}(k_1^{-1})$. With $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ fixed, the problem Eq. 5.34 can be explicitly solved for $\widehat{\alpha}$,

$$\widehat{\alpha} = \sigma_1(\alpha \mathbf{u}_{K_1} \mathbf{v}'_{K_2} + \mathbf{\Delta}_{K_1K_2}) - \lambda_1, \quad (5.35)$$

which gives us a constraint on the signal strength α and the tuning parameter λ_1 .

So far, we have constructed $\widehat{\mathbf{X}}_{K_1K_2}$ and $\widehat{\mathbf{S}}_{K_1K_2}$. We need to verify that there is a matrix $\widehat{\mathbf{Z}}$ that satisfies Eq. 5.32 by plugging back $\widehat{\mathbf{X}}_{K_1K_2}$ and $\widehat{\mathbf{S}}_{K_1K_2}$ into Eq. 5.33. We will construct

$$\widehat{\mathbf{Z}} = \begin{pmatrix} \widehat{\mathbf{Z}}_{K_1K_2} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{Z}}_{K_1^c K_2^c} \end{pmatrix}. \quad (5.36)$$

From Eq. 5.33, we observe that

$$(\hat{\alpha} + \lambda_1)\hat{\mathbf{u}}_{K_1}\hat{\mathbf{v}}'_{K_2} - \alpha\mathbf{u}_{K_1}\mathbf{v}'_{K_2} - \mathbf{\Delta}_{K_1K_2} = \lambda_1\hat{\mathbf{Z}}_{K_1K_2}.$$

It follows that we need $\lambda_1 = \Omega(\sqrt{k_1} + \sqrt{k_2})$ to ensure that $\sigma_1(\hat{\mathbf{Z}}_{K_1K_2}) \leq 1$.

We have already seen in the proof of Theorem 5.4.3 that $\hat{\mathbf{S}}_{K_1^c K_2} = \lambda_2^{-1}\mathbf{\Delta}_{K_1^c K_2}$, $\hat{\mathbf{S}}_{K_1 K_2^c} = \lambda_2^{-1}\mathbf{\Delta}_{K_1 K_2^c}$ and $\hat{\mathbf{S}}_{K_1^c K_2^c} = \lambda_2^{-1}\mathbf{\Delta}_{K_1^c K_2^c}$ are valid blocks of a subdifferential of the ℓ_1 norm. Plugging back into Eq. 5.33, it follows that $\hat{\mathbf{Z}}_{K_1^c K_2^c} = \mathbf{0}$.

We can conclude that under the conditions of Theorem 5.4.3 on the size of the bicluster and the signal strength β with $\lambda_1 = \mathcal{O}(\sqrt{k_1} + \sqrt{k_2})$ and $\lambda_2 = \frac{\beta}{2\sqrt{k_1 k_2}}$, the solution $\hat{\mathbf{X}}$ of Eq. 5.31 is of rank one and correctly recovers (K_1, K_2) .

We can also show a similar result to Theorem 5.4.4, which establishes that the signal strength β cannot be much smaller than the one given in Theorem 5.4.3. From Eq. 5.33 follows that

$$\sigma_1(\mathbf{\Delta}_{K_1^c K_2^c} - \lambda_2\hat{\mathbf{S}}_{K_1^c K_2^c}) \leq \lambda_1$$

is necessary for $\hat{\mathbf{X}}$ to be of rank one and to correctly recover (K_1, K_2) . From Eq. 5.35, we have that $\lambda < \sigma_1(\alpha\mathbf{u}_{K_1}\mathbf{v}'_{K_2} + \mathbf{\Delta}_{K_1K_2})$ for a solution to be of rank 1. Since $\sigma_1(\alpha\mathbf{u}_{K_1}\mathbf{v}'_{K_2} + \mathbf{\Delta}_{K_1K_2}) < \alpha + 2(\sqrt{k_1} + \sqrt{k_2})$ with high probability, we have that $\lambda < \alpha + 2(\sqrt{k_1} + \sqrt{k_2})$. However, it was shown in the proof of Theorem 5.4.4 that

$$\begin{aligned} \min_{\|\mathbf{s}_{K_1^c K_2^c}\|_\infty \leq 1} \max_{\|\mathbf{x}\|_2=1} \mathbf{x}' \left[\begin{pmatrix} \mathbf{0} & \mathbf{A}_{K_1^c K_2^c} \\ \mathbf{A}'_{K_1^c K_2^c} & \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{0} & \mathbf{S}_{K_1^c K_2^c} \\ \mathbf{S}'_{K_1^c K_2^c} & \mathbf{0} \end{pmatrix} \right] \mathbf{x} \\ > \alpha + 2(\sqrt{k_1} + \sqrt{k_2}) \end{aligned} \quad (5.37)$$

with probability tending to 1.

Chapter 6

Recovering Block-structured Activations Using Compressive Measurements

In this chapter of the thesis we consider the problems of detection and localization of a contiguous block of weak activation in a large matrix, from a small number of noisy, possibly adaptive, compressive (linear) measurements. This is closely related to the problem of compressed sensing, where the task is to estimate a sparse vector using a small number of linear measurements. Contrary to results in compressed sensing, where it has been shown that neither adaptivity nor contiguous structure help much, we show that for reliable *localization* the magnitude of the weakest signals is strongly influenced by both structure and the ability to choose measurements adaptively while for *detection* neither adaptivity nor structure reduce the requirement on the magnitude of the signal. We characterize the precise tradeoffs between the various problem parameters, the signal strength and the number of measurements required to reliably detect and localize the block of activation. In each case the sufficient conditions are complemented with information theoretic lower bounds.

6.1 Introduction

Compressive measurements provide a very efficient means of recovering signals that are sparse in some basis or frame. Specifically, several papers, including those of Candès and Tao [40, 41], Donoho [64], and Candès and Wakin [42] have shown that it is possible to recover, in an ℓ_2 sense, a k -sparse vector in n dimensions using only $\mathcal{O}(k \log n)$ incoherent compressive measurements, instead of measuring all of the n coordinates. This is a novel and important paradigm with applications in a wide range of scientific areas. Along with ℓ_2 recovery, researchers have also considered the problems of *detection* and *localization* of a sparse signal corrupted by additive noise, the former task logically preceding the latter. The problem of detection is to test whether all components of the vector are zero. Arias-Castro et al. [10], Duarte et al. [66], Haupt and Nowak [94], Ingster et al. [99] and Arias-Castro [8] studied detection of sparse vectors from compressive measurements. The problem of localization is to identify coordinates of the

non-zero elements of a signal. Wainwright [197, 198] studied information theoretic limits and localization properties of the LASSO procedure. More recently, researchers have contributed two important refinements: 1) by considering a sparse *structured* signal (such as a signal consisting of adjacent coordinates or a block) [9, 23, 175] and 2) by allowing for the possibility of taking adaptive measurements, i.e., where subsequent measurements are designed based on past observations [see, e.g., 9, 39, 56, 93, 136]. However, almost all of this work has been focused on recovery or detection of (structured or unstructured) sparse data *vectors* from (passive or adaptive) compressed measurements.

In this chapter we focus on the unexplored problems of detection and localization for data matrices from compressive measurements. We are concerned with signals that are both sparse and highly structured, taking the form of a sub-matrix of a larger matrix with contiguous row and column indices. Data matrices have been considered in the context of low-rank matrix completion [see, e.g., 112, 144], where recovery in Frobenius norm is studied. The problems of detection and localization for data matrices that are observed directly were studied previously. See, for example, [29, 37, 38, 111, 182]. However, compressive measurement schemes were not investigated. If the activation is unstructured, the treatment of data matrices is exactly equivalent to the treatment of data vectors. However, in the structured case the problem is rather different, as we will show. Data matrices with signals that are both sparse and highly structured form a natural model for several real-world activations such as when we have a group of genes (belonging to a common pathway for instance) co-expressed under the influence of a set of similar drugs [208], or when we have groups of patients exhibiting similar symptoms [140], or when we have sets of malware with similar signatures [100], etc. However, in many of these applications, it is difficult to measure, compute or store all the entries of the data matrix. For example, measuring expression levels of all genes under all possible drugs is expensive, or recording the signatures of each individual malware is computationally demanding as it might require stepping through the entire malware code. However, if we have access to linear combinations of matrix entries (i.e. compressive measurements) such as combined expression of multiple genes under the influence of multiple drugs then we might need to only make and store few such measurements, while still being able to infer the existence or location of the activated block of the data matrix. Thus, the goal is to detect or recover the activated block (set of co-expressed genes and drugs or malware with similar signatures) using only few compressive measurements of the data matrix, instead of observing the entire data matrix directly. We consider both the passive (non-adaptive) and active (adaptive) measurements. The non-adaptive measurements are random or pre-specified linear combinations of matrix entries. In other cases, such as mixing drugs, we might be able to adapt the measurement process by using feedback to sequentially design linear combinations that are more informative.

Extensions to a setup where there is a non-contiguous sub-matrix or block of activation are also interesting, but beyond the scope of this thesis. Bhamidi et al. [29], Butucea and Ingster [37], Butucea et al. [38], Kolar et al. [111], Sun and Nobel [182] study a problem where a large noisy matrix is observed directly, i.e., not through compressed measurements, and the block of activation is non-contiguous. In such a setting, tight upper and lower bounds are derived for the localization problem. However, passive and adaptive *compressive* measurement schemes have to the best of our knowledge not yet been investigated.

Table 6.1: Summary of known results for the sparse vector case, where the length of the vector is n and the number of active elements is k . The number of measurements is m and μ/σ represents SNR per element of the activated elements.

	Detection	Localization
Passive	$\frac{\mu}{\sigma} \asymp \sqrt{\frac{n}{mk^2}}$	$\frac{\mu}{\sigma} \asymp \sqrt{\frac{n \log n}{m}}$, $m \succ k \log n$ Wainwright [197]
Active	Arias-Castro [8]	Arias-Castro et al. [9] Davenport and Arias-Castro [56] Malloy and Nowak [136]

Summary of results in this chapter. Using information theoretic tools, we establish *lower bounds* on the minimum number of compressive measurements and the weakest signal-to-noise ratio (SNR) needed to detect the presence of an activated block of positive activation, as well as to localize the activated block, using both non-adaptive and adaptive measurements. We also demonstrate minimax optimal *upper bounds* through detectors and estimators that can guarantee consistent detection and localization of weak block-structured activations using few non-adaptive and adaptive compressive measurements.

Our results indicate that adaptivity and structure play a key role and provide significant improvements over non-adaptive and unstructured cases for localization of the activated block in the data matrix setting. This is unlike the vector case where contiguous structure and adaptivity have been shown to provide minor, if any, improvement. We describe the results for the sparse vector case in related work section below. A summary of the SNR needed for detection and localization of an unstructured sparse vector using passive and adaptive compressive measurements is given in Table 6.1.

In our setting we take compressive measurements of a data *matrix* of size $n = (n_1 \times n_2)$, the activated block is of size $k = (k_1 \times k_2)$, with minimum SNR per entry of μ/σ , and we have a budget of m compressive measurements with each measurement matrix constrained to have unit Frobenius norm. Table 6.2 describes our main findings (for the case when $n_1 = n_2$ and $k_1 = k_2$ and paraphrasing for clarity) and compare the scalings under which passive and active, detection and localization are possible.

For detection, akin to the vector setting, structure and adaptivity play no role. The structured data matrix setting requires an SNR scaling of $\sqrt{n_1 n_2 / (m k_1^2 k_2^2)}$ for both non-adaptive and adaptive cases, which is same as the SNR needed to detect a $k_1 k_2$ -sparse non-negative vector of length $n_1 n_2$ as demonstrated in the paper [8]. Thus, the structure of the activation pattern as well as the power of adaptivity offer no advantage in the detection problem.

For localization of the activated block, the structured data matrix setting requires an SNR scaling as $\sqrt{n_1 n_2 / (m \min(k_1, k_2))}$ using non-adaptive compressive measurements. In contrast, the unstructured setting requires a higher SNR of $\sqrt{n_1 n_2 \log(n_1 n_2) / m}$ where $m \geq k_1 k_2 \log(n_1 n_2)$ as demonstrated in the paper [197]. Structure, without adaptivity already yields a factor of

Table 6.2: Summary of main findings for the case when $n = n_1 \times n_2$ ($n_1 = n_2$) and $k = k_1 \times k_2$ ($k_1 = k_2$), where the size of the matrix is $n_1 \times n_2$ and the size of the activation block is $k_1 \times k_2$. The number of measurements is m and μ/σ represents SNR per element of the activated block.

	Detection	Localization	
Passive	$\frac{\mu}{\sigma} \asymp \sqrt{\frac{n_1 n_2}{m k_1^2 k_2^2}}$	$\frac{\mu}{\sigma} \asymp \sqrt{\frac{n_1 n_2}{m \min(k_1, k_2)}}$	Theorems 6.4.1 and 6.4.2
Active	Theorems 6.3.1 and 6.3.2	$\frac{\mu}{\sigma} \asymp \frac{1}{\sqrt{m}} \max\left(\sqrt{\frac{n_1 n_2}{k_1^2 k_2^2}}, \frac{1}{\sqrt{\min(k_1, k_2)}}\right)$	Theorems 6.5.1 and 6.5.2

$\sqrt{\min(k_1, k_2)}$ reduction in the smallest SNR that still allows for reliable localization. Moreover, adaptivity in the compressive measurement design yields further improvements: with adaptive measurements, identifying the activated block requires a much weaker SNR of

$$\max\left(\sqrt{n_1 n_2 / (m k_1^2 k_2^2)}, \sqrt{1 / (m \min(k_1, k_2))}\right)$$

for the weakest entry in the data matrix. In contrast, for the sparse vector case, Arias-Castro et al. [9] showed that adaptive compressive measurements cannot localize the non-zero components if the SNR is smaller than $\sqrt{n_1 n_2 / m}$. A matching upper bound was provided using compressive binary search in the papers [56] and [136] for localization of a single non-zero entry in the vector. Thus, exploiting structure of the activations and designing adaptive linear measurements can both yield significant gains if the activation corresponds to a contiguous block in a data matrix.

Related Work. This chapter of the thesis builds on a number of fairly recent contributions on detection, localization and recovery of a sparse and weak unstructured signal by adaptive compressive measurements. In the paper [9], the authors show that the adaptive compressive scheme offers improvements over the passive scheme which, in terms of the mean-squared error (MSE) and localization, are limited to a $\log(n)$ factor. The authors also provide a general proof strategy for minimax analysis under adaptive measurements. Arias-Castro [8] further applies this strategy to the problem of detection of an unstructured and structured sparse and weak vector signal under compressive adaptive measurements. Malloy and Nowak [136] shows that a compressive version of standard binary search achieves minimax performance for localization in a one-sparse vector. The work of Wainwright [197] which is based on analyzing the performance of an exhaustive search procedure under passive measurements, is relevant to our analysis of passive localization. Our analysis provides a generalization of these results to the case of a *structured* signal embedded as a small contiguous block in a large matrix.

While in this chapter we focus on detection and localization, some other papers have considered estimation of sparse vectors in the MSE sense using adaptive compressive measurements. For example, Arias-Castro et al. [9] establishes fundamental lower bounds on the MSE in a linear regression framework, while Haupt et al. [93] demonstrates upper bounds using compressive distilled sensing. Baraniuk et al. [23] and Soni and Haupt [175] have analyzed different forms of structured sparsity in the vector setting, e.g. if the non-zero locations in a data vector form non-overlapping or partially-overlapping groups or are tree-structured. Finally, Negahban and

Wainwright [144] and Koltchinskii et al. [112] have considered a measurement model identical to ours in the setting of low-rank matrix completion, but in that setting the matrix under consideration is not assumed to be a structured sparse matrix and the theoretical guarantees are with respect to the Frobenius norm. Furthermore, Kolar et al. [111] illustrate that penalization using the sum of nuclear and ℓ_1 norm cannot be used for localization in a related model.

When data matrix is observed directly, Butucea and Ingster [37] study the problem of detection, while Kolar et al. [111] and Butucea et al. [38] study the problem of localization. Sun and Nobel [182] and Bhamidi et al. [29] characterize largest average submatrices of the data matrix under the null hypothesis that the signal is not present. Results in those papers do not carry over to a setting where a data matrix is accessed through compressive measurements, as already seen in the vector case [8].

The rest of this chapter is organized as follows. We describe the problem set up and notation in Section 6.2. We study the detection problem in Section 6.3, for both adaptive and non-adaptive schemes. Section 6.4 is devoted to the non-adaptive localization, while Section 6.5 is focused on adaptive localization. Finally, in Section 6.6 we present and discuss some simulations that support our findings. Detailed proofs are given in Section 6.7.

6.2 Preliminaries

In this chapter we denote $[n]$ to be the set $\{1, \dots, n\}$. For a vector $\mathbf{a} \in \mathbb{R}^n$, we denote $\text{supp}(\mathbf{a}) = \{j : a_j \neq 0\}$ the support set, $\|\mathbf{a}\|_q$, $q \in [1, \infty)$, the ℓ_q -norm defined as $\|\mathbf{a}\|_q = (\sum_{i \in [n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $\|\mathbf{a}\|_0 = |\text{supp}(\mathbf{a})|$ and $\|\mathbf{a}\|_\infty = \max_{i \in [n]} |a_i|$. For a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we denote $\|\mathbf{A}\|_F$ the Frobenius norm defined as

$$\|\mathbf{A}\|_F = \left(\sum_{i \in [n_1], j \in [n_2]} a_{ij}^2 \right)^{1/2}.$$

For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = \mathcal{O}(b_n)$ to denote that $a_n < Cb_n$ for some finite positive constant C . We also denote $a_n = \mathcal{O}(b_n)$ to be $b_n \gtrsim a_n$. If $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$, we denote it to be $a_n \asymp b_n$. The notation $a_n = o(b_n)$ is used to denote that $a_n b_n^{-1} \rightarrow 0$.

Let $A \in \mathbb{R}^{n_1 \times n_2}$ be a signal matrix with unknown entries. We are interested in a highly *structured* setting where a *contiguous* block of the matrix A of size $(k_1 \times k_2)$ has entries all equal to $\mu > 0$, while all the other elements of A are equal to zero. We denote the coordinate set of all contiguous blocks, of size $k_1 \times k_2$ with

$$\mathcal{B} = \left\{ I_r \times I_c : \begin{array}{l} I_r \text{ and } I_c \text{ are contiguous subsets of } [n_1] \text{ and } [n_2], \\ |I_r| = k_1, |I_c| = k_2 \end{array} \right\}. \quad (6.1)$$

Then $A = (a_{ij})$ with $a_{ij} = \mu \mathbb{I}\{(i, j) \in B^*\}$ for some (unknown) $B^* \in \mathcal{B}$, where \mathbb{I} is the indicator function. Some of our results extend to the case when the activation is positive, but not constant on B^* , as we discuss below. Note that we assume the size $(k_1 \times k_2)$ is known.

We consider the following observation model under which m noisy linear measurements of A are available

$$y_i = \text{tr}(AX_i) + \epsilon_i, \quad i = 1, \dots, m, \quad (6.2)$$

where $\epsilon_1, \dots, \epsilon_m \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$, with $\sigma > 0$ known, and the sensing matrices $(X_i)_{i \in [m]}$ are normalized to satisfy either $\|X_i\|_F \leq 1$ or $\mathbb{E}\|X_i\|_F^2 = 1$, i.e., every measurement has the same amount of energy. These are similar assumptions as made in the papers [56] and [39].

Under the observation model in Eq. 6.2, we study two tasks: (1) detecting whether a contiguous block of positive signal exists in A and (2) identifying the block B^* , that is, the localization of B^* . We develop efficient algorithms for these two tasks that provably require the smallest number of measurements, as explained below. The algorithms are designed for one of two measurement schemes: (1) the measurement scheme can be implemented in an adaptive or sequential fashion, that is, actively, by letting each X_i to be a (possibly randomized) function of $(y_j, X_j)_{j \in [i-1]}$, and (2) the measurement matrices are chosen all at once or ignoring the outcomes in previous measurements, that is, passively.

Detection. The detection problem concerns checking whether a positive contiguous block exists in A . As we will show later, we can detect the presence of a contiguous block with a much smaller number of measurements than is required for localizing its position. Formally, detection is a hypothesis testing problem with a composite alternative of the form

$$\begin{aligned} H_0: & \quad A = 0_{n_1 \times n_2} \\ H_1: & \quad A = (a_{ij}) \text{ with } a_{ij} = \mu \mathbb{I}_{\{(i,j) \in B^*\}}, \quad B^* \in \mathcal{B}. \end{aligned} \quad (6.3)$$

A test T is a measurable function of the observations $(y_i)_{i \in [m]}$ and the measurements matrices $(X_i)_{i \in [m]}$, which takes values in $\{0, 1\}$, with $T = 1$ if the null hypothesis is rejected and $T = 0$ otherwise. For any test T , we define its risk as

$$R^{\text{det}}(T) \equiv \mathbb{P}_0 [T((y_i, X_i)_{i \in [m]}) = 1] + \max_{B^* \in \mathcal{B}} \mathbb{P}_{B^*} [T((y_i, X_i)_{i \in [m]}) = 0],$$

where \mathbb{P}_0 and \mathbb{P}_B denote the joint probability distributions of $((y_i, X_i)_{i \in [m]})$ under the null hypothesis and when the activation pattern is B , respectively. The risk $R(T)$ measures the maximal sum of type I and type II errors over the set of alternatives. The overall difficulty of the detection problem is quantified by the *minimax risk*

$$R^{\text{det}} \equiv \inf_T R^{\text{det}}(T),$$

where the infimum is taken over all tests. For a sufficiently small SNR, the minimax risk is bounded away from zero by a large constant, which implies that no test can distinguish H_0 from H_1 . In Section 6.3 we will precisely characterize the boundary for SNR $\frac{\mu}{\sigma}$ below which no test can distinguish H_0 and H_1 .

Localization. The localization problem concerns the recovery of the true activation pattern B^* . Let Ψ be an estimator of B^* , i.e., a measurable function of $(y_i, X_i)_{i \in [m]}$ taking values in \mathcal{B} . We define the risk of any such estimator as

$$R^{\text{loc}}(\Psi) = \max_{B^* \in \mathcal{B}} P_{B^*} [\Psi((y_i, X_i)_{i \in [m]}) \neq B^*],$$

while the *minimax risk*

$$R^{\text{loc}} \equiv \inf_{\Psi} R^{\text{loc}}(\Psi)$$

of the localization problem is the minimal risk over all such estimators Ψ . Like in the detection task, the minimax risk specifies the minimal risk of any localization procedure. By standard arguments, the evaluation of the minimax localization risk also proceeds by first reducing the localization problem to a hypothesis testing problem (see for example the book by Tsybakov [192]).

Below we will provide a sharp characterization, through information theoretic lower bounds and tractable estimators, of the minimax detection and localizations risks as functions of tuples of $(n_1, n_2, k_1, k_2, m, \mu, \sigma)$ and for both the active and passive sampling schemes. Our results identify precisely both the minimal SNR given a budget of m possibly adaptive measurements, and the minimal number of measurements m for a given SNR in order to achieve successful detection and localization.

Along with a careful and detailed minimax analysis, we also describe procedures for detection and localization in both the active and passive case whose risks match the minimax rates.

6.3 Detection of contiguous blocks

In this section, we derive minimax rates for detection.

6.3.1 Lower bound

The following theorem gives a lower bound on the SNR needed to distinguish H_0 and H_1 .

Theorem 6.3.1. *Fix any $0 < \alpha < 1$. Based on m (possibly adaptive) measurements, if*

$$\mu \leq \sigma(1 - \alpha) \sqrt{\frac{16(n_1 - k_1)(n_2 - k_2)}{mk_1^2 k_2^2}},$$

then $R^{\text{det}} \geq \alpha$.

The lower bound on possibly *adaptive* procedures is established by analyzing the risk of the (optimal) likelihood ratio test under a uniform prior over the alternatives. Careful modifications of standard arguments are necessary to account for adaptivity. We closely follow the approach of Arias-Castro [8] who established the analogue of Theorem 6.3.1 in the vector setting.

6.3.2 Upper bound

We now demonstrate the sharpness of the result established in the previous section. We choose the sensing matrices passively as $X_i = (n_1 n_2)^{-1/2} \mathbf{1}_{n_1} \mathbf{1}'_{n_2}$ and consider the following test

$$T((y_i)_{i \in [m]}) = \mathbb{I} \left\{ \sum_i y_i > \sigma \sqrt{2m \log(\alpha^{-1})} \right\}. \quad (6.4)$$

Theorem 6.3.2. *Assume that $k_1 \leq cn_1$ and $k_2 \leq cn_2$ for some $c \in (0, 1)$. If*

$$\mu \geq \sigma \sqrt{\frac{8n_1 n_2 \log(\alpha^{-1})}{mk_1^2 k_2^2}},$$

then $R^{\det}(T) \leq \alpha$, where T is the test defined in Eq. 6.4.

The results of Theorem 6.3.1 and Theorem 6.3.2 establish that the minimax rate for detection under the model in Eq. 6.2 is $\mu \asymp \sigma (k_1 k_2)^{-1} \sqrt{m^{-1} n_1 n_2}$, under the (mild) assumption that $k_1 \leq cn_1$ and $k_2 \leq cn_2$ for any constant $0 < c < 1$. It is worth pointing out that the structure of the activation pattern *does not* play any role in the minimax detection problem, since the rate matches the known bounds for detection in the unstructured vector case [8]. We will contrast this to the localization problem below. Furthermore, the procedure that achieves the adaptive lower bound (upto constants) is non-adaptive, indicating that adaptivity can not help much in the detection problem.

We also note that results established in this section continue to hold when the activation is positive, but not constant on B^* , with $\min_{(i,j) \in B^*} a_{ij}$ replacing μ .

6.4 Localization from passive measurements

In this section, we address the problem of estimating a contiguous block of activation B^* from noisy linear measurements as in equation 6.2, when the measurement matrices $(X_i)_{i \in [m]}$ are independent with i.i.d. entries having a $\mathcal{N}(0, (n_1 n_2)^{-1})$ distribution. The variance of the elements is set so that $\mathbb{E} \|X_i\|_F^2 = 1$.

6.4.1 Lower bound

The following theorem gives a lower bound on the SNR needed for any procedure to localize B^* .

Theorem 6.4.1. *There exist positive constants $C, \alpha > 0$ independent of the problem parameters (k_1, k_2, n_1, n_2) , such that if*

$$\mu \leq C \sigma \sqrt{\frac{n_1 n_2}{m} \max \left(\frac{1}{\min(k_1, k_2)}, \frac{\log \max(n_1 - k_1, n_2 - k_2)}{k_1 k_2} \right)},$$

then $R^{\text{loc}} \geq \alpha > 0$.

The proof is based on a standard technique described in Chapter 2.6 of the book by Tsybakov [192]. We start by identifying a subset of matrices that are hard to distinguish. Once a suitable finite set is identified, tools for establishing lower bounds on the error in multiple-hypothesis testing can be directly applied. These tools only require computing the Kullback-Leibler (KL) divergence between the induced distributions, which in our case are two multivariate normal distributions.

The two terms in the lower bound feature two aspects of our construction, the first term arises from considering two matrices that overlap considerably, while the second term arises from considering matrices that do not overlap at all of which there are possibly a very large number. These constructions and calculations are described in detail in Section 6.7.

6.4.2 Upper bound

We will investigate a procedure that searches over all contiguous blocks of size $(k_1 \times k_2)$ as defined in Eq. 6.1 and outputs the one minimizing the squared error. Specifically, let the loss function $f : \mathcal{B} \mapsto \mathbb{R}$ be

$$f(B) := \min_{\mu} \sum_{i \in [m]} \left(\mu \sum_{(a,b) \in B} X_{i,ab} - y_i \right)^2, \quad (6.5)$$

where $X_{i,ab}$ denotes element in row a and column b of the i^{th} sensing matrix. Then the estimated block \hat{B} is defined as

$$\hat{B} := \operatorname{argmin}_{B \in \mathcal{B}} f(B). \quad (6.6)$$

Note that the minimization problem above requires solving $O(n_1 n_2)$ univariate regression problems and can be implemented efficiently for reasonably large matrices.

The following result characterizes the SNR needed for \hat{B} to correctly identify B^* .

Theorem 6.4.2. *There exist positive constants $C_1, C_2 > 0$ independent of the problem parameters (k_1, k_2, n_1, n_2) , such that if $m \geq C_1 \log \max(n_1 - k_1, n_2 - k_2)$ and*

$$\mu \geq C_2 \sigma \sqrt{\frac{n_1 n_2}{m} \log(2/\alpha) \max \left(\frac{\log \max(k_1, k_2)}{\min(k_1, k_2)}, \frac{\log \max(n_1 - k_1, n_2 - k_2)}{k_1 k_2} \right)},$$

for $0 < \alpha \leq 1$, then $R^{\text{loc}}(\hat{B}) \leq \alpha$, where \hat{B} is defined in Eq. 6.6.

Comparing to the lower bound in Theorem 6.4.1, we observe that the procedure outlined in this section achieves the lower bound up to constants and a $\log(\max(k_1, k_2))$ factor. Under the scaling $\max(k_1, k_2) \geq \log \max(n_1 - k_1, n_2 - k_2)$, we obtain that the *passive* minimax rate for localization of the active blocks B^* is $\mu \asymp \tilde{O}(\sigma \sqrt{(m \min(k_1, k_2))^{-1} n_1 n_2})$. In this and subsequent uses, the \tilde{O} notation hides a $\sqrt{\log \max(k_1, k_2)}$ factor.

This establishes that the SNR needed for passive localization is considerably larger than the bound we saw earlier for passive detection. This should be contrasted to the unstructured normal means problem, where the bounds for localization and detection differ only in constants [65].

The block structure of the activation allows us, even in the passive setting, to localize much weaker signals. A straightforward adaptation of results on the LASSO [198] suggest that if the non-zero entries are spread out (say at random) then we would require $\mu \asymp O\left(\sigma\sqrt{\frac{n_1 n_2}{m}}\right)$ for localization.

One could extend the analysis in this section to data matrices with non-constant activation as in the paper [197]. Furthermore, one can adapt to the unknown size of the activation block. In particular, one can perform exhaustive search procedure for all possible sizes of activation blocks. Let \mathcal{B}_{k_1, k_2} denote the coordinate set of all contiguous blocks of size $k_1 \times k_2$. Then the estimated block

$$\widehat{B} = \underset{B \in \cup_{k_1, k_2} \mathcal{B}_{k_1, k_2}}{\operatorname{argmin}} f(B)$$

adapts to the unknown size of the activation if the signal strength satisfies the condition in Theorem 6.4.2. This can be verified by small modifications to the proof of Theorem 6.4.2.

The non-contiguous case

Suppose that the block of activation B^* belongs to the collection $\widetilde{\mathcal{B}}$, where

$$\widetilde{\mathcal{B}} = \{I_r \times I_c : I_r \subset [n_1], I_c \subset [n_2], |I_r| = k_1, |I_c| = k_2\},$$

so that the activation block is not necessarily a contiguous block. This collection contains less structure than the collection \mathcal{B} , but we can still localize much weaker signals compared to completely unstructured case. Slight modification of proofs¹ of Theorem 6.4.1 and Theorem 6.4.2 yields the following.

Theorem 6.4.3. *Let $\widetilde{B} := \operatorname{argmin}_{B \in \widetilde{\mathcal{B}}} f(B)$. There exists a constant C_1 such that if the signal strength satisfies*

$$\mu \geq C_1 \sigma \sqrt{\frac{n_1 n_2}{m} \log(2/\alpha) \frac{\log(n_1 - k_1)(n_2 - k_2)}{k_1 + k_2}}, \quad (6.7)$$

then $R^{\operatorname{loc}}(\widetilde{B}) \leq \alpha$, for any $0 < \alpha \leq 1$.

Conversely, there exists constants $C_2, \alpha > 0$ such that if

$$\mu \leq C_2 \sigma \sqrt{\frac{n_1 n_2}{m} \max\left(\frac{\log(n_1 - k_1)}{k_2}, \frac{\log(n_2 - k_2)}{k_1}, \frac{\log\binom{n_1 - k_1}{k_1} \binom{n_2 - k_2}{k_2}}{k_1 k_2}\right)}, \quad (6.8)$$

then $R^{\operatorname{loc}} \geq \alpha > 0$.

¹A sketch of the derivation is given in Section 6.7.7

Therefore, we conclude that even without contiguous blocks, the additional structure helps for the problem of localization.

6.5 Localization from active measurements

In this section, we study localization of B^* using adaptive procedures, that is, the measurement matrix X_i may be a function of $(y_j, X_j)_{j \in [i-1]}$.

6.5.1 Lower bound

A lower bound on the SNR needed for any active procedure to localize B^* is given as follows.

Theorem 6.5.1. *Fix any $0 < \alpha < 1$. Given m adaptively chosen measurements, if*

$$\mu < \sigma(1-\alpha) \max \left(\sqrt{\frac{2 \max((n_1 - k_1)(n_2/2 - k_2), (n_1/2 - k_1)(n_2 - k_2))}{mk_1^2 k_2^2}}, \sqrt{\frac{8}{m \min(k_1, k_2)}} \right)$$

then $R^{\text{loc}} \geq \alpha$.

The proof is based on information theoretic arguments applied to specific pairs of hypotheses that are hard to distinguish. The two terms in the lower bound reflect the two important sources of hardness of the problem of localization. The first term reflects the difficulty of approximately localizing the block of activation. This term grows at the same rate as the detection lower bound, and its proof is similar. Given a coarse localization of the block we still need to exactly localize the block. The hardness of this problem gives rise to the second term in the lower bound. The term is independent of n_1 and n_2 but has a considerably worse dependence on k_1 and k_2 .

6.5.2 Upper bound

The upper bound is established by analyzing the procedures described in Algorithms 1 and 2 for approximate and exact localization. Algorithm 1 is used to approximately locate the activation block, that is, it locates a $8k_1 \times 8k_2$ block that contains the activation block with high probability. The algorithm essentially performs compressive binary search ([56]) on a collection of non-overlapping blocks that partition the matrix. It is run on four collections, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ and \mathcal{D}_4

Algorithm 3 Approximate localization

input Measurement budget $m \geq \log p$, ordered collection of size^a p of blocks \mathcal{D} of size $(u_1 \times u_2)$

Initial support: $J_0^{(1)} \equiv \{1, \dots, p\}$, $s_0 \equiv \log p$

For each s in $1, \dots, \log_2 p$

1. Allocate: $m_s \equiv \lfloor (m - s_0) s 2^{-s-1} \rfloor + 1$.
2. Split: $J_1^{(s)}$ and $J_2^{(s)}$, left and right half collections of blocks of $J_0^{(s)}$.
3. Sensing matrix: $X_s = \sqrt{\frac{2^{-(s_0-s+1)}}{u_1 u_2}}$ on $J_1^{(s)}$, $X_s = -\sqrt{\frac{2^{-(s_0-s+1)}}{u_1 u_2}}$ on $J_2^{(s)}$ and 0 otherwise.
4. Measure: $y_i^{(s)} = \text{tr}(AX_s) + z_i^{(s)}$ for $i \in [1, \dots, m_s]$.
5. Update support: $J_0^{(s+1)} = J_1^{(s)}$ if $\sum_{i=1}^{m_s} y_i^{(s)} > 0$ and $J_0^{(s+1)} = J_2^{(s)}$ otherwise.

output The single block in $J_0^{(s_0+1)}$.

^aWe assume p is dyadic to simplify our presentation of the algorithm.

Algorithm 4 Exact localization (of columns)

input Measurement budget m , a sub-matrix $B \in \mathbb{R}^{4k_1 \times 4k_2}$, success probability δ

1. Measure: $y_i^c = (4k_1)^{-1/2} \sum_{l=1}^{4k_1} B_{lc} + z_i^c$ for $i = \{1, \dots, m/5\}$ and $c \in \{1, k_2 + 1, 2k_2 + 1, 3k_2 + 1\}$.
2. Let $l = \text{argmax}_c \sum_{i=1}^{m/5} y_i^c$, $r = l + k_2$, $m_b = \lfloor \frac{m}{6 \log_2 k_2} \rfloor$.
3. While $r - l \geq 1$
 - (a) Let $c = \lfloor \frac{r+l}{2} \rfloor$.
 - (b) Measure $y_i^c = (4k_1)^{-1/2} \sum_{l=1}^{4k_1} B_{lc} + z_i^c$ for $i = \{1, \dots, m_b\}$.
 - (c) If^a $\sum_{i=1}^{m_b} y_i^c \geq \mathcal{O} \left(\sqrt{\log \left(\frac{\log k_2}{\delta} \right) \frac{m_b \sigma^2}{\log k_2}} \right)$ then $l = c$, otherwise $r = c$.

output Set of columns $\{l - k_2 + 1, \dots, l\}$.

^aThe exact constants appear in the proof of Theorem 6.5.2.

defined as²

$$\begin{aligned}
\mathcal{D}_1 &\equiv \{B_{1,1} := [1, \dots, 2k_1] \times [1, \dots, 2k_2], B_{1,2} := [2k_1 + 1, \dots, 4k_1] \times [1, \dots, 2k_2] \\
&\quad \dots, B_{1,n_1 n_2 / 4k_1 k_2} := [n_1 - 2k_1, \dots, n_1] \times [n_2 - 2k_2, \dots, n_2]\} \\
\mathcal{D}_2 &\equiv \{B_{2,1} := [k_1, \dots, 3k_1] \times [k_2, \dots, 3k_2], B_{2,2} := [3k_1 + 1, \dots, 5k_1] \times [k_2, \dots, 3k_2] \\
&\quad \dots, B_{2,n_1 n_2 / 4k_1 k_2} := [n_1 - k_1, \dots, n_1, 1, \dots, k_1] \times [n_2 - k_2, \dots, n_2, 1, \dots, k_2]\} \\
\mathcal{D}_3 &\equiv \{B_{3,1} := [k_1, \dots, 3k_1] \times [1, \dots, 2k_2], B_{3,2} := [3k_1 + 1, \dots, 5k_1] \times [1, \dots, 2k_2] \\
&\quad \dots, B_{3,n_1 n_2 / 4k_1 k_2} := [n_1 - k_1, \dots, n_1, 1, \dots, k_1] \times [n_2 - 2k_2, \dots, n_2]\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{D}_4 &\equiv \{B_{4,1} := [1, \dots, 2k_1] \times [k_2, \dots, 3k_2], B_{4,2} := [2k_1 + 1, \dots, 4k_1] \times [k_2, \dots, 3k_2] \\
&\quad \dots, B_{4,n_1 n_2 / 4k_1 k_2} := [n_1 - 2k_1, \dots, n_1] \times [n_2 - k_2, \dots, n_2, 1, \dots, k_2]\}.
\end{aligned}$$

\mathcal{D}_1 is a partition of the matrix into disjoint blocks of size $(2k_1 \times 2k_2)$, \mathcal{D}_3 is a similar partition shifted down by k_1 rows, \mathcal{D}_4 is shifted to the right by k_2 columns and \mathcal{D}_2 is both shifted down by k_1 rows and to the right by k_2 columns. Figure 6.1 illustrates this.

Notice, that one of these collections must include a block that contains the *full* block of activation. Algorithm 1 applied four times returns four blocks, one of which as we show contains the full activation block with high probability.

Algorithm 2 is used next to precisely locate the activation block within one of the four coarser blocks identified by Algorithm 1. Algorithm 2 itself works in several stages: in the first stage the procedure measures a small number of columns, exactly one of which is active, repeatedly, to identify the active column with high probability. The next stage finds the first non-active column to the left and right by testing columns using a binary search (halving) procedure. In this way, all the active columns are located. Finally, Algorithm 2 is repeated on the rows to identify the active rows.

The following theorem states that Algorithm 1 and Algorithm 2 succeed in localization of the active block with high probability if the SNR is large enough.

Theorem 6.5.2. *If*

$$\mu \geq \sigma \sqrt{\log(1/\alpha)} \tilde{O} \left(\max \left(\sqrt{\frac{n_1 n_2}{m k_1^2 k_2^2}}, \sqrt{\frac{1}{\min(k_1, k_2) m}} \right) \right),$$

and $m \geq 3 \log(n_1 n_2)$ then $R(\hat{B}) \leq \alpha$, where \hat{B} is the block output by the algorithms.

As before, the \tilde{O} hides a $\sqrt{\log \max(k_1, k_2)}$ factor, and our upper bound matches the lower bound up to this factor. It is worth noting that for small activation blocks (when the first term dominates) our active localization procedure achieves the *detection* limits. This is the best result we could hope for. For larger activation blocks, the lower bound indicates that *no* procedure can achieve the detection rate. The active procedure still remains significantly more efficient than the passive

²For simplicity, we assume n_1 is a multiple of $2k_1$ and n_2 of $2k_2$

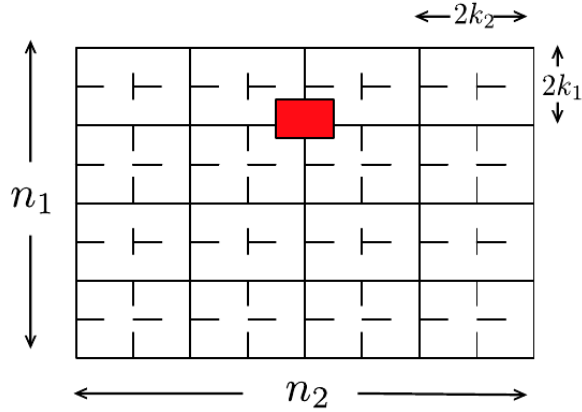


Figure 6.1: The collection of blocks \mathcal{D}_1 is shown in solid lines and the collection \mathcal{D}_2 is shown in dashed lines. The collections \mathcal{D}_3 and \mathcal{D}_4 overlap with these and are not shown. The $(k_1 \times k_2)$ block of activation is shown in red.

one, and even in this case is able to localize signals that are weaker by a (large) $\sqrt{n_1 n_2}$ factor. This is not the case for compressed sensing of vectors as shown in the paper of Arias-Castro et al. [9]. The great potential for gains from adaptive measurements is clearly seen in our model which captures the fundamental interplay between *structure* and *adaptivity*.

6.6 Experiments

In this section, we perform a set of simulation studies to illustrate finite sample performance of the proposed procedures. We let $n_1 = n_2 = n$ and $k_1 = k_2 = k$. Theorem 6.4.2 and Theorem 6.5.2 characterize the SNR needed for the passive and active identification of a contiguous block, respectively. We demonstrate that the scalings predicted by these theorems are sharp by plotting the probability of successful recovery against appropriately rescaled SNR and showing that the curves for different values of n and k line up.

Experiment 1. Figure 6.2 shows the probability of successful localization of B^* using \hat{B} defined in Eq. 6.6 plotted against $n^{-1}\sqrt{km} * \text{SNR}$, where the number of measurements $m = 100$. Each plot in Figure 6.2 represents different relationship between k and n ; in the first plot, $k = \Theta(\log n)$, in the second $k = \Theta(\sqrt{n})$, while in the third plot $k = \Theta(n)$. The dashed vertical line denotes the threshold position for the scaled SNR at which the probability of success is larger than 0.95. We observe that irrespective of the problem size and the relationship

between n and k , Theorem 6.4.2 tightly characterizes the minimum SNR needed for successful identification.

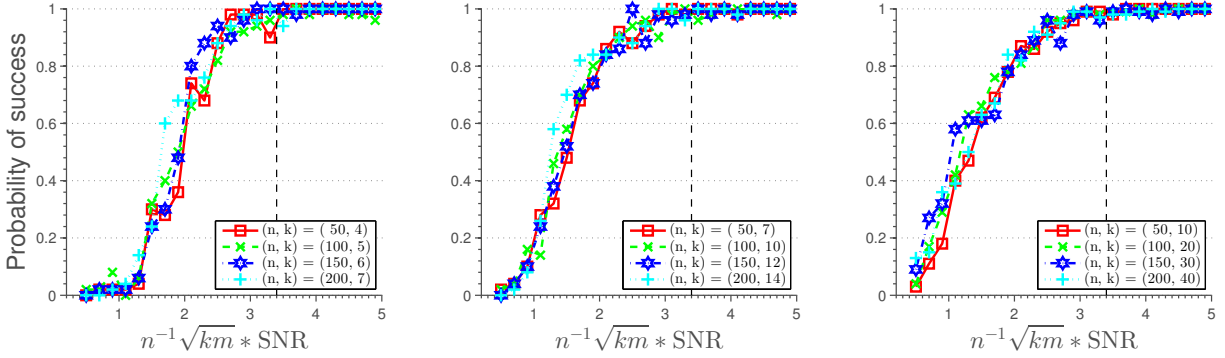


Figure 6.2: Probability of success with passive measurements (averaged over 100 simulation runs).

Experiment 2. Figure 6.3 shows the probability of successful localization of B^* using the procedure outlined in Section 5.2., with $m = 500$ adaptively chosen measurements, plotted against the scaled SNR. The SNR is scaled by $n^{-1}\sqrt{mk}^2$ in the first two plots where $k = \Theta(\log n)$ and $k = \Theta(\sqrt{n})$ respectively, while in the third plot the SNR is scaled by $\sqrt{mk}/\log k$ as $k = \Theta(n)$. The dashed vertical line denotes the threshold position for the scaled SNR at which the probability of success is larger than 0.95. We observe that Theorem 6.5.2 sharply characterizes the minimum SNR needed for successful identification.

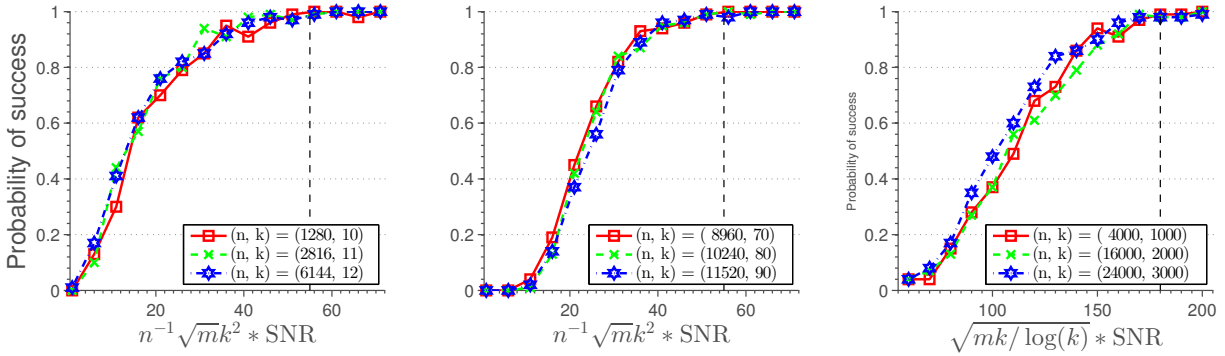


Figure 6.3: Probability of success with adaptively chosen measurements (averaged over 100 simulation runs).

6.7 Technical proofs

We now give detailed proofs of the results of this chapter. Throughout the proofs, we will denote c_1, c_2, \dots positive constants that may change their value from line to line.

6.7.1 Proof of Theorem 6.3.1

We lower bound the Bayes risk of any test T . Recall, the null and alternate hypothesis, defined in Eq. 6.3,

$$\begin{aligned} H_0: & A = 0_{n_1 \times n_2} \\ H_1: & A = (a_{ij}) \text{ with } a_{ij} = \mu \mathbb{I}_{\{(i,j) \in B\}}, B \in \mathcal{B}. \end{aligned}$$

We will consider a uniform prior over the alternatives π , and bound the average risk

$$R_\pi(T) = \mathbb{P}_0[T = 1] + \mathbb{E}_{A \sim \pi} \mathbb{P}_A[T = 0],$$

which provides a lower bound on the worst case risk of T .

Under the prior π , the hypothesis testing problem becomes to distinguish

$$\begin{aligned} H_0: & A = 0_{n_1 \times n_2} \\ H_1: & A = (a_{ij}) \text{ with } a_{ij} = \mathbb{E}_{B \sim \pi} \mu \mathbb{I}_{\{(i,j) \in B\}}. \end{aligned}$$

Both H_0 and H_1 are simple and the likelihood ratio test is optimal by the Neyman-Pearson lemma. The likelihood ratio is

$$L \equiv \frac{\mathbb{E}_\pi \mathbb{P}_A[(y_i, X_i)_{i \in [m]}]}{\mathbb{P}_0[(y_i, X_i)_{i \in [m]}]} = \frac{\mathbb{E}_\pi \prod_{i=1}^m \mathbb{P}_A[y_i | X_i]}{\prod_{i=1}^m \mathbb{P}_0[y_i | X_i]},$$

where the second equality follows by decomposing the probabilities by the chain rule and observing that $P_0[X_i | (y_j, X_j)_{j \in [i-1]}] = P_A[X_i | (y_j, X_j)_{j \in [i-1]}]$, since the sampling strategy (whether active or passive) is the same irrespective of the true hypothesis.

The likelihood ratio can be further simplified as

$$L = \mathbb{E}_\pi \exp \left(\sum_{i=1}^m \frac{2y_i \text{tr}(AX_i) - \text{tr}(AX_i)^2}{2\sigma^2} \right).$$

The average risk of the likelihood ratio test

$$R_\pi(T) = 1 - \frac{1}{2} \|\mathbb{E}_\pi \mathbb{P}_A - \mathbb{P}_0\|_{TV}$$

is determined by the total variation distance between the mixture of alternatives from the null.

By Pinsker's inequality [192],

$$\|\mathbb{E}_\pi \mathbb{P}_A - \mathbb{P}_0\|_{TV} \leq \sqrt{KL(\mathbb{P}_0, \mathbb{E}_\pi \mathbb{P}_A)/2}$$

and

$$\begin{aligned}
KL(\mathbb{P}_0, \mathbb{E}_\pi \mathbb{P}_A) &= -\mathbb{E}_0 \log L \\
&\leq -\mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 \frac{2y_i \text{tr}(AX_i) - \text{tr}(AX_i)^2}{2\sigma^2} \\
&= \mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 \frac{\text{tr}(AX_i)^2}{2\sigma^2} \\
&\leq \frac{m}{2\sigma^2} \sup_{\|X\|_F \leq 1} \mathbb{E}_\pi \text{tr}(AX_i) := \frac{m}{2\sigma^2} \|C\|_{op},
\end{aligned}$$

where the first inequality follows by applying the Jensen's inequality followed by Fubini's theorem, and the second inequality follows using the fact that $\|X_i\|_F^2 = 1$, where $C \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$.

To describe the entries of C , consider the invertible map τ from a linear index in $\{1, \dots, n_1 n_2\}$ to an entry of A . Now, $C_{ii} = \mu^2 \mathbb{E}_\pi P_A[A_{\tau(i)} = 1]$ and $C_{ij} = \mu^2 \mathbb{E}_\pi P_A[A_{\tau(i)} = 1, A_{\tau(j)} = 1]$.

To bound the operator norm of C we make two observations. Firstly, because of the contiguous structure of the activation pattern, in any row of C there are at most $k_1 k_2$ non-zero entries. Secondly, each non-zero entry in C is of magnitude at most $\mu^2 k_1 k_2 / (n_1 - k_1)(n_2 - k_2)$.

Now, note that

$$\|C\|_{op} \leq \max_j \sum_k |C_{jk}| \leq \mu^2 k_1^2 k_2^2 / (n_1 - k_1)(n_2 - k_2)$$

from which we obtain a bound on the KL divergence.

Now, this gives us that

$$R_\pi(T) \geq 1 - k_1 k_2 \mu \sqrt{\frac{m}{16(n_1 - k_1)(n_2 - k_2)}},$$

proving the lower bound on the minimax risk.

6.7.2 Proof of Theorem 6.3.2

Define $t = \frac{1}{\sqrt{m}} \sum_{i=1}^m y_i$. It is easy to see that under H_0 , $t \sim \mathcal{N}(0, \sigma^2)$ while under H_1 , $t \sim \mathcal{N}(\sqrt{\frac{m}{n_1 n_2}} k_1 k_2 \mu, \sigma^2)$. The theorem now follows from an application of standard Gaussian tail bounds in Eq. 5.25.

6.7.3 Proof of Theorem 6.5.1

The proof will proceed via two separate constructions. At a high level these constructions are intended to capture the difficulty of exactly and approximately localizing the activation block.

Construction 1 - approximate localization: Let us define three distributions: \mathbb{P}_0 corresponding to no bicluster, \mathbb{P}_1 which is a uniform mixture over the distributions induced by having the top-left corner of the bicluster in the left half of the matrix and \mathbb{P}_2 which is a uniform mixture over the distributions induced by having the top-left corner of the bicluster in the right half of the matrix.

We first upper bound the total variation between \mathbb{P}_1 and \mathbb{P}_2 . This results directly in a lower bound for the problem of distinguishing whether the top-left corner of the bicluster is in the left or right half of the matrix, which in turn is a lower bound for the localization of the bicluster.

Now notice that,

$$\begin{aligned} \|\mathbb{P}_1 - \mathbb{P}_2\|_{TV}^2 &\leq 2\|\mathbb{P}_0 - \mathbb{P}_1\|_{TV}^2 + 2\|\mathbb{P}_0 - \mathbb{P}_2\|_{TV}^2 \\ &\leq KL(\mathbb{P}_0, \mathbb{P}_1) + KL(\mathbb{P}_0, \mathbb{P}_2). \end{aligned}$$

Notice that $KL(\mathbb{P}_0, \mathbb{P}_1)$ is exactly the quantity we have to upper bound to produce a lower bound on the signal strength for detecting whether a block of activation is in the left half of the matrix or not. At least from a lower bound perspective this reduces the problem of localization to that of detection. We can now apply a slight modification of the proof of Theorem 6.3.1 to obtain that

$$KL(\mathbb{P}_0, \mathbb{P}_1) = KL(\mathbb{P}_0, \mathbb{P}_2) \leq \frac{m\mu^2 k_1^2 k_2^2}{(n_1 - k_1)(n_2/2 - k_2)}.$$

Noting that the minimax risk R for distinguishing \mathbb{P}_1 from \mathbb{P}_2

$$R = 1 - \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_2\|_{TV} \geq 1 - \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1 - k_1)(n_2/2 - k_2)}}.$$

Construction 2 - exact localization: Without loss of generality we assume $k_1 \leq k_2$. Consider, two distributions \mathbb{P}_1 and \mathbb{P}_2 , where \mathbb{P}_1 is induced by matrix A_1 when the activation block $B = B_1 = [1, \dots, k_1][1, \dots, k_2]$ and \mathbb{P}_2 is induced by matrix A_2 when the activation block $B = B_2 = [1, \dots, k_1][2, \dots, k_2 + 1]$.

Now, following the same argument as in the proof of Theorem 6.3.1, we have

$$\begin{aligned} KL(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m \left(-\frac{1}{2\sigma^2} [(y_i - \text{tr}(A_1 X_i))^2 - (y_i - \text{tr}(A_2 X_i))^2] \right) \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m [\text{tr}(A_2 X_i)^2 - \text{tr}(A_1 X_i)^2 + 2y_i \text{tr}(A_1 X_i) - 2y_i \text{tr}(A_2 X_i)] \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m \left(\underbrace{\text{tr}(A_2 X_i) - \text{tr}(A_1 X_i)}_{t_i} \right)^2 = \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m t_i^2. \end{aligned}$$

Now, with some abuse of notation,

$$\begin{aligned} t_i &= \mu \left(\sum_{j \in B_1 \setminus B_2} X_{ij} - \sum_{j \in B_2 \setminus B_1} X_{ij} \right) \\ &\leq \mu \left(\sum_{j \in B_1 \Delta B_2} |X_{ij}| \right). \end{aligned}$$

By using Cauchy-Schwarz we get

$$t_i^2 \leq 2\mu^2 k_1 \sum_{j \in B_1 \Delta B_2} X_{ij}^2 \leq 2\mu^2 k_1$$

since $\|X_i\|_F^2 = 1$.

This gives us that,

$$KL(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{mk_1\mu^2}{\sigma^2}.$$

Together with a similar construction for the case when $k_2 \leq k_1$ we get

$$KL(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{m \min(k_1, k_2)\mu^2}{\sigma^2}.$$

Once again noting (by Pinsker's theorem),

$$R \geq 1 - \sqrt{KL(\mathbb{P}_1, \mathbb{P}_2)/8} \geq 1 - \sqrt{\frac{m \min(k_1, k_2)\mu^2}{8\sigma^2}}.$$

Combining the approximate and exact localization bounds we get,

$$R \geq \max \left(1 - \sqrt{\frac{m \min(k_1, k_2)\mu^2}{8\sigma^2}}, 1 - \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1 - k_1)(n_2/2 - k_2)}} \right).$$

Thus, we get for any $0 < \alpha < 1$, $R \geq \alpha$ if

$$\min \left(\sqrt{\frac{m \min(k_1, k_2)\mu^2}{8\sigma^2}}, \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1 - k_1)(n_2/2 - k_2)}} \right) \leq 1 - \alpha.$$

6.7.4 Proof of Theorem 6.4.1

Without loss of generality we assume $k_1 \leq k_2$. Consider, two distributions \mathbb{P}_1 and \mathbb{P}_2 , where \mathbb{P}_1 is induced by matrix A_1 when the activation block $B = B_1 = [1, \dots, k_1] \times [1, \dots, k_2]$ and \mathbb{P}_2 is induced by matrix A_2 when the activation block $B = B_2 = [1, \dots, k_1] \times [2, \dots, k_2 + 1]$.

Following the proof of Theorem 6.5.1.

$$\begin{aligned} \text{KL}(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \log \frac{\mathbb{P}_1}{\mathbb{P}_2} \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m (\text{tr}(A_2 X_i) - \text{tr}(A_1 X_i))^2 \\ &= \frac{\mu^2 m k_1}{\sigma^2 n_1 n_2}, \end{aligned} \tag{6.9}$$

using the fact that X_i is a random Gaussian matrix with independent entries of variance $\frac{1}{n_1 n_2}$.

Now, note that the minimax risk

$$R \geq 1 - \sqrt{\text{KL}(\mathbb{P}_1, \mathbb{P}_2)/8}.$$

For the second part of the theorem, we consider $\mathbb{P}_2, \dots, \mathbb{P}_{t+1}$, where $t = (n_1 - k_1)(n_2 - k_2)$, each of which is induced by a B which does not overlap with B_1 .

The same calculation now gives

$$\text{KL}(\mathbb{P}_1, \mathbb{P}_j) \leq \frac{\mu^2 m k_1 k_2}{\sigma^2 n_1 n_2}. \tag{6.10}$$

Now, applying the multiple hypothesis version of Fano's inequality (see Theorem 2.5 in the book [192]) we conclude the proof.

6.7.5 Proof of Theorem 6.4.2

Let $z_{i,B} = \sum_{(a,b) \in B} X_{i,ab}$ and $\mathbf{z}_B = (z_{1,B}, \dots, z_{m,B})'$. With this, we can write the loss function defined in Eq. 6.5 as

$$f(B) := \min_{\hat{\mu}_B} \|\hat{\mu}_B \mathbf{z}_B - \mathbf{y}\|_2^2. \tag{6.11}$$

Let $\Delta(B) = f(B) - f(B^*)$ and observe that an error is made if $\Delta(B) < 0$ for $B \neq B^*$. Therefore,

$$\mathbb{P}[\text{error}] = \mathbb{P}[\cup_{B \in \mathcal{B} \setminus B^*} \{\Delta(B) < 0\}].$$

Under the conditions of the theorem, we will show that $\Delta(B) > 0$ for all $B \in \mathcal{B} \setminus B^*$ with large probability.

The following lemma shows that for any fixed B , the event $\{\Delta(B) < 0\}$ occurs with exponentially small probability.

Lemma 6.7.1. Fix any $B \in \mathcal{B} \setminus B^*$. Then

$$\mathbb{P}[\Delta(B) < 0] \leq \exp\left(-c_1 \frac{\mu^2 m |B^* \setminus B|}{\sigma^2 n_1 n_2}\right) + c_2 \exp(-c_3 m). \quad (6.12)$$

From the second term in Eq. 6.12, we obtain a lower bound on the sample size m . Using the union bound, it is sufficient that m satisfies

$$c_1(n_1 - k_1)(n_2 - k_2) \exp(-c_2 m) \leq \delta/2,$$

which gives us the lower bound as $m \geq C \log \max(n_1 - k_1, n_2 - k_2)$.

Define $N(l) = |\{B \in \mathcal{B} : |B \Delta B^*| = l\}|$ to be the number of elements in \mathcal{B} whose symmetric difference with B^* is equal to l . Note that $N(l) = \mathcal{O}(1)$ for any l . Using the union bound

$$\begin{aligned} & \mathbb{P}[\cup_{B \in \mathcal{B}} \{\Delta(B) < 0\}] \\ & \leq \sum_{B \in \mathcal{B}, |B \Delta B^*| = 2k_1 k_2} \exp\left(-c_1 \frac{\mu^2 k_1 k_2 m}{\sigma^2 n_1 n_2}\right) + \sum_{l < 2k_1 k_2} N(l) \exp\left(-c_1 \frac{\mu^2 l m}{\sigma^2 n_1 n_2}\right) \\ & \leq c_2(n_1 - k_1)(n_2 - k_2) \exp\left(-c_1 \frac{\mu^2 k_1 k_2 m}{\sigma^2 n_1 n_2}\right) + c_3 k_1 k_2 \exp\left(-c_1 \frac{\mu^2 \min(k_1, k_2) m}{\sigma^2 n_1 n_2}\right). \end{aligned} \quad (6.13)$$

Choosing

$$\mu = c_1 \sigma \sqrt{\frac{n_1 n_2}{m} \log(2/\delta) \max\left(\frac{\log \max(k_1, k_2)}{\min(k_1, k_2)}, \frac{\log \max(n_1 - k_1, n_2 - k_2)}{k_1 k_2}\right)}$$

each term in Eq. 6.13 will be smaller than $\delta/2$, with an appropriately chosen constant c_1 .

We finish the proof of the theorem, by proving Lemma 6.7.1.

Proof of Lemma 6.7.1. For any $B \in \mathcal{B}$, let

$$\begin{aligned} \hat{\mu}_B &= \operatorname{argmin}_{\hat{\mu}_B} \|\hat{\mu}_B \mathbf{z}_B - \mathbf{y}\|_2^2 \\ &= \|\mathbf{z}_B\|_2^{-2} \mathbf{z}'_B \mathbf{y}. \end{aligned}$$

Note that $\hat{\mu}_{B^*} = \mu + \|\mathbf{z}_{B^*}\|_2^{-2} \mathbf{z}'_{B^*} \boldsymbol{\epsilon}$.

Let

$$\begin{aligned} \mathbf{H}_B &= \|\mathbf{z}_B\|_2^{-2} \mathbf{z}_B \mathbf{z}'_B \\ \mathbf{H}_B^\perp &= \mathbf{I} - \|\mathbf{z}_B\|_2^{-2} \mathbf{z}_B \mathbf{z}'_B \end{aligned}$$

be the projection matrices and write

$$\begin{aligned} f(B^*) &= \|\mathbf{H}_{B^*}^\perp \boldsymbol{\epsilon}\|_2^2 \\ f(B) &= \|\mathbf{H}_B^\perp (\mathbf{z}_{B^*} \mu + \boldsymbol{\epsilon})\|_2^2 = \|\mathbf{H}_B^\perp \boldsymbol{\epsilon}\|_2^2 + \mu^2 \|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2 + 2\boldsymbol{\epsilon}' \mathbf{H}_B^\perp \mathbf{z}_{B^*} \mu. \end{aligned}$$

Now,

$$\Delta(B) = \underbrace{\|\mathbf{H}_B^\perp \epsilon\|_2^2 - \|\mathbf{H}_{B^*}^\perp \epsilon\|_2^2}_{T_1} + \underbrace{\mu^2 \|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2 + 2\epsilon' \mathbf{H}_B^\perp \mathbf{z}_{B^*} \mu}_{T_2}.$$

Conditional on \mathbf{X} , $\|\mathbf{H}_B^\perp \epsilon\|_2^2 \mid \mathbf{X} \sim \sigma^2 \chi_{m-1}^2$ and $\|\mathbf{H}_{B^*}^\perp \epsilon\|_2^2 \mid \mathbf{X} \sim \sigma^2 \chi_{m-1}^2$ [see Theorem 3.4.4 in 137]. Since the conditional distributions do not depend on \mathbf{X} , they are the same as the marginal distributions. Therefore, $T_1 \sim \sigma^2(V_1 - V_2)$ where $V_1, V_2 \sim \chi_{m-1}^2$.

$$\mathbb{P} \left[|T_1| \geq \frac{\sigma^2(m-1)\eta}{2} \right] \leq 2\mathbb{P} \left[|\chi_{m-1}^2 - m + 1| \geq \frac{(m-1)\eta}{4} \right] \leq 2 \exp \left(-\frac{3(m-1)\eta^2}{256} \right) \quad (6.14)$$

using Eq. 6.18, as long as $\eta \in [0, 2)$.

To analyze the term T_2 , we condition on \mathbf{X} , so that

$$T_2 \mid \mathbf{X} \sim \mathcal{N}(\tilde{\mu}, 4\sigma^2 \tilde{\mu})$$

where $\tilde{\mu} = \mu^2 \|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2$. This gives

$$\mathbb{P}[T_2 \leq \tilde{\mu}/2 \mid \mathbf{X}] = \mathbb{P}[\mathcal{N}(0, 1) \geq \sqrt{\tilde{\mu}}/(4\sigma) \mid \mathbf{X}].$$

Next, we show how to control $\|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2$. Writing $\mathbf{z}_{B^*} = \mathbf{z}_B - \mathbf{z}_{B \setminus B^*} + \mathbf{z}_{B^* \setminus B}$, simple algebra gives

$$\begin{aligned} & \|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2 \\ &= \|\mathbf{H}_B^\perp \mathbf{z}_{B^* \setminus B}\|_2^2 + \|\mathbf{H}_B^\perp \mathbf{z}_{B \setminus B^*}\|_2^2 - 2\mathbf{z}'_{B^* \setminus B} \mathbf{H}_B^\perp \mathbf{z}_{B \setminus B^*} \\ &= \|\mathbf{H}_B^\perp \mathbf{z}_{B^* \setminus B}\|_2^2 + \|\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B}\|_2^2 - \|\mathbf{z}_{B^* \setminus B}\|_2^2 - \frac{((\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B})' \mathbf{z}_B)^2 - (\mathbf{z}'_{B^* \setminus B} \mathbf{z}_B)^2}{\|\mathbf{z}_B\|_2^2} \\ &\geq \|\mathbf{H}_B^\perp \mathbf{z}_{B^* \setminus B}\|_2^2 + \|\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B}\|_2^2 - \|\mathbf{z}_{B^* \setminus B}\|_2^2 - \frac{((\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B})' \mathbf{z}_B)^2}{\|\mathbf{z}_B\|_2^2}. \end{aligned}$$

Define the event

$$\begin{aligned} \mathcal{E}(\eta) &= \left\{ \|\mathbf{H}_B^\perp \mathbf{z}_{B^* \setminus B}\|_2^2 \geq \frac{(1-\eta)(m-1)|B^* \setminus B|}{n_1 n_2} \right\} \cap \left\{ \|\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B}\|_2^2 \geq \frac{(1-\eta)2m|B^* \setminus B|}{n_1 n_2} \right\} \\ &\quad \cap \left\{ \|\mathbf{z}_{B^* \setminus B}\|_2^2 \leq \frac{(1+\eta)m|B^* \setminus B|}{n_1 n_2} \right\} \cap \left\{ \|\mathbf{z}_B\|_2^2 \geq \frac{(1-\eta)m|B|}{n_1 n_2} \right\} \\ &\quad \cap \left\{ |(\mathbf{z}_{B \setminus B^*} - \mathbf{z}_{B^* \setminus B})' \mathbf{z}_B| \leq \frac{(1+\eta)m|B^* \setminus B|}{n_1 n_2} \right\}, \end{aligned}$$

such that, using standard concentration results from Section 6.7.8,

$$\mathbb{P}[\mathcal{E}(\eta)^C] \leq c_1 \exp(-c_2 m \eta^2).$$

On the event $\mathcal{E}(\eta)$ we have that

$$\begin{aligned} \|\mathbf{H}_B^\perp \mathbf{z}_{B^*}\|_2^2 &\geq \frac{m|B^* \setminus B|}{n_1 n_2} \left[3(1-\eta) - (1+\eta) - \frac{(1+\eta)^2 |B^* \setminus B|}{1-\eta} \frac{1}{|B|} \right] - \frac{(1-\eta)|B^* \setminus B|}{n_1 n_2} \\ &\geq c_1 \frac{m|B^* \setminus B|}{n_1 n_2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{P}[T_2 \leq \tilde{\mu}/2 | \mathbf{X}] &\leq \mathbb{P} \left[\mathcal{N}(0, 1) \geq c_1 \frac{\mu}{\sigma} \sqrt{\frac{m|B^* \setminus B|}{n_1 n_2}} \right] + \mathbb{P}[\mathcal{E}^C] \\ &\leq \exp \left(-c_1 \frac{\mu^2 m |B^* \setminus B|}{\sigma^2 n_1 n_2} \right) + c_2 \exp(-c_3 m \eta^2). \end{aligned} \tag{6.15}$$

Combining Eq. 6.14 and Eq. 6.15 completes the proof. \square

6.7.6 Proof of Theorem 6.5.2

As with the lower bound the localization algorithm and analysis is naturally divided into two phases. An approximate localization phase and an exact localization one. We will analyze each of these in turn. To ease presentation we will assume n_1 is a dyadic multiple of $2k_1$ and n_2 a dyadic multiple of $2k_2$. Straightforward modifications are possible when this is not the case.

Approximate localization: The approximate localization phase proceeds by a modification of the compressive binary search (CBS) procedure of Malloy and Nowak [136] (see also Davenport and Arias-Castro [56]) on the matrix A .

We will run this modified CBS procedure four times on sets of blocks of the matrix A . The four sets are

$$\begin{aligned} \mathcal{D}_1 &\equiv \{B_{1,1} := [1, \dots, 2k_1] \times [1, \dots, 2k_2], B_{1,2} := [2k_1 + 1, \dots, 4k_1] \times [1, \dots, 2k_2] \\ &\quad \dots, B_{1, n_1 n_2 / 4k_1 k_2} := [n_1 - 2k_1, \dots, n_1] \times [n_2 - 2k_2, \dots, n_2]\} \\ \mathcal{D}_2 &\equiv \{B_{2,1} := [k_1, \dots, 3k_1] \times [k_2, \dots, 3k_2], B_{2,2} := [3k_1 + 1, \dots, 5k_1] \times [k_2, \dots, 3k_2] \\ &\quad \dots, B_{2, n_1 n_2 / 4k_1 k_2} := [n_1 - k_1, \dots, n_1, 1, \dots, k_1] \times [n_2 - k_2, \dots, n_2, 1, \dots, k_2]\} \\ \mathcal{D}_3 &\equiv \{B_{3,1} := [k_1, \dots, 3k_1] \times [1, \dots, 2k_2], B_{3,2} := [3k_1 + 1, \dots, 5k_1] \times [1, \dots, 2k_2] \\ &\quad \dots, B_{3, n_1 n_2 / 4k_1 k_2} := [n_1 - k_1, \dots, n_1, 1, \dots, k_1] \times [n_2 - 2k_2, \dots, n_2]\} \end{aligned}$$

and

$$\begin{aligned} \mathcal{D}_4 &\equiv \{B_{4,1} := [1, \dots, 2k_1] \times [k_2, \dots, 3k_2], B_{4,2} := [2k_1 + 1, \dots, 4k_1] \times [k_2, \dots, 3k_2] \\ &\quad \dots, B_{4, n_1 n_2 / 4k_1 k_2} := [n_1 - 2k_1, \dots, n_1] \times [n_2 - k_2, \dots, n_2, 1, \dots, k_2]\}. \end{aligned}$$

Notice that the entire block of activation is always *fully* contained in one of these blocks. The output of the CBS procedure when run on these four collections is four blocks - one from each

collection. We define an approximate localization *error* to be the event in which none of the blocks returned fully contains the block of activation.

Without loss of generality let us assume that the activation block is fully contained in some block from the first collection. Once we have fixed the collection of blocks the CBS procedure is invariant to reordering of the blocks, so without loss of generality we can consider the case when the activation block is contained in B_{11} .

The analysis proceeds exactly as in the paper [136]. We only outline the differences arising from having a block of activation as opposed to a single activation in a vector, and refer the reader to Malloy and Nowak [136] for the details.

The binary search procedure on the first collection of blocks proceeds for

$$s_0 \equiv \log \left(\frac{n_1 n_2}{4k_1 k_2} \right)$$

rounds. Now, we can bound the probability of error of the procedure by a union bound as

$$\mathbb{P}_e \leq \sum_{s=1}^{s_0} P[w^s < 0]$$

where

$$w^s \sim \mathcal{N} \left(\frac{m_s 2^{(s-1)/2} k_1 k_2 \mu}{\sqrt{n_1 n_2}}, m_s \sigma^2 \right).$$

Recall, the allocation scheme: for $m \geq 2s_0$, $m_s \equiv \lfloor (m - s_0) s 2^{-s-1} \rfloor + 1$ and observe that $\sum_{s=1}^{s_0} m_s \leq m$.

Now, using the Gaussian tail bound

$$P[N(0, 1) > t] \leq \frac{1}{2} \exp(-t^2/2)$$

we see that

$$\mathbb{P}_e \leq \frac{1}{2} \sum_{s=1}^{s_0} \exp \left(-\frac{m_s 2^s k_1^2 k_2^2 \mu^2}{4n_1 n_2 \sigma^2} \right).$$

Now, observe that $m_s \geq (m - s_0) s 2^{-s-1}$ and $m \geq 2s_0$, so $m_s \geq m s 2^{-s-2}$.

It is now straightforward to verify that if

$$\mu \geq \sqrt{\frac{16\sigma^2 n_1 n_2}{m k_1^2 k_2^2} \log \left(\frac{1}{2\delta} + 1 \right)}$$

we have $\mathbb{P}_e \leq \delta$. We apply this procedure 4 times (once on each collection).

Let us revisit what we have shown so far: if μ is large enough then one of the four runs of the CBS procedure will return a block of size $(2k_1 \times 2k_2)$ which fully contains the block of activation, with probability at least $1 - 4\delta$.

Exact localization: We collect all the rows and columns returned by the 4 runs of the CBS procedure. In the $1 - 4\delta$ probability event described above, we have a block of at most $(8k_1 \times 8k_2)$ which contains the full block of activation (for simplicity we disregard the fact that we know that the block is actually in one of two $(4k_1 \times 4k_2)$ blocks, i.e. we assume the worst case that none of the returned blocks overlap in their rows or columns and we explore the off-diagonal blocks).

Let us first identify the active columns. First, notice that exactly one of the following columns: $\{1, k_2 + 1, 2k_2 + 1, \dots, 7k_2 + 1\}$ must be active.

Let us devote $8m$ measurements to identifying the active column amongst these. The procedure is straightforward: measure each column m times, and pick the one that has the largest total signal.

It is easy to show that the active column results in a draw from $\mathcal{N}(\sqrt{\frac{k_1}{8}}\mu m, m\sigma^2)$ and the non-active columns result in draws from $\mathcal{N}(0, m\sigma^2)$.

Using the same Gaussian tail bound as before it is easy to show that if

$$\mu \geq \sqrt{\frac{64\sigma^2}{k_1 m} \log(4/\delta)}$$

we successfully find the active column with probability at least $1 - \delta$.

So far, we have identified an active column and localized the columns of the activation block to one of $2k_2$ columns. We will use m more measurements to find the remaining active columns. Rather, than test each of the $2k_2$ columns we will do a binary search. This will require us to test at most $t \equiv 2\lceil \log k_2 \rceil \leq 3 \log k_2$ columns, and we will devote $m/(3 \log k_2)$ measurements to each column. We will need to threshold these measurements at

$$\sqrt{\log\left(\frac{3 \log k_2}{\delta}\right) \frac{2m\sigma^2}{3 \log k_2}}$$

and declare a row as active if its average is larger than this.

It is easy to show that this binary search procedure successfully finds all active columns with probability at least $1 - \delta$ if

$$\mu \geq \sqrt{\frac{32\sigma^2 \log k_2}{mk_1} \log\left(\frac{3 \log k_2}{\delta}\right)}.$$

We repeat this procedure to identify the active rows.

Putting everything together: Total number of measurements used:

1. Four rounds of CBS: $4m$
2. Identifying first active column and first active row: $16m$

3. Identifying remaining active rows and columns: $2m$

This is a total of $22m$ measurements. Each of these steps fails with a probability at most δ , for a total of 8δ .

Now, re-adjusting constants we obtain, if

$$\mu \geq \max \left(\sqrt{\frac{352\sigma^2 n_1 n_2}{m k_1^2 k_2^2} \log \left(\frac{4}{\delta} + 1 \right)}, \sqrt{\frac{1408\sigma^2 \log \max(k_1, k_2)}{m \min(k_1, k_2)} \log \left(\frac{24 \log \max(k_1, k_2)}{\delta} \right)} \right)$$

then we successfully localize the matrix with probability at least $1 - \delta$.

Stated more succinctly we require

$$\mu \geq \tilde{O} \left(\max \left(\sqrt{\frac{\sigma^2 n_1 n_2}{m k_1^2 k_2^2}}, \sqrt{\frac{\sigma^2}{\min(k_1, k_2) m}} \right) \right).$$

This matches the lower bound up to $\log k$ factors.

6.7.7 Proof of Eq. 6.7 and Eq. 6.8

Proof of Eq. 6.7 follows the same line as the proof of Theorem 6.4.2. We have

$$\begin{aligned} \mathbb{P}[\text{error}] &= \mathbb{P}[\cup_{B \in \mathcal{B} \setminus B^*} \{\Delta(B) < 0\}] \\ &\leq \sum_{i=0}^{k_1} \binom{k_1}{i} \binom{n_1 - k_1}{k_1 - i} \sum_{j=0}^{k_2} \binom{k_2}{j} \binom{n_2 - k_2}{k_2 - j} \exp \left(-c_1 \frac{(\mu^*)^2 m (k_1 k_2 - ij)}{\sigma^2 n_1 n_2} \right) \\ &\quad + \sum_{i=0}^{k_1} \binom{k_1}{i} \binom{n_1 - k_1}{k_1 - i} \sum_{j=0}^{k_2} \binom{k_2}{j} \binom{n_2 - k_2}{k_2 - j} c_2 \exp(-c_3 m). \end{aligned}$$

The argument given in the proof of Theorem 2 in the paper [111] gives us Eq. 6.7 if $m \geq C \log \max \left(\binom{n_1}{k_1}, \binom{n_2}{k_2} \right)$. Proof of Eq. 6.8 follows the proof of Theorem 1 in the paper [111] with the appropriate KL divergences derived in Eq. 6.9 and Eq. 6.10.

6.7.8 Some concentration bounds

We now state some useful results on tail bounds of various random quantities used throughout this chapter.

Tail bounds for Chi-squared variables

Throughout the chapter we use one of the following tail bounds for central χ^2 random variables. These are well known and proofs can be found in the original papers.

Lemma 6.7.2 ([121]). Let $X \sim \chi_d^2$. For all $x \geq 0$,

$$\mathbb{P}[X - d \geq 2\sqrt{dx} + 2x] \leq \exp(-x) \quad (6.16)$$

$$\mathbb{P}[X - d \leq -2\sqrt{dx}] \leq \exp(-x). \quad (6.17)$$

Lemma 6.7.3 ([103]). Let $X \sim \chi_d^2$, then

$$\mathbb{P}[|d^{-1}X - 1| \geq x] \leq \exp(-\frac{3}{16}dx^2), \quad x \in [0, \frac{1}{2}]. \quad (6.18)$$

The following result provide a tail bound for non-central χ^2 random variable with non-centrality parameter ν .

Lemma 6.7.4 ([31]). Let $X \sim \chi_d^2(\nu)$, then for all $x > 0$

$$\mathbb{P}[X \geq (d + \nu) + 2\sqrt{(d + 2\nu)x} + 2x] \leq \exp(-x) \quad (6.19)$$

$$\mathbb{P}[X \leq (d + \nu) - 2\sqrt{(d + 2\nu)x}] \leq \exp(-x). \quad (6.20)$$

Using the above results, we have a tail bound for sum of product-normal random variables.

Lemma 6.7.5. Let $Z = (Z_a, Z_b) \sim \mathcal{N}_2(0, 0, \sigma_{aa}, \sigma_{bb}, \sigma_{ab})$ be a bivariate Normal random variable and let $(z_{ia}, z_{ib}) \stackrel{iid}{\sim} Z, i = 1, \dots, n$. Then for all $t \in [0, \nu_{ab}/2)$

$$\mathbb{P}\left[\left|n^{-1} \sum_i z_{ia}z_{ib} - \sigma_{ab}\right| \geq t\right] \leq 4 \exp\left(-\frac{3nt^2}{16\nu_{ab}^2}\right), \quad (6.21)$$

where $\nu_{ab} = \max\{(1 - \rho_{ab})\sqrt{\sigma_{aa}\sigma_{bb}}, (1 + \rho_{ab})\sqrt{\sigma_{aa}\sigma_{bb}}\}$.

Proof. Let $z'_{ia} = z_{ia}/\sqrt{\sigma_{aa}}$. Then using Eq. 6.18

$$\begin{aligned} & \mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n z_{ia}z_{ib} - \sigma_{ab}\right| \geq t\right] \\ &= \mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n z'_{ia}z'_{ib} - \rho_{ab}\right| \geq \frac{t}{\sqrt{\sigma_{aa}\sigma_{bb}}}\right] \\ &= \mathbb{P}\left[\left|\sum_{i=1}^n ((z'_{ia} + z'_{ib})^2 - 2(1 + \rho_{ab})) - ((z'_{ia} - z'_{ib})^2 - 2(1 - \rho_{ab}))\right| \geq \frac{4nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}\right] \\ &\leq \mathbb{P}\left[\left|\sum_{i=1}^n ((z'_{ia} + z'_{ib})^2 - 2(1 + \rho_{ab}))\right| \geq \frac{2nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}\right] \\ &\quad + \mathbb{P}\left[\left|\sum_{i=1}^n ((z'_{ia} - z'_{ib})^2 - 2(1 - \rho_{ab}))\right| \geq \frac{2nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}\right] \\ &\leq 2\mathbb{P}\left[\left|\chi_n^2 - n\right| \geq \frac{nt}{\nu_{ab}}\right] \leq 4 \exp(-\frac{3nt^2}{16\nu_{ab}^2}), \end{aligned}$$

where $\nu_{ab} = \max\{(1 - \rho_{ab})\sqrt{\Sigma_{aa}\Sigma_{bb}}, (1 + \rho_{ab})\sqrt{\Sigma_{aa}\Sigma_{bb}}\}$ and $t \in [0, \nu_{ab}/2)$. \square

Corollary 6.7.6. *Let Z_1 and Z_2 be two independent standard Normal random variables and let $X_i \stackrel{iid}{\sim} Z_1 Z_2$, $i = 1 \dots n$. Then for $t \in [0, 1/2)$*

$$\mathbb{P}\left[\left|n^{-1} \sum_{i \in [n]} X_i\right| > t\right] \leq 4 \exp\left(-\frac{3nt^2}{16}\right). \quad (6.22)$$

Chapter 7

Minimax Rates for Homology Inference

In this chapter we begin our foray into statistical topics in topological data analysis. This chapter considers minimax bounds for estimating the homology of a sub-manifold while the next chapter considers the problem of learning the cluster tree of a density supported on or near a sub-manifold.

Often, high dimensional data lie close to a low-dimensional sub-manifold and it is of interest to understand the geometry of these sub-manifolds. The homology groups of a manifold are important topological invariants that provide an algebraic summary of the manifold. These groups contain rich topological information, for instance, about the connected components, holes, tunnels and sometimes the dimension of the manifold. In this chapter, we consider the statistical problem of estimating the homology of a manifold from noisy samples under several different noise models. We derive upper and lower bounds on the minimax risk for this problem. Our upper bounds are based on estimators which are constructed from a union of balls of appropriate radius around carefully selected points. In each case we establish complementary lower bounds using Le Cam's lemma. Finally, we establish tight asymptotic lower bounds by a direct analysis of the likelihood ratio test on a pair of suitably chosen hypotheses.

7.1 Introduction

Let M be a d -dimensional manifold embedded in \mathbb{R}^D where $d \leq D$. The *homology groups* $\mathcal{H}(M)$ of M [92] are an algebraic summary of the properties of M . The homology groups of a manifold describe its topological features such as its connected components, holes, tunnels, etc.

In machine learning, there is much focus on clustering. However, the clusters are only the zeroth order homology and hence only scratch the surface of the topological information in a dataset. Extracting information beyond clustering is known as topological data analysis. It is worth emphasizing that the homology groups are topological invariants of a manifold that can be *efficiently* computed [58, 59]. Examples of applications of homology inference have been growing rapidly

in the last few years. Homology inference has found application in medical imaging and neuroscience [50, 173], sensor networks [60, 171], landmark-based shape data analyses [80], proteomics [166], microarray analysis [61] and cellular biology [110]. The books by [68, 151, 211] contain various case studies in applications in fields ranging from computational biology to geophysics.

In this chapter we study the problem of estimating the homology of a manifold M from a noisy sample Y_1, \dots, Y_n . Specifically, we bound the minimax risk

$$R_n \equiv \inf_{\hat{\mathcal{H}}} \sup_{Q \in \mathcal{Q}} Q^n \left(\hat{\mathcal{H}} \neq \mathcal{H}(M) \right) \quad (7.1)$$

where the infimum is over *all* estimators $\hat{\mathcal{H}}$ of the homology of M and the supremum is over appropriately defined classes of distributions \mathcal{Q} for Y . Note that $0 \leq R_n \leq 1$ with $R_n = 1$ meaning that the problem is hopeless. Bounding the minimax risk is equivalent to bounding the *sample complexity* of the best possible estimator, defined by

$$n(\epsilon) = \min \{ n : R_n \leq \epsilon \}$$

where $0 < \epsilon < 1$.

7.1.1 Related Work

Other work on statistical homology includes that of Chazal et al. [47] who show under certain conditions the homology estimate of a manifold from a sample is stable under noise perturbation that is small in a Wasserstein sense. Kahle [106] studies the homology of random geometric graphs and proves many threshold and central limit theorems for their homology. Adler et al. [3] study the homology induced by the level sets of certain Gaussian random fields. There is also a large literature on manifold denoising that focuses on aspects of the manifold not related to homology; see for instance the paper [95] and references therein.

Our upper bounds mainly generalize those in the work of Niyogi, Smale and Weinberger (henceforth NSW) [146, 147]. They establish a general result showing that when *all* the samples are dense in a thin region surrounding the manifold, a union of appropriately sized balls around the samples can be used to construct an accurate estimate of the homology with high probability. Under a variety of different noise models we will show that even when *all* the samples are not close to the manifold it is possible to “clean” the samples (essentially removing those in regions of low-density) and be left with samples which are dense in a thin region around the manifold.

In the case of additive noise with general noise distributions however, we cannot expect too many samples to fall close to the manifold. We will show that when the noise distribution is known one can use a statistical deconvolution procedure to obtain a “deconvolved measure” concentrated around the manifold from which we can in turn draw a small number of samples and apply the cleaning procedure described above to them. Deconvolution has been extensively studied in the

	Noise Model				
	Noiseless	Clutter	Tubular	Additive Gaussian	General additive (τ fixed)
Upper Bound	NSW	This chapter	NSW	This chapter	This chapter
Lower Bound	This chapter	This chapter	This chapter	This chapter	This chapter

Table 7.1: Summary of our contributions

statistical literature (see the paper [70] and references therein). Most related to our application is the work of Koltchinskii [113] who uses deconvolution to estimate the dimension and cluster tree of a distribution supported on a submanifold. We defer a detailed comparison to Section 7.5.4 after the necessary preliminaries have been introduced.

To the best of our knowledge these are the first lower and upper minimax bounds for the problem of inferring the homology of a manifold. There are a few existing results on upper bounds. A summary of previous results and the results in this chapter are in Table 1.

Outline. In Section 7.2 we describe the statistical model. In Section 7.3 we give a brief description of homology. In Section 7.4 we give an overview of our techniques. We derive the minimax rates for the four noise settings in Section 7.5. Technical proofs are contained in Section 7.7.

7.2 Statistical Model

We assume that the sample $\{Y_1, \dots, Y_n\} \subset \mathbb{R}^D$ constitutes a set of “noisy” observations of an unknown d -dimensional manifold M , with $d < D$, whose homology we seek to estimate. The distribution of the sample depends on the properties of the manifold M as well as on the type of sampling noise, which we describe below by formulating various statistical models for sampling data from manifolds.

Notation. We let $B_r^k(x)$ denote a k -dimensional ball of radius r centered at x . When $k = D$, we write $B_r(x)$ instead of $B_r^D(x)$. For any set M and any $\sigma > 0$ define $\text{tube}_\sigma(M) = \bigcup_{x \in M} B_\sigma(x)$. Let v_k denote the volume of the k -dimensional unit ball. Finally, for clarity we let $c_1, c_2, \dots, C_1, C_2, \dots$ denote various positive constants whose value can be different in different expressions. The constants will be specified in the corresponding proofs.

Manifold Assumptions. We assume that the unknown manifold M is a d -dimensional smooth compact Riemannian manifold without boundary embedded in the compact set $\mathcal{X} = [0, 1]^D$. We further assume that the volume of the manifold is bounded from above by a constant which can depend on the dimensions d, D , i.e. we assume $\text{vol}(M) \leq C_{D,d}$. We will also make the further assumption that $D > d$. The main regularity condition we impose on M is that its *condition number* be not too small. The *condition number* $\kappa(M)$ (see the paper of Niyogi et al. [146]) is the largest number τ such that the open normal bundle about M of radius r is imbedded in \mathbb{R}^D

for every $r < \tau$. For $\tau > 0$ let

$$\mathcal{M} \equiv \mathcal{M}(\tau) = \left\{ M : \kappa(M) \geq \tau \right\}$$

denote the set of all such manifolds with condition number no smaller than τ . A manifold with large condition number does not come too close to being self-intersecting. We consider the collection

$$\mathcal{P} \equiv \mathcal{P}(\mathcal{M}) \equiv \mathcal{P}(\mathcal{M}, a)$$

of all probability distributions supported over manifolds M in \mathcal{M} having densities f with respect to the volume form on M uniformly bounded from below by a constant $a > 0$, i.e. $0 < a \leq f(x) < \infty$ for all $x \in M$. For expositional clarity we treat a as a *fixed* constant although our upper and lower bounds match in their dependence on a .

The Noise Models. We consider four noise models and, for each of them, we specify a class \mathcal{Q} of probability distributions for the sample.

1. *Noiseless.* We observe data $Y_1, \dots, Y_n \sim P$ where $P \in \mathcal{P}$. In this case,

$$\mathcal{Q} = \mathcal{Q}(\tau) = \mathcal{P}.$$

2. *Clutter Noise.* We observe data Y_1, \dots, Y_n from the mixture

$$Q = (1 - \pi)U + \pi P$$

where, $P \in \mathcal{P}$, $0 \leq \pi \leq 1$ and U is a uniform distribution on \mathcal{X} . The points drawn from U are called background clutter. Then

$$\mathcal{Q} = \mathcal{Q}(\pi, \tau) = \left\{ Q = (1 - \pi)U + \pi P : P \in \mathcal{P} \right\}.$$

Notice that $\pi = 1$ reduces to the noiseless case.

3. *Tubular Noise.* We observe $Y_1, \dots, Y_n \sim Q_{M,\sigma}$ where $Q_{M,\sigma}$ is uniform on a tube of size σ around M . In this case

$$\mathcal{Q} = \mathcal{Q}(\sigma, \tau) = \left\{ Q_{M,\sigma} : M \in \mathcal{M} \right\}.$$

4. *Additive Noise.* The data are of the form $Y_i = X_i + \epsilon_i$, where $X_1, \dots, X_n \sim P$, for some $P \in \mathcal{P}$, and $\epsilon_1, \dots, \epsilon_n$ are a sample from a noise distribution Φ . Note that $Q = P \star \Phi$, that is, Q is the convolution of P and Φ . We consider two cases:

- (a) Φ is a D -dimensional Gaussian with mean $(0, \dots, 0)$ and covariance $\sigma^2 I$, with $\sigma \ll \tau$. Define

$$\mathcal{Q} = \mathcal{Q}(\sigma, \tau) = \left\{ Q = P \star \Phi : P \in \mathcal{P} \right\}.$$

- (b) Φ is any known noise distribution whose Fourier transform is bounded away from 0 but with the added restriction that we only consider manifolds with τ being a fixed constant. Then

$$\mathcal{Q} = \mathcal{Q}(\Phi) = \left\{ Q = P \star \Phi : P \in \mathcal{P}_\tau \right\}.$$

where \mathcal{P}_τ is the subset of \mathcal{P} comprised of distributions supported on manifolds M with condition number at least as large as the *fixed* value τ .

The noise model used in the paper [147] is to take the noise at any point to be only along the normal fibres; this seems unnatural and we will not consider that model here.

In almost all of the distribution classes considered we allow for τ to vanish as n gets bigger, which is equivalent to letting the difficulty of the statistical problem increase with the sample size. To this end, we will also analyze the quantity

$$\tau_n \equiv \tau_n(\epsilon) = \inf\{\tau : R_n \leq \epsilon\}$$

which corresponds to the smallest condition number that permits accurate estimation. We call this the *resolution*.

7.3 Homology

Often in this chapter we will use phrases like “the homology of the union of balls around samples”. In this section we explain this usage and discuss briefly *simplicial homology* (see Hatcher (2001) for a detailed treatment) and its computation.

The homology \mathcal{H} of a space M is a collection of groups that correspond to topological features of M . We will consider the case when M is a compact Riemannian manifold. In what follows, it might help the reader’s intuition to imagine that we are starting with a dense sample of points U on the manifold and building a collection of simplices from these points. The union of balls $\bigcup_{y \in U} B_\epsilon(y)$ gives a geometric approximation to the underlying manifold. This is however a continuous (infinite) collection of points. To make computation tractable we need to be able to reduce the computation of homology from a continuous space to its discretization. The Čech complex (a particular *simplicial complex*, see Figure 7.3) which is described below gives a discrete representation of the union of balls. A classic result in topology called the Nerve Theorem [92] states that the homology of $\bigcup_{y \in U} B_\epsilon(y)$ is identical to the homology of the corresponding Čech complex.

We now describe a simplicial complex and its homology. A *simplicial complex* is a hereditary set system \mathcal{K} over a vertex set V , i.e. $\sigma \subset \sigma' \in \mathcal{K}$ implies that $\sigma \in \mathcal{K}$. The *dimension* of a simplex σ is $|\sigma| - 1$; singletons are 0-simplices or vertices, pairs in \mathcal{K} are 1-simplices or edges, triples are 2-simplices or triangles, etc. A p -chain is a formal sum of p -simplices. The coefficients are taken in $\mathbb{Z}/2\mathbb{Z}$, the integers mod 2.¹ Thus, chains may be viewed as subsets of simplices and addition

¹In general, homology may be defined over any ring, but we stick with \mathbb{Z}_2 for ease of exposition and computation.

(mod 2) as symmetric difference of sets. Addition of chains forms an abelian group called the *chain group* C_p with 0 denoting the empty chain.

A p -simplex $\sigma = \{v_0, \dots, v_p\}$ has $p + 1$ simplices of dimension $p - 1$ on its boundary, denoted $\sigma_i = \sigma \setminus \{v_i\}$. The *boundary* of a simplex is

$$\partial_p \sigma = \sum_{i=0}^p \sigma_i.$$

The *boundary operator* $\partial_p : C_p \rightarrow C_{p-1}$ is the natural extension of the boundary of a simplex to the boundary of a chain: $\partial_p c = \sum_{\sigma \in c} \partial_p \sigma$.

The kernel and image of the boundary operator are two important subgroups of the chain group: *the cycle group*:

$$Z_p = \ker \partial_p = \{z \in C_p : \partial_p z = 0\},$$

and *the boundary group*:

$$B_p = \text{im } \partial_p = \{\partial_{p+1} c : c \in C_{p+1}\}.$$

The *cycles* Z_p are those chains that have boundary 0. The *boundary cycles* B_p are those p -chains that are the boundary of some $p + 1$ -chain. It is easy to check that $\partial_{p-1} \partial_p c = 0$ and thus $B_p \subset Z_p \subset C_p$. See Figure 7.1.

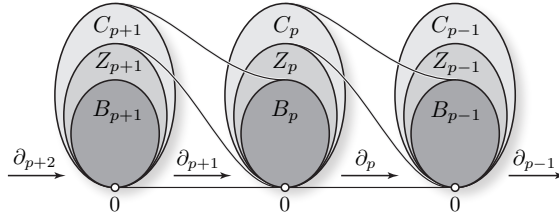


Figure 7.1: Relationship between chains C_p , cycles $Z_p = \ker \partial_p$ and boundaries $B_p = \text{im } \partial_{p+1}$. The chains C_p are just collections of simplices. The chains in Z_p are the cycles. The cycles in B_p are the cycles that happen to be boundaries of chains in C_{p+1} .

Two cycles $z_1, z_2 \in Z_p$ are *homologous* if $z_1 - z_2 \in B_p$, i.e. their difference is the boundary of a $p + 1$ -chain. The p th homology group H_p is defined as the quotient group Z_p/B_p . That is, the homology group is a collection of equivalence classes of cycles. The first homology group H_0 corresponds to connected components (clusters). The next homology group H_1 corresponds to non-bounding cycles (or loops). Higher order homology groups correspond to equivalence classes of higher dimensional cycles.² The homology of \mathcal{K} is the collection \mathcal{H} of all its homology groups.

The Čech complex is a specific simplicial complex defined as follows. Fix some $\epsilon > 0$ and a set of points $S \subset \mathbb{R}^D$. The Čech complex consists of all simplices σ such that $\bigcap_{x \in \sigma} B_\epsilon(x) \neq \emptyset$ where $B_\epsilon(x)$ is a ball of radius ϵ centered at x . See Figure 7.3.

² Intuitively, boundary cycles are “filled in” cycles and two cycles are homologous if one cycle can be deformed into the other cycle.



Figure 7.2: The sum of two 1-cycles is another 1-cycle. Here the cycles are homologous because their sum (in \mathbb{Z}_2) is the boundary of a 2-chain of triangles.

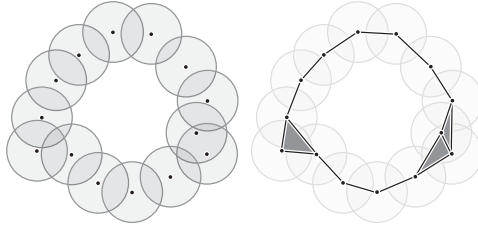


Figure 7.3: A union of balls and its corresponding Čech complex.

Since the coefficient ring is a field, the computations may be completely described by linear algebra. The groups C_p , Z_p , B_p , and H_p are vector spaces and the boundary operators are linear maps. It is possible to efficiently compute the homology groups of a simplicial complex in time polynomial in the size of the complex. The algorithm only involves row reduction on the matrix representations of ∂_p .

7.4 Preliminaries

In this section we briefly describe some of the main techniques we use to obtain upper and lower bounds.

7.4.1 Techniques for lower bounds

The *total variation distance* between two measures P and Q is defined by $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ where the supremum is over all measurable sets. It can be shown that $\text{TV}(P, Q) = P(G) - Q(G) = 1 - \int \min(P, Q)$ where $G = \{y : p(y) \geq q(y)\}$ and p and q are the densities of P and Q with respect to any measure μ that dominates both P and Q .

We shall make repeated use of Le Cam's lemma which we now state (see, e.g., Lemma 1 in the paper [209]).

Lemma 7.4.1 (Le Cam). *Let \mathcal{Q} be a set of distributions. Let $\theta(Q)$ take values in a metric space with metric ρ . Let $Q_1, Q_2 \in \mathcal{Q}$ be any pair of distributions in \mathcal{Q} . Let Y_1, \dots, Y_n be drawn iid*

from some $Q \in \mathcal{Q}$ and denote the corresponding product measure by Q^n . Then

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \left[\rho(\hat{\theta}, \theta(Q)) \right] \geq \frac{1}{8} \rho(\theta(Q_1), \theta(Q_2)) (1 - \text{TV}(Q_1, Q_2))^{2n}$$

where the infimum is over all estimators.

Le Cam's lemma makes precise the intuition that if there are *distinct* members of the class \mathcal{Q} for which the data generating distributions are close then the statistical problem is hard given a small sample.

When we apply Le Cam's lemma in this chapter, Q_1 and Q_2 will be associated with two different manifolds M_1 and M_2 . We will take $\theta(Q)$ to be the homology of the manifold and $\rho(\theta(Q_1), \theta(Q_2)) = 1$ if the homologies are the different and $\rho(\theta(Q_1), \theta(Q_2)) = 0$ if the homologies are the same. The subtlety of establishing *tight* lower bounds boils down to the task of finding a set of distributions in the class \mathcal{Q} for which the homology of the underlying submanifolds are distinct but whose empirical distributions are hard to distinguish from a small number of samples.

We will use two representative manifolds M_1 and M_2 in the application of LeCam's lemma which we describe here. See Figure 7.4. The manifold M_1 is a pair of $1 - \tau$ d -balls (shown in blue) embedded 2τ apart in \mathbb{R}^D joined smoothly at their ends (shown in red). The manifold M_2 is a pair of d -annuli (shown in blue) embedded 2τ apart with outer radius $1 - \tau$ and inner radius 4τ , smoothly joined at both the inner and outer ends (shown in red). It is clear from the construction that both these manifolds are d -dimensional compact, have no boundary and have condition number τ . It is also the case that $\mathcal{H}(M_1) \neq \mathcal{H}(M_2)$.

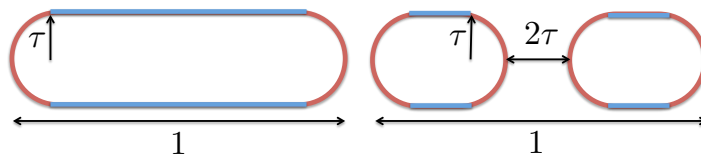


Figure 7.4: The two manifolds M_1 and M_2 , with $d = 1$, $D = 2$

If there exist two manifolds M_1 and M_2 with corresponding distributions Q_1 and Q_2 in \mathcal{Q} such that (i) $\mathcal{H}(M_1) \neq \mathcal{H}(M_2)$ and (ii) $Q_1 = Q_2$ then we say that the model \mathcal{Q} is *non-identifiable*. In this case, recovering the homology is impossible and we write $R_n = 1$ and $n(\epsilon) = \infty$.

When directly analyzing the likelihood ratio test we will lower bound the minimax risk by considering a related *testing* problem.

Before describing the hypotheses we describe the null and alternate manifolds. The null manifold M_0 is a collection of m , d -spheres of radius τ , denoted S_1, \dots, S_m , with centers on one face of the unit hypercube in $d + 1$ dimensions (M_0 is embedded in a space of dimension D which is of

dimension at least $d + 1$), with spacing between adjacent centers $= 4\tau$. It is easy to see that

$$m = O\left(\frac{1}{(4\tau)^d}\right)$$

because the manifold must be completely in $[0, 1]^D$, and that the manifold has condition number at least $1/\tau$. We will use

$$m = \Theta\left(\frac{1}{(4\tau)^d}\right)$$

for the lower bound we construct in this chapter. Let P_0 denote the uniform distribution on M_0 .

The alternate manifolds are a collection $\{M_{1i} : i \in \{1, \dots, m\}\}$, where M_{1i} is M_0 with S_i removed. Let π denote the uniform distribution on $\{1, \dots, m\}$, and P_{1i} denote the uniform distribution on M_{1i} .

We need to ensure that the density f is lower bounded by a constant. Note that the total d -dimensional volume of M_0 is $v_d\tau^d m$, and so

$$f(x) \geq \frac{1}{v_d\tau^d m}$$

where v_d is the volume of the d -dimensional unit ball. This is $\Omega(1)$ as desired. A similar argument works for M_{1i} .

Consider the following testing problem:

$$\begin{aligned} H_0 &: \mathbf{X} \sim P_0 \\ H_1 &: \mathbf{X} \sim P_{1i} \text{ with } i \sim \pi. \end{aligned}$$

A test T , is a measurable function of \mathbf{X} , in particular $T : \mathbf{X} \rightarrow \{0, 1\}$, and its risk is defined as

$$R_n^T := \mathbb{P}_{H_0}(T(\mathbf{X}) = 1) + \mathbb{P}_{H_1}(T(\mathbf{X}) = 0).$$

The relationship between testing and estimation is standard [126]. In our case it is easy to see that the estimation minimax risk we are interested in satisfies,

$$R_n^T \leq 2R_n$$

and so it suffices to lower bound R_n^T to obtain a lower bound on R_n . This relation is a straightforward consequence of the fact that $\mathcal{H}(M_0) \neq \mathcal{H}(M_{1i})$ for every i (since they have different number of connected components), and so any estimator can be used in the testing problem described.

The optimal test for the hypothesis testing problem described is the likelihood ratio test,

$$T(\mathbf{X}) = 0 \text{ if and only if } L(\mathbf{X}) \leq 1$$

where

$$L(\mathbf{X}) = \frac{L_1(\mathbf{X})}{L_0(\mathbf{X})}$$

where $L_1(\mathbf{X})$ and $L_0(\mathbf{X})$ are likelihoods of the data under the alternate and null respectively. To prove a strong lower bound all we need to do is to show that this test has a large risk. We do this in Section 7.6.

7.4.2 Techniques for upper bounds

To establish an upper bound we need to construct an estimator that achieves the upper bound. In the noiseless and tubular noise cases the samples are in a thin region around the manifold and our estimator is constructed from a union of balls (of a carefully chosen radius) around the sample points.

In the case of clutter noise and additive Gaussian noise samples are concentrated around the manifold but a few samples may be quite far away from the manifold. In these cases our upper bounds are obtained by analyzing the performance of the Algorithm 5 (CLEAN) with a carefully specified threshold and radius, which is used to remove points in regions of low density far away from the manifold. Our estimator is then constructed from a union of balls around the remaining points. In the case of additive noise with general known distribution the samples are

Algorithm 5 CLEAN

- IN: $(X_i)_{i=1}^n$, threshold t , radius r
 - 1. Construct a graph G_r with nodes $\{X_i\}_{i=1}^n$. Include edge (X_i, X_j) , if $\|X_i - X_j\| \leq r$.
 - 2. Mark all vertices with degree $d_i \leq (n-1)t$.
 - OUT: All unmarked vertices
-

not expected to be concentrated around the manifold. We will first use deconvolution to estimate a deconvolved measure \widehat{P}_n which we will show is densely concentrated in a thin region around the manifold. We will then draw samples from this measure, clean them and construct a union of balls of appropriate radius around the remaining samples, and show that this set has the right homology with high probability.

We now briefly review statistical deconvolution. We refer the interested reader to the work of Fan [70] for more details and to the paper [113] for an application related to ours. The procedure is similar to kernel density estimation with a kernel modified to account for the additive noise. For symmetric noise distributions Φ , we consider two kernels \mathcal{K} and Ψ such that $\mathcal{K} \star \Phi = \Psi$, where \star denotes convolution. The deconvolution estimator is

$$\widehat{P}_n(A) = 1/n \sum_{i=1}^n \mathcal{K}(Y_i - A).$$

It is easy to verify that $E\widehat{P}_n = P \star \Psi$ similar to regular kernel density estimation with the kernel Ψ . In the noiseless case we can even take $\mathcal{K} = \Psi = \delta_0$ (a Dirac at 0) and get back the

empirical distribution of the sample. More generally, we will be interested in Ψ that satisfies $\Psi\{x : |x| \geq \epsilon\} \leq \gamma$ for ϵ and γ that we will later specify.

In each of the above cases our final estimator is constructed from a union of balls around appropriate points, and our theorems will show that these have the correct homology with high probability. To compute the homology one would construct the corresponding Čech complex and compute its “boundary matrices” (as described in Section 7.3). Recovering the homology from these matrices consists of linear algebraic manipulation. There are several fast algorithms to compute the homology (either exactly [58] or approximately [59]) of the Čech complexes from large point sets in high dimensions.

7.5 Minimax Rates

We now derive the minimax rates for homology estimation under the four noise models described in section 7.2. We will first give minimax rates for all noise models with lower bounds obtained using Le Cam’s lemma. The sample complexity lower bounds differ from the corresponding upper bounds by a logarithmic factor. In Section 7.6 we will show a tighter analysis and derive an asymptotic lower bound for the noiseless case that eliminates this discrepancy. We will not consider the extension to the other noise models in this thesis but they are straightforward.

There are three quantities of interest: the minimax risk R_n , the resolution τ_n and the sample complexity $n(\epsilon)$. We write $R_n \asymp a_n$ (similarly for $\tau_n \asymp a_n$) if there are positive constants c and C such that $c \leq R_n/a_n \leq C$ for all large n . Similarly, we write $n(\epsilon) \asymp a(\epsilon)$ if there are positive constants c and C such that $c \leq n(\epsilon)/a(\epsilon) \leq C$ for all small ϵ . Our analysis will show that the rates (as a function of n) are typically polynomial for the resolution and exponential for the risk. We will often match upper and lower bounds on sample complexity and resolution only up to logarithmic factors, and correspondingly those on the risk up to polynomial factors. In this case we will use the notation $R_n \asymp^* a_n$, $\tau_n \asymp^* a_n$, and $n(\epsilon) \asymp^* a(\epsilon)$.

It is worth emphasizing at this point that despite the fact that we use two specific manifolds in the application of Le Cam’s lemma, the resulting lower bound holds for *all* manifolds in \mathcal{M} and *all* distributions in \mathcal{Q} . Le Cam’s lemma allows one to get a lower bound that holds for *any* estimator by using *two* carefully chosen distributions in \mathcal{Q} . The upper bounds are from specific estimators and they establish an upper bound on the number of samples to estimate the homology of any manifold in our class.

7.5.1 Noiseless Case

Theorem 7.5.1. *For all $\tau \leq \tau_0(a, d)$, in the noiseless case the minimax rate,*

$$R_n \asymp^* e^{-n\tau^d},$$

where $\tau_0(a, d)$ is a constant which depends on a and d . Also,

$$n(\epsilon) \asymp^* \tau^{-d} \log(1/\epsilon)$$

and

$$\tau_n \asymp^* \left(\frac{1}{n} \log(1/\epsilon) \right)^{1/d}.$$

We provide proof sketches for the lower and upper bounds on R_n separately.

Lower Bound: Proof Sketch

To obtain a lower bound on the minimax risk over the class $\mathcal{Q}(\tau)$ we will consider the two carefully chosen manifolds M_1 and M_2 described earlier.

We further need to specify the density on each of the manifolds, and we choose two densities from \mathcal{P} so that the data distributions are as similar as possible while respecting the constraint $f(x) \geq a$. The construction is described in more detail in the Section 7.7.1, but for now it suffices to notice that the two densities can be constructed to differ only on the sets $W_1 = M_1 \setminus M_2$ and $W_2 = M_2 \setminus M_1$ and can be made as low as a on one of these sets. A straightforward calculation shows that

$$\text{TV}(p_1, p_2) \leq a \max(\text{vol}(W_1), \text{vol}(W_2)) \leq C_d a \tau^d$$

where the constant C_d depends on d . Now, we apply Le Cam's lemma to obtain that

$$R_n \geq \frac{1}{8} (1 - C_d a \tau^d)^n \geq \frac{1}{8} e^{-2C_d n a \tau^d}$$

for all $\tau \leq \tau_0(a, d)$. $\tau_0(a, d)$ is a constant depending on a and d . The lower bound of Theorem 7.5.1 follows.

Upper Bound: Proof Sketch

In the noiseless case the samples are densely concentrated around the manifold and our estimator is constructed from a union of balls of radius $\tau/2$ around the sample points. The upper bound on the minimax risk follows from a straightforward modification of the results of Niyogi et al. [146]. For completeness, we reproduce an adaptation of their main homology inference theorem (Theorem 3.1) here.

Lemma 7.5.2. [NSW] Let $0 < \epsilon < \tau$ and let $U = \bigcup_{i=1}^n B_\epsilon(X_i)$. Let $\hat{\mathcal{H}} = \mathcal{H}(U)$. Let

$$\zeta_1 = \frac{\text{vol}(M)}{a \cos^d \theta_1 \text{vol}(B_{\epsilon/4}^d)},$$

$$\zeta_2 = \frac{\text{vol}(M)}{a \cos^d \theta_2 \text{vol}(B_{\epsilon/8}^d)},$$

$$\theta_1 = \sin^{-1} \frac{\epsilon}{8\tau} \quad \text{and} \quad \theta_2 = \sin^{-1} \frac{\epsilon}{16\tau}.$$

Then for all

$$n > \zeta_1 \left(\log(\zeta_2) + \log\left(\frac{1}{\delta}\right) \right),$$

$$\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta.$$

By assumption $\text{vol}(M) \leq C_{D,d}$ for some constant $C_{D,d}$ depending on d and D . To obtain a sample complexity bound we simply choose $\epsilon = \tau/2$ and this gives us

$$n(\epsilon) \leq C_1/(a\tau^d)(C_2 \log(1/(a\tau^d)) + \log(1/\epsilon))$$

which matches the lower bound upto the factor of $\log(1/\tau)$. Further calculation (see Section 7.7.1) then shows that as desired

$$R_n \leq \frac{C_1}{\tau^d} \exp(-C_2 n a \tau^d)$$

for appropriate constants C_1, C_2 , and

$$\tau_n \leq C \left(\frac{\log n \log(1/\epsilon)}{an} \right)^{1/d}.$$

This establishes Theorem 7.5.1.

7.5.2 Clutter Noise

Theorem 7.5.3. For all $\tau \leq \tau_0(a, d)$, in the clutter noise case,

$$R_n \asymp^* e^{-n\pi\tau^d},$$

where $\tau_0(a, d)$ is a constant which depends on a and d . Also,

$$n(\epsilon) \asymp^* \left(\frac{1}{\pi\tau^d} \log(1/\epsilon) \right)$$

and

$$\tau_n \asymp^* \left(\frac{1}{n\pi} \log(1/\epsilon) \right)^{1/d}.$$

Lower Bound: Proof Sketch

The lower bound for the class $\mathcal{Q}(\pi, \tau)$ follows via the same construction as in the noiseless case. In the calculation of the total variation distance (see Section 7.7.1) we have instead

$$\text{TV}(q_1, q_2) \leq \pi a \max(\text{vol}(W_1), \text{vol}(W_2)) \leq C_d \pi a \tau^d$$

where C_d depends on d . As before the lower bound follows from the application of Le Cam's lemma.

Upper Bound: Proof Sketch

As a preliminary step we clean the data samples to eliminate points that are far away from, while retaining those close to, the manifold. Our analysis shows that Algorithm 5 will achieve this, with high probability for a carefully chosen threshold and radius. We then show that taking a union of balls of the appropriate radius around the remaining points will give us the correct homology, with high probability. We give an outline here and defer details to Section 7.7.1.

1. We define two regions

$$A = \text{tube}_r(M) \text{ and } B = \mathbb{R}^D \setminus \text{tube}_{2r}(M)$$

where

$$r < \frac{(\sqrt{9} - \sqrt{8})\tau}{2}.$$

2. We then invoke Algorithm CLEAN on the data with threshold

$$t = \left(\frac{v_D s^D (1 - \pi)}{\text{vol}(\text{Box})} + \frac{\pi a v_d r^d \cos^d \theta}{2} \right)$$

and radius $2r$. Let I be the set of vertices returned.

3. Through careful analysis we show that with high probability I contains *all* the vertices from the region A and *none* of the points in region B .
4. We further show that the retained points form a thin dense cover of the manifold M , i.e. $\left\{ M \subset \bigcup_{i \in I} B_{2r}(X_i) \right\}$.
5. Using a straightforward corollary of Lemma 7.5.2 we show that this thin dense cover can be used to recover the homology of M with high probability.

Formally, in Section 7.7.1 we prove the following lemma,

Lemma 7.5.4. *If $n > \max(N_1, N_2)$, and $r < (\sqrt{9} - \sqrt{8})\frac{\tau}{2}$ where $N_1 = 4\kappa \log(\kappa)$*

$$\text{with } \kappa = \max \left(1 + \frac{200}{3\zeta} \log \left(\frac{2}{\delta} \right), 4 \right)$$

$$\text{and } N_2 = \frac{1}{\zeta} \left(\log \left(\frac{\text{vol}(M)}{\cos^d(\theta) v_d r^d} \right) + \log \left(\frac{2}{\delta} \right) \right)$$

where $\zeta = \pi a v_d r^d \cos^d(\theta)$ and $\theta = \sin^{-1}(r/2\tau)$, then after cleaning the points $\{X_i : i \in I\}$ are all in $\text{tube}_{2r}(M)$ and are $2r$ dense in M . Let $U = \bigcup_{i \in I} B_w(X_i)$ with $w = r + \frac{\tau}{2}$ and let $\widehat{\mathcal{H}} = \mathcal{H}(U)$. We have that $\widehat{\mathcal{H}} = \mathcal{H}(M)$ with probability at least $1 - \delta$.

Taking $r = (\sqrt{9} - \sqrt{8})\tau/4$, we obtain the sample complexity bound,

$$n(\epsilon) \leq \frac{C_1}{\pi \tau^d} \left(\log \frac{C_2}{\tau^d} + \log(C_3/\epsilon) \right).$$

Given this sample complexity upper bound, the upper bounds on minimax risk and resolution follow identical arguments to the noiseless case (Section 7.7.1).

7.5.3 Tubular Noise

Under this noise model we get samples uniformly from a tubular region of width σ around the manifold. This model highlights an important phenomenon in high-dimensions. Although, we receive samples *uniformly* from a full D dimensional shape these samples concentrate tightly around a d dimensional manifold. We show that with some care we can still reconstruct the homology at a rate independent of D .

Theorem 7.5.5. *Under the tubular noise model we establish the following cases.*

1. If $\sigma \geq 2\tau$ then the model is non-identifiable and hence, $R_n = 1$ and $n(\epsilon) > \infty$.
2. If $\sigma \leq C_0\tau$, with C_0 small and $\tau \leq \tau_0(a, d)$, then

$$R_n \asymp^* e^{-n\tau^d},$$

where $\tau_0(a, d)$ is a constant which depends on a and d . Also,

$$n(\epsilon) \asymp^* 1/\tau^d$$

and

$$\tau_n \asymp^* \left(\frac{1}{n} \log(1/\epsilon) \right)^{1/d}.$$

Remark 7.5.6. *The case when σ is very close to τ is significantly more involved since it involves the exact calculation of the volume of the tubular region and establishing tight upper and lower bounds here is an open problem we are attempting to address in current work.*

Lower bound: Proof Sketch

1. When $d < D$ and $\sigma \geq 2\tau$ the two manifolds M_1 and M_2 that we have considered thus far are still identifiable because even when $\sigma \geq \tau$ M_2 has a “dimple” along the co-dimensions that M_1 does not. To show that the class \mathcal{Q} is still not identifiable we require a different construction. Consider the manifolds M_1 and M_2 with two points placed above and below the manifold at a distance 2τ above their centers along each of the co-dimensions. Denote these new manifolds M'_1 and M'_2 . It is clear that $\mathcal{H}(M'_1) \neq \mathcal{H}(M'_2)$, however $\mathcal{Q}'_1 = \mathcal{Q}'_2$ since the extra points hide the “dimple” and the two manifolds cannot be distinguished.
2. When $d < D$, and $\sigma \leq C_0\tau$ we return to our old constructions of M_1 and M_2 . There is however an important difference in that the two manifolds differ on full D -dimensional sets, and one might suspect that $TV(q_1, q_2) = O(\tau^D)$ or perhaps $O(\sigma^{D-d}\tau^d)$. As we show in Section 7.7.1 however, $TV(q_1, q_2)$ is still $O(\tau^d)$, and we recover an identical lower bound to the noiseless case.

Upper bound: Proof Sketch

We are interested in case when $\sigma \leq C_0\tau$ (in particular $\sigma < \tau/24$ will suffice). Our proof will involve two main steps which we sketch here.

1. We first show that if we consider balls of sufficiently large radius ϵ (compared to σ) then the probability mass in these balls is $O(\epsilon^d)$. This is a manifestation of the phenomenon alluded to earlier: inside large enough balls the mass is concentrated around the lower dimensional manifold. Precisely, define $k_\epsilon = \inf_{p \in M} Q(B_\epsilon(p))$. In Lemma 7.7.4, we show that, if $\epsilon \gg \sigma$ is large, k_ϵ is of order $\Omega(\epsilon^d)$.
2. There is however a disadvantage to considering balls that are too large. The homology of the union of balls around the samples may no longer have the right homology. Using tools from NSW, we show that we can balance these two considerations for manifolds with high condition number, i.e. provided $\sigma < \tau/24$, we can choose balls that are both large relative to σ and whose union still has the correct homology.

We will prove the following main lemma in Section 7.7.

Lemma 7.5.7. *Let N_ϵ be the ϵ -covering number of the submanifold M . Let $U = \bigcup_{i=1}^n B_{\epsilon+\tau/2}(X_i)$. Let $\widehat{\mathcal{H}} = \mathcal{H}(U)$. Then if*

$$n > \frac{1}{k_\epsilon} (\log(N_\epsilon) + \log(1/\delta)),$$

$\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$ as long as $\sigma \leq \epsilon/2$ and $\epsilon < \frac{(\sqrt{9}-\sqrt{8})\tau}{2}$.

Notice, that we require $\sigma < \frac{(\sqrt{9}-\sqrt{8})\tau}{4}$ which is satisfied if $\sigma < \tau/24$ (for instance). To obtain the upper bound set $\epsilon = 2\sigma$, and observe that $N_\epsilon = O(1/\epsilon^d) = O(1/\tau^d)$ and $k_\epsilon = O(\epsilon^d) = O(\tau^d)$. This gives us that if

$$n \geq \frac{C_1}{\tau^d} \left(\log \left(\frac{C_2}{\tau^d} \right) + \log \left(\frac{1}{\delta} \right) \right)$$

we recover the right homology with probability at least $1 - \delta$. The upper bound on minimax risk and resolution follows from similar arguments to those made previously.

7.5.4 Additive Noise

For additive noise we consider two cases. In the first case, we derive the minimax rates for additive *Gaussian* noise under the somewhat restrictive assumption that $C\sqrt{D}\sigma < \tau$. This problem is related of the problem of separating mixtures of Gaussians (which corresponds to the case where the manifold is a collection of points and 2τ is the distance between the closest pair). In this case have the following theorem.

Theorem 7.5.8. *For all $\tau \leq \tau_0(a, d)$ and $8\sqrt{D}\sigma < \tau$,*

$$R_n \asymp^* e^{-n\tau^d},$$

where $\tau_0(a, d)$ is a constant which depends on a and d . Also,

$$n(\epsilon) \asymp^* (1/\tau^d) \log(1/\epsilon)$$

and

$$\tau_n \asymp^* ((1/n) \log(1/\epsilon))^{1/d}.$$

As in the clutter noise case we need to first clean the data and then take a union of balls around the points which survive. We analyze this procedure in Section 7.7.

Deconvolution

Here we consider more general *known* noise distributions but work over the class of distributions $\mathcal{Q}(\Phi)$ over manifolds with τ fixed. We first use deconvolution to estimate a deconvolved measure \widehat{P}_n which is concentrated around the manifold. We then draw samples from this measure, clean them and construct a union of balls H around these samples, and show that H has the right homology with high probability. The class of noise distributions we will consider satisfy the following assumption on its density.

Assumption 4. Denote $\rho(R) = \inf_{|t|_\infty \leq R} |\Phi^*(t)|$, where $R > 0$, $|t|_\infty = \max_{1 \leq j \leq m} |t_j|$ and $\Phi^*(t)$ is the Fourier transform of the symmetric noise density Φ . We assume $\rho(R) > 0$.

This is a standard assumption in the literature on deconvolution (see [70, 113]), since as described deconvolution requires us to divide by the Fourier transform of the noise which needs to be bounded away from 0 for the procedure to be well behaved. The assumption is satisfied by a variety of noise distributions including Gaussian noise. Our main result says that for this broad class of noise distributions the deconvolution procedure described above will achieve an optimal rate of convergence.

Theorem 7.5.9. In the additive noise case with τ fixed for Φ satisfying Assumption 4. $R_n \asymp e^{-n}$. Hence, $n(\epsilon) \asymp \log(1/\epsilon)$.

Lower Bound: Proof Sketch To obtain the lower bound one can consider the same construction from the previous subsection with additive Gaussian noise. If τ is taken to be fixed we obtain the desired bound.

Upper Bound: Proof Sketch Our proof of the upper bound follows similar lines to that of Koltchinskii [113]. We deviate in two significant aspects. Koltchinskii only assumes an upper bound on the density, which he shows is sufficient to estimate weak geometric characteristics like the dimension of the manifold. To show that we can accurately reconstruct its homology we require both an upper and lower bound and our methods are quite different. Koltchinskii uses an epsilon net of the *entire* compact set containing the manifold critically in his construction and his procedure is thus not implementable/practical. Our algorithm instead draws a small number of samples from the deconvolved measure and uses those to estimate the homology resulting in a practical procedure. We prove the following upper bound.

Lemma 7.5.10. Given n samples from $\mathcal{Q}(\Phi)$ with Φ satisfying Assumption 4, there exist $C_1, C_2, c_1 > 0$ such that $P(\mathcal{H}(H) \neq \mathcal{H}(M)) \leq C_1 e^{-c_1 n}$, where H is a union of balls of radius $\frac{5\epsilon + \tau}{2}$ centered around $m \geq C_2 n$ samples drawn from the deconvolved measure \widehat{P}_n with a kernel Ψ with parameters γ, ϵ (specified in the proof). The samples are cleaned using the deconvolved measure by considering balls of radius 4ϵ at a threshold 2γ .

Remark 7.5.11. The cleaning procedure we use here is different from the Algorithm CLEAN. We remove points around which a ball of appropriate radius has low probability mass under the deconvolved measure. This is equivalent to using the deconvolved measure in place of the k -NN

density estimate implicitly constructed by the CLEAN procedure.

Simple calculations show that this lemma together with the lower bound give the exponential minimax rate described in Theorem 7.5.9.

7.6 Tight lower bound

In the previous sections we used Le Cam's lemma to establish the lower bound

$$R_n = \Omega(\exp(-n\tau^d))$$

for $d \geq 1$ and $D > d$.

In this section we use a different construction based on the direct analysis of the likelihood ratio test to show that

$$R_n = \Omega\left(\frac{1}{\tau^d} \exp(-n\tau^d)\right),$$

as $n \rightarrow \infty$ thus establishing rate optimal asymptotic minimax bounds for the problem. The techniques we use here extend in a straightforward way to the noisy settings. Although, we do not consider the extension here non-asymptotic bounds are also straightforward.

7.6.1 Coupon collector lower bound

We begin with a theorem from the book [141].

Lemma 7.6.1 (Theorem 3.8 of [141]). *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then for any constant $c \in \mathbb{R}$, and $m = n \log n - cn$,*

$$\lim_{n \rightarrow \infty} \mathbb{P}(X > m) = 1 - \exp(-\exp(c)).$$

7.6.2 Main result

Theorem 7.6.2. *For any constant $\delta < 1$, we have*

$$R_n \geq \Omega\left(\min\left(\frac{1}{\tau^d} \exp(-n\tau^d), \delta\right)\right)$$

as $n \rightarrow \infty$.

Proof. Notice that since

$$m = \Theta\left(\frac{1}{(4\tau)^d}\right)$$

the theorem is implied by the statement that

$$n = m \log m + m \log \left(\frac{1}{\delta} \right) \implies R_n \geq c\delta$$

for some constant c . We will focus on proving this claim.

Let us consider the case when samples are drawn according to P_0 . From Lemma 7.6.1 we have that if

$$n = m \log m + m \log \left(\frac{1}{\delta} \right)$$

then the probability with which we do not see a point in each of the m spheres is

$$1 - \exp(-\exp(-\log 1/\delta)) \geq c\delta$$

since $\delta < 1$, for some constant c . It is easy to see that if we do not see a point in each of the m spheres then

$$L(\mathbf{X}) \geq \frac{1}{m} \frac{1}{(1 - 1/m)^n} := T_{m,n}.$$

When $n = m \log m + m \log \left(\frac{1}{\delta} \right)$,

$$T_{m,n} \rightarrow \frac{1}{\delta} > 1$$

so asymptotically the likelihood ratio test always rejects the null.

From this we can see the probability of a Type I error $\rightarrow c\delta$, and $R_n^T \geq c\delta$, which gives

$$R_n \geq \frac{c}{2}\delta$$

as desired. □

7.7 Technical proofs

7.7.1 Key technical lemmas from NSW

We will need two technical lemmas, which follow from the paper [146].

Lemma 7.7.1 (Ball volume lemma, Lemma 5.3 in [146]). *Let $p \in M$. Now consider $A = M \cap B_\epsilon(p)$. Then*

$$\text{vol}(A) \geq (\cos(\theta))^d \text{vol}(B_\epsilon^d(p))$$

where $B_\epsilon^d(p)$ is the a d -dimensional ball in the tangent space at p , $\theta = \sin^{-1} \frac{\epsilon}{2r}$.

Next, consider a collection of balls $\{B_r(p_i)\}_{i=1,\dots,n}$ centered around points p_i on the manifold and such that $M \subset \cup_{i=1}^n B_r(p_i)$.

Lemma 7.7.2 (Sampling lemma, Lemma 5.1 in [146]). Let $A_i = B_r(p_i)$ be a collection of sets such that $\cup_{i=1}^l A_i$ forms a minimal cover of M . If $Q(A_i) \geq \alpha$, and

$$n > \frac{1}{\alpha} \left(\log l + \log \left(\frac{2}{\delta} \right) \right)$$

then w.p. at least $1 - \delta/2$, each A_i contains at least one sample point, and $M \subset \cup_{i=1}^n B_{2r}(x_i)$. Further we have that $l \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d r^d}$.

Proofs for the noiseless case

Lower bound Here we describe the densities on the two manifolds M_1 and M_2 . There are two sets of interest to us: $W_1 = M_1 \setminus M_2$ which corresponds to the two ‘‘holes’’ of radius 4τ in the annulus, and $W_2 = M_2 \setminus M_1$ which corresponds to the d -dimensional piece added to smoothly join the inner pieces of the two annuli in M_2 .

By construction, $\text{vol}(W_1) = 2v_d(4\tau)^d$ where v_d is the volume of the unit d -ball. $\text{vol}(W_2)$ is somewhat tricky to calculate exactly due to the curvature of W_2 but it is easy to see that $\text{vol}(W_2)$ is also $O(\tau^d)$ with the constant depending on d .

One of the densities is constructed in the following way, on the set of larger volume (between W_1 and W_2) we set $f(x) = a$, and evenly distribute the rest of the mass over the remaining portion of the manifold (we are guaranteed that the mass on the rest of the manifold is at least a since otherwise the constraint $f(x) \geq a$ can never be satisfied).

The other density is constructed to be equal (to the first density) outside the set on which the two manifolds differ. The remaining mass is spread evenly on the set where they do differ. We are again guaranteed that $f(x) \geq a$ by construction.

Let us now calculate the TV between these two densities. This is just the integral of the difference of the densities over the set where one of the densities is larger. Since the two densities are equal outside $W_1 \cup W_2$ and disjoint over $W_1 \cup W_2$ it is clear that

$$TV(p_1, p_2) = a \max(\text{vol}(W_1), \text{vol}(W_2)) \leq O(a\tau^d)$$

with the constant depending on d . The lower bound follows from the calculations described previously.

Upper bound The NSW lemma tells us that for

$$n > \zeta_1 \left(\log(\zeta_2) + \log \left(\frac{1}{\delta} \right) \right),$$

with

$$\zeta_1 = \frac{\text{vol}(M)}{a \cos^d \theta_1 \text{vol}(B_{\epsilon/4}^d)},$$

$$\zeta_2 = \frac{\text{vol}(M)}{\cos^d \theta_2 \text{vol}(B_{\epsilon/8}^d)},$$

$$\theta_1 = \sin^{-1} \frac{\epsilon}{8\tau} \quad \text{and} \quad \theta_2 = \sin^{-1} \frac{\epsilon}{16\tau},$$

we have $\mathbb{P}(\widehat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$.

By assumption, we have $\text{vol}(M) \leq C$. We further take $\epsilon = \tau/2$. It is clear that in ζ_1 and ζ_2 all terms except the ball volumes are constant. This gives us that $\zeta_1 = C_1/(a\tau^d)$ and $\zeta_2 = C_2/(a\tau^d)$.

Now, the NSW lemma can be restated as if

$$n = \frac{C_1}{\tau^d} \left(\log \frac{C_2}{\tau^d} + \log(1/\delta) \right)$$

we recover the homology with probability at least $1 - \delta$. Notice that this means that the minimax risk $\leq \delta$.

A straightforward rearrangement of this gives us

$$R_n \leq C_2/(a\tau^d) \exp(-na\tau^d/C_1)$$

for appropriate C_1, C_2 . To bound the resolution we rewrite this as

$$R_n \leq \exp \left(-\frac{na\tau^d}{C_1} + \log \left(\frac{C_2}{a\tau^d} \right) \right).$$

One can verify that if

$$\tau^d \leq C \frac{\log n \log(1/\epsilon)}{n}$$

for an appropriately large C , we have $R_n \leq \epsilon$ as desired.

Proofs for the clutter noise case

Lower bound This is a straightforward extension of the noiseless case. The densities are constructed in an identical manner. The contribution to the densities from the clutter noise is identical in each case. As in the analysis for the noiseless case we bound the total variation distance between the two densities. We have an additional factor of π which is the mixture weight of the component corresponding to the density on the manifold.

$$TV(q_1, q_2) = \pi a \max(\text{vol}(W_1), \text{vol}(W_2)) \leq C_d \pi a \tau^d.$$

Given this bound the calculations are identical to those in the noiseless case.

Upper bound As a preliminary step we will need to clean the data to eliminate points that are far away from the manifold. Our analysis will show that Algorithm 5 will achieve this, with high

probability. We will then show that taking a union of balls of the appropriate radius around the remaining points will give us the correct homology, with high probability.

Let $a = \inf_{x \in M} f(x)$, which is strictly positive by assumption. Define,

$$A = \text{tube}_r(M) \text{ and } B = \mathbb{R}^D - \text{tube}_{2r}(M)$$

where

$$r < \frac{(\sqrt{9} - \sqrt{8})\tau}{2}.$$

Following Niyogi et al. [147], we define

$$\alpha_s = \inf_{t \in A} Q(B_s(t)) \text{ and } \beta_s = \sup_{t \in B} Q(B_s(t))$$

where $s = 2r$. Then

$$\alpha_s \geq \frac{v_D s^D (1 - \pi)}{\text{vol}(\text{Box})} + \pi a v_d r^d \cos^d \theta = \alpha$$

and

$$\beta_s \leq \frac{v_D s^D (1 - \pi)}{\text{vol}(\text{Box})} = \beta$$

where $\theta = \sin^{-1}(\frac{r}{2r})$. The second term of the bound on α_s follows in two steps: first observe that for any point x in A , $B_s(x) \supseteq B_r(t)$ where t is the closest point on M to x . Now, we use Lemma 7.7.1 to bound $Q(B_r(t))$.

We will now invoke Algorithm CLEAN on the data with threshold

$$t = \left(\frac{v_D s^D (1 - \pi)}{\text{vol}(\text{Box})} + \frac{\pi a v_d r^d \cos^d \theta}{2} \right)$$

and radius $2r$. Let I be the set of vertices returned.

Define the events

$$\mathcal{E}_1 = \left\{ \{X_i : i \in I\} \supseteq \{X_i \in A\} \text{ and } \{X_i : i \in I^c\} \supseteq \{X_i \in B\} \right\}$$

and

$$\mathcal{E}_2 = \left\{ M \subset \bigcup_{i \in I} B_{2r}(X_i) \right\}.$$

We will show that \mathcal{E}_1 and \mathcal{E}_2 both hold with high probability.

For \mathcal{E}_1 to hold, we need β to be not too close to α , in particular $\beta < \alpha/2$ will suffice. This happens with probability 1, for τ small if $d < D$. By Lemma 7.7.8, \mathcal{E}_1 happens with probability at least $1 - \delta/2$, provided that $n > 4\kappa \log \kappa$, where

$$\kappa = \max \left(1 + \frac{200}{3\pi a v_d r^d \cos^d(\theta)} \log \left(\frac{2}{\delta} \right), 4 \right).$$

Now we turn to \mathcal{E}_2 . Let $p_1, \dots, p_N \in M$ be such that $B_r(p_1), \dots, B_r(p_N)$ forms a minimal covering of M . From Lemma 7.7.2, we have that $N \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d r^d}$. Let $A_j = B_r(p_j)$. Then

$$\begin{aligned} Q(A_j) &\geq \frac{v_D S^D (1 - \pi)}{\text{vol}(\text{Box})} + \pi a v_d r^d \cos^d(\theta) \\ &\geq \pi a v_d r^d \cos^d(\theta) \equiv \gamma. \end{aligned}$$

Using again Lemma 7.7.2, if $n > \frac{1}{\gamma} (\log N + \log(\frac{2}{\delta}))$, then with probability at least $1 - \delta/2$, each A_i contains at least one sample point, and hence $M \subset \bigcup_{i \in I} B_{2r}(X_i)$, which implies that \mathcal{E}_2 holds.

Combining these we are now ready to again apply the main result from NSW. We restate this lemma in a slightly different form here.

Lemma 7.7.3. [NSW] *Let S be a set of points in the tubular neighborhood of radius R around M . Let $U = \bigcup_{x \in S} B_\epsilon(x)$. If S is R -dense in M then $\widehat{\mathcal{H}}(U) = \mathcal{H}(M)$ for all $R < (\sqrt{9} - \sqrt{8})\tau$, if $\epsilon = \frac{R + \tau}{2}$.*

Combining the previously established facts with the lemma above we obtain Lemma 7.5.4. Taking $r = (\sqrt{9} - \sqrt{8})\tau/4$ in that lemma, we can see that if $n \geq \frac{C_1}{\pi \tau^d} (\log \frac{C_2}{\tau^d} + \log(C_3/\epsilon))$ then we recover the correct homology with probability at least $1 - \epsilon$.

This is a sample complexity upper bound. Corresponding upper bounds on the minimax risk and resolution follow the arguments of the noiseless case.

Proofs for the tubular noise case

Lower bound In this setting we get samples uniformly in a full dimensional tube around the manifold. We are interested in the case when $\sigma \leq C_0 \tau$ for a small constant C_0 .

Let us denote the density q_1 at a point in the tube around M_1 by θ_1 and the density q_2 around M_2 by θ_2 . Since, it is not straightforward to decide whether $\theta_1 \leq \theta_2$ or not we will need to consider both possibilities. We will show the calculations assuming $\theta_1 \leq \theta_2$ (the other calculation follows similarly).

Now, remember from the definition of total variation $TV = q_1(G) - q_2(G)$ where G is the set where $q_1 > q_2$. We need an upper bound on total variation and so it suffices to use $TV \leq q_1(G^+) - q_2(G^-)$ where G^+ and G^- are sets containing and contained in G respectively.

Since, $\theta_1 < \theta_2$ we have G is contained in the holes (of radius 4τ) of the two annuli, and G contains a strip of width at least $2\tau - 2\sigma$ in these holes. These are G^+ and G^- .

We need to upper bound the mass under q_1 in G^+ and lower bound the mass under q_2 in G^- . We can now follow the a similar argument to the one made below (in the tubular noise upper bound) to obtain bounds on the various volumes. In each case, the volume of the tubular region is $\Omega(\text{vol}(M)\sigma^{D-d})$, and both M_1 and M_2 have constant volume, in particular $c_1 \leq \text{vol}(M) \leq C_1$. Giving us that the tubular region has volume $\Omega(\sigma^{D-d})$.

It is also clear that both G^+ and G^- have volumes that are $\Omega(\sigma^{D-d}\tau^d)$ (these can be calculated *exactly* since they are cylindrical with no additional curvature but we will not need this here). Here we use that σ is not too close to τ (and in particular is at most a constant fraction of τ).

Since q_1 and q_2 are both uniform in their respective tubes, it follows that

$$TV(q_1, q_2) \leq \Omega \left(\frac{\sigma^{D-d}\tau^d}{\sigma^{D-d}} \right) = \Omega(\tau^d).$$

Notice, that we assumed $\theta_1 \leq \theta_2$ above. The other calculation is nearly identical and we will not reproduce it here.

Upper bound Denote by M_σ the tube of radius σ around M . Recall that we are interested in the case when $\sigma \ll \tau$, and $\epsilon = \tau/2$.

Lemma 7.7.4. *If $\epsilon \gg \sigma$ (in particular $\epsilon \geq 2\sigma$ will suffice)*

$$k_\epsilon = \Omega(\epsilon^d).$$

Proof. For any $p \in M$,

$$Q(B_\epsilon(p)) = \frac{\text{vol}(B_\epsilon(p) \cap M_\sigma)}{\text{vol}(M_\sigma)}.$$

We will prove the claim by deriving an upper bound on the denominator and a lower bound on the numerator using packing/covering arguments, both bounds holding uniformly in p .

Upper bound on $\text{vol}(M_\sigma)$

We consider a covering of M by γ -balls of d dimensions, and denote the number of balls required N_γ , and the centers \mathcal{C}_γ . It is clear N_γ is bounded by the number of balls of radius $\gamma/2$ one can pack in M . A simple volume argument then gives

$$N_\gamma \leq C \frac{\text{vol}(M)}{(\gamma/2)^d},$$

for some constant C . Given this covering of M , it is easy to see that $\gamma + \sigma$ D -dimensional balls around each of the centers in \mathcal{C}_γ covers the tubular region. Thus, we have

$$\text{vol}(M_\sigma) \leq v_D N_\gamma (\gamma + \sigma)^D \leq v_D C \frac{\text{vol}(M)}{(\gamma/2)^d} (\gamma + \sigma)^D,$$

for any γ . Selecting $\gamma = \sigma$, we have

$$\text{vol}(M_\sigma) \leq C_{D,d} \text{vol}(M) \sigma^{D-d}$$

for some constant $C_{D,d}$ depending on the manifold and ambient dimensions, independent of σ .

Lower bound on $\text{vol}(B_\epsilon(p) \cap M_\sigma)$

Define

$$\begin{aligned} A(p) &= M \cap B_{\epsilon-\sigma}(p), \\ B(p) &= M \cap B_\epsilon(p), \\ B_\sigma(p) &= M_\sigma \cap B_\epsilon(p). \end{aligned}$$

Denote with N_σ the number of points we can “pack” in $A(p)$ such that the distance between any two points is at least 2σ . Denote the points themselves by the set \mathcal{C} . Then,

$$\text{vol}(B_\sigma) \geq N_\sigma v_D \sigma^D$$

where v_D is the volume of the unit ball in D -dimensions. To see this just note that every point that is at most σ away from any point in \mathcal{C} is contained in B_σ , and these sets are disjoint so the union of σ balls around \mathcal{C} is contained in B_σ .

Now, to prove a lower bound on N_σ we invoke some ideas from [146]. Consider, the map f described in Lemma 5.3 in [146], which projects the manifold onto its tangent space, and observe its action on $A(p)$. It is clear by their discussion that this map projects the manifold onto a superset of a ball of radius $(\epsilon - \sigma) \cos \theta$, for $\theta = \sin^{-1}(\frac{\epsilon - \sigma}{2\tau})$. In addition to being invertible, this map is a projection, and only shrinks distances between points. So if we can derive a lower bound on the number of points we can “pack” in this projection then it is also a lower bound on N_σ . Now, the set is just a ball in d -dimensions of radius $(\epsilon - \sigma) \cos \theta$. Using, the fact that 2σ balls around each of the points in \mathcal{C} must cover this set a simple volume argument shows

$$N_\sigma (2\sigma)^d \geq v_d ((\epsilon - \sigma) \cos \theta)^d,$$

i.e.

$$N_\sigma \geq C_{D,d} \left(\frac{(\epsilon - \sigma) \cos \theta}{\sigma} \right)^d,$$

which gives a lower bound.

Putting the upper and lower bound together, we get

$$\begin{aligned} k_\epsilon &= \inf_{p \in M} Q(B_\epsilon(p)) \\ &\geq C'_{D,d} \frac{1}{\text{vol}(M) \sigma^{D-d}} \left(\frac{(\epsilon - \sigma) \cos \theta}{\sigma} \right)^d \sigma^D \\ &= C'_{D,d} \frac{[(\epsilon - \sigma) \cos \theta]^d}{\text{vol}(M)}, \end{aligned}$$

for some quantity $C'_{D,d}$, independent of σ . □

We will prove the following main lemma.

Lemma 7.7.5. *Let N_ϵ be the ϵ -covering number of the submanifold M . Let $U = \bigcup_{i=1}^n B_{\epsilon+\tau/2}(X_i)$. Let $\hat{\mathcal{H}} = \mathcal{H}(U)$. Then if*

$$n > \frac{1}{k_\epsilon} (\log(N_\epsilon) + \log(1/\delta)),$$

$\mathbb{P}(\hat{\mathcal{H}} \neq \mathcal{H}(M)) < \delta$ as long as

$$\sigma \leq \epsilon/2 \text{ and } \epsilon < \frac{(\sqrt{9} - \sqrt{8})\tau}{2}.$$

Proof. This is a straightforward consequence of Lemma 7.7.3 and Lemma 7.7.2. □

Proof of Theorem 7.5.8 (additive case)

Lower Bound

From Lemma 7.7.9 we see that convolution only decreases the total variation distance, and so the lower bound for the noiseless case is still valid here.

Upper Bound

We will again proceed by a similar argument to the clutter noise case. Let $\sqrt{D}\sigma < r$, $R = 8r$ and $s = 4r$ and set $\alpha_s = \inf_{p \in A} Q(B_s(p))$ and $\beta_s = \sup_{p \in B} Q(B_s(p))$, where $A = \text{tube}_r(M)$, $B = \mathbb{R}^D - \text{tube}_R(M)$.

As in the clutter noise case, we will need the two events \mathcal{E}_1 and \mathcal{E}_2 to hold with high probability.

We will use the following version of a common χ^2 inequality, established by Niyogi et al. [147].

Lemma 7.7.6. *For a D -dimensional Gaussian random vector*

$$\mathbb{P}(\|\epsilon\| > \sqrt{T}) \leq (ze^{1-z})^{D/2}$$

where $z = \frac{T}{D\sigma^2}$

Using this inequality,

$$\mathbb{P}(\|\epsilon\| \geq 4r) \leq (16 \exp\{-15\})^{D/2} \equiv t$$

and

$$\mathbb{P}(\|\epsilon\| \geq 2r) \leq (4 \exp\{-3\})^{D/2} \equiv \gamma.$$

Observe that these are both constants. Next, it is easy to see that

$$\alpha_s \geq Q(B_{s-r}(p)) \geq av_d r^d (\cos \theta)^d (1 - \gamma) \equiv \alpha,$$

where $\theta = \sin^{-1}(r/(2\tau))$, and

$$\beta_s \leq v_D (8r)^D t \equiv \beta.$$

As in the clutter noise, we need β to be sufficiently smaller than α if we are to successfully clean the data. As we are interested in the case when r is small, if $D > d$ then we can take $\beta \leq \alpha/2$, while, if $D = d$ then we will need that the dimension is quite large (observe that both γ and t tend to zero rapidly as D grows).

We are now in a position to invoke the Lemma 7.7.8 to ensure \mathcal{E}_1 holds with high probability for n large enough. Further, one can see that the mass of an $r/2$ -ball close to manifold is at least

$$Q(A_i) \geq av_d (1 - \gamma) (\cos \theta)^d (r/2)^d$$

for $\theta = \sin^{-1}(r/(4\tau))$. This quantity is also $O(r^d)$ as desired, and for n large enough we can ensure \mathcal{E}_2 holds with high probability. Under the condition on σ , and r we have $r \leq \frac{(\sqrt{9}-\sqrt{8})\tau}{8}$. At this point we can invoke Theorem 5.1 from the paper [147] to see that for $n \succ^* \frac{1}{\tau^d}$ we recover the correct homology with high probability.

Deconvolution

Upper bound Recall, that the kernel Ψ satisfies

$$\Psi\{x : |x| \geq \epsilon\} \leq \gamma \tag{7.2}$$

with ϵ and γ being small constants that we will specify in our proof.

The starting point of our proof will be a uniform concentration result from Koltchinskii [113].

Lemma 7.7.7. *Consider the event*

$$A = \{\max_x |\widehat{P}_n(B_{2\epsilon}(x)) - \widehat{P}_\Psi(B_{2\epsilon}(x))| < \gamma\}.$$

For any small constants ϵ and γ , there exists $q \in (0, 1)$ such that

$$P(A^c) \leq 4q^n.$$

This lemma tells us that the deconvolved measure is uniformly close to a smoothed (by the kernel Ψ) version of the true density.

Our first step will be to draw

$$m > \frac{1}{\omega} \left(\log l + \log \left(\frac{2}{\delta} \right) \right)$$

samples from \widehat{P}_n , where $\omega = \inf_{x \in M} \widehat{P}_n(B_{2\epsilon}(x))$, and l is the 2ϵ covering number of the manifold, and $\delta = 8q^n$. Denote, this sample Z . We know that $l \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d(2\epsilon)^d}$.

Let us first show that we can choose ϵ and γ so that ω is at least a small positive constant.

$$\begin{aligned} \omega &= \inf_{x \in M} \widehat{P}_n(B_{2\epsilon}(x)) \\ &\geq \inf_{x \in M} P_\Psi(B_{2\epsilon}(x)) - \gamma. \end{aligned}$$

Notice that,

$$P_\Psi(B_{2\epsilon}) \geq P(B_\epsilon)\Psi(x : |x| \leq \epsilon).$$

So, we have,

$$\omega \geq \inf_{x \in M} P(B_\epsilon(x))(1 - \gamma) - \gamma.$$

Using the ball volume lemma we have,

$$\omega \geq av_d \epsilon^d \cos^d \theta (1 - \gamma) - \gamma$$

where $\theta = \sin^{-1}(\epsilon/2\tau)$. Notice, that τ is a fixed constant, and ϵ and γ are constants to be chosen appropriately. It is clear that for $\gamma \leq C_{d,\tau}\epsilon$, with $C_{d,\tau}$ small we have

$$\omega \geq c$$

for a small constant c which depends on τ, d and our choices of ϵ and γ .

We now use the sampling lemma 7.7.2 to conclude that w.p. at least $1 - 4q^n$,

1. The m samples are 4ϵ dense around M .
2. $M \subset \cup_{i=1}^m B_{4\epsilon}(x_i)$

Our next step will be a cleaning step. This cleaning procedure differs from the Algorithm CLEAN in that we use the deconvolved measure to clean the data. In particular, we will remove all points from Z for which $\widehat{P}_n(B_{4\epsilon}(Z_i)) \leq 2\gamma$. Denote the remaining points by W . Our estimator will then be constructed from

$$H = \bigcup B_{\frac{5\epsilon+\tau}{2}}(W_i).$$

To analyze this cleaning procedure, we use the uniform concentration lemma 7.7.7 above, and consider the case when event A happens.

1. **All points far away from M are eliminated:** In particular, for any point x if we have

$$\text{dist}(B_{4\epsilon}(x), M) \geq \epsilon$$

then the corresponding point is eliminated.

To see this is simple. We eliminated all points with deconvolved empirical mass $\widehat{P}_n(B_{4\epsilon}) < 2\gamma$. Since, we are assuming event A happened, we have for any remaining point $P_\Psi(B_{4\epsilon}) > \gamma$. Now, we have that

$$\Psi\{x : |x| \geq \epsilon\} \leq \gamma.$$

From this we see that some part of $B_{4\epsilon}$ must be within ϵ of M , and we have arrived at a contradiction.

2. **All points close to M are kept:** In particular, for any point x if

$$\text{dist}(x, M) \leq 2\epsilon$$

then the corresponding point is kept.

We need to show $\widehat{P}_n(B_{4\epsilon}(x)) \geq 2\gamma$. Notice, that $\widehat{P}_n(B_{4\epsilon}(x)) \geq \widehat{P}_n(B_{2\epsilon}(\pi(x)))$ where $\pi(x)$ is the projection of x onto M . This quantity is just ω .

To finish, we need to show that we can choose ϵ and γ such that $\omega \geq 2\gamma$. Since, $\omega \geq av_d \epsilon^d \cos^d \theta (1 - \gamma) - \gamma$ which as a function of γ is continuous, bounded from below by a constant depending on τ, d and ϵ and monotonically increasing as γ decreases we have for γ small enough

$$\omega \geq 2\gamma.$$

3. **The set H has the right homology:** We have shown that the cleaning eliminates all points outside a tube of radius 5ϵ , and further keeps all points in a tube of radius 2ϵ . From the sampling result we know the points that we keep are 4ϵ dense and that $M \subset \cup_{i=1}^m B_{4\epsilon}(x_i)$. We can now apply lemma 7.7.3 to conclude that H has the right homology provided

$$\epsilon < \frac{(\sqrt{9} - \sqrt{8})\tau}{5}.$$

Since τ is a fixed constant we can always choose ϵ small enough to satisfy this condition. To review, we need to select γ and ϵ to satisfy three conditions

- (a) $\omega \geq av_d \epsilon^d \cos^d \theta (1 - \gamma) - \gamma$ has to be atleast a small positive constant.
- (b) $\omega \geq 2\gamma$.
- (c) $\epsilon < \frac{(\sqrt{9} - \sqrt{8})\tau}{5}$.

Each of these can be satisfied by choosing γ and ϵ small enough.

Now, returning to m . We have

$$m > \frac{1}{\omega} \left(\log l + \log \left(\frac{2}{\delta} \right) \right)$$

where $\omega = \inf_{x \in M} \widehat{P}_n(B_{2\epsilon}(x))$, and l is the 2ϵ covering number of the manifold $l \leq \frac{\text{vol}(M)}{\cos^d(\theta)v_d(2\epsilon)^d}$, and $\delta = 8q^n$. It is clear that all terms except those in n are constant. In particular it is easy to see that

$$m \geq Cn$$

for C large enough is sufficient.

From this we can conclude with probability at least $1 - 8q^n$ our procedure will construct an estimator with the correct homology. Since, $q \in (0, 1)$ the success probability can be re-written as at least $1 - e^{-cn}$ for c small enough. Together this gives us the deconvolution lemma.

7.7.2 Additional technical lemmas

The cleaning lemma

In this section we sharpen Lemma 4.1 of Niyogi et al. [147], also known as the A-B lemma, by using Bernstein's inequality instead of Hoeffding's inequality. This modification is crucial to obtain minimax rates.

Lemma 7.7.8. *Let $\beta_s \leq \beta < \alpha/2 \leq \alpha_s/2$. If $n > 4\beta \log \beta$, where*

$$\beta = \max \left(1 + \frac{200}{3\alpha} \log \left(\frac{1}{\delta} \right), 4 \right),$$

then procedure CLEAN($\frac{\alpha+\beta}{2}$) will remove all points in region B and keep all points in region A with probability at least $1 - \delta$.

Proof. We use the notation established in section 7.5.2. We first analyze the set A .

For a point X_i in A , let $q = q(i) = Q(B_s(X_i))$, and define,

$$Z_j = \mathbb{I}(X_j \in B_s(X_i)), \quad j \neq i,$$

where \mathbb{I} denotes the indicator function. Notice that the random variables $\{Z_j, j \neq i\}$ are independent Bernoulli with common mean q .

We will consider two cases.

Case 1: $\alpha \leq q \leq 2\alpha$.

Notice that if

$$q - \frac{1}{n-1} \sum_{j \neq i} Z_j \leq \frac{\alpha}{4}$$

the point X_i will not be removed. By Bernstein's inequality, the probability that X_i will instead be removed is

$$\begin{aligned} \mathbb{P} \left(q - \frac{1}{n-1} \sum_{j \neq i} Z_j \geq \frac{\alpha}{4} \right) &\leq \exp \left\{ -\frac{1}{2} \frac{(n-1)(\alpha/4)^2}{2\alpha + \alpha/12} \right\} \\ &\leq \exp \left\{ -\frac{3}{200} (n-1)\alpha \right\}. \end{aligned}$$

Case 2: $q > 2\alpha$.

In this case if

$$q - \frac{1}{n-1} \sum_{j \neq i} Z_j \leq q - \frac{3\alpha}{4}$$

the point X_i will be removed. Another application of Bernstein's inequality yields

$$\begin{aligned} \mathbb{P} &\left(q - \frac{1}{n-1} \sum_{j \neq i} Z_j \geq q - \frac{3\alpha}{4} \right) \\ &\leq \exp \left\{ -\frac{1}{2} \frac{(n-1)(q - 3\alpha/4)^2}{q + (q - 3\alpha/4)/3} \right\} \\ &\leq \exp \left\{ -\frac{1}{2} (n-1) \left[\frac{q}{2} + \frac{9\alpha^2}{32p} - \frac{3\alpha}{4} \right] \right\} \\ &\leq \exp \left\{ -\frac{(n-1)\alpha}{8} \right\}. \end{aligned}$$

Now, consider a point X_i in the region B, and define q and the Z_j s in an identical way. This time if

$$\frac{1}{n-1} \sum_{j \neq i} Z_j - q \leq \frac{\alpha}{4},$$

the point X_i will not be removed. By Bernstein's inequality,

$$\begin{aligned} \mathbb{P} \left(\frac{1}{n-1} \sum_{j \neq i} Z_j - q \geq \frac{\alpha}{4} \right) &\leq \exp \left\{ -\frac{1}{2} \frac{(n-1)(\alpha/4)^2}{\alpha/2 + \alpha/12} \right\} \\ &\leq \exp \left\{ -\frac{3}{56} (n-1)\alpha \right\}. \end{aligned}$$

Putting all the pieces together, we obtain that the cleaning procedure succeeds on all points with probability at least $n \exp \left\{ -\frac{3}{200} (n-1)\alpha \right\}$. This requires,

$$\begin{aligned} n-1 &> \frac{200}{3\alpha} \left(\log n + \log \left(\frac{1}{\delta} \right) \right) \quad \text{i.e.} \\ n &> 1 + \frac{200}{3\alpha} \log \left(\frac{1}{\delta} \right) + \frac{200}{3\alpha} \log n. \end{aligned}$$

If $\delta < 1/2$, then $1 + \frac{200}{3\alpha} \log \left(\frac{1}{\delta} \right) > \frac{200}{3\alpha}$, so it is enough to solve

$$n > x + x \log n$$

with $x = 1 + \frac{200}{3\alpha} \log \left(\frac{1}{\delta} \right)$. The result of the lemma follows. \square

Convolution only decreases total variation

Lemma 7.7.9. *Let P and Q two probability measures in \mathbb{R}^D with common dominating measure μ . Then,*

$$\text{TV}(P \star \Phi, Q \star \Phi) \leq C_\Phi \text{TV}(P, Q).$$

where \star denotes deconvolution and Φ is a probability measure on \mathbb{R}^D .

Proof. This is a standard result, but we provide a proof for completeness. Let $p \star \phi$ denote the Lebesgue density of the probability distribution $P \star \Phi$, i.e.

$$p \star \phi(z) = \int \phi(z-x)p(x)d\mu(x), \quad z \in \mathbb{R}^D.$$

Similarly, $q \star \phi$ denotes the analogous quantity for $Q \star \Phi$. Then,

$$\begin{aligned}
2\text{TV}(P \star \Phi, Q \star \Phi) &= \int_{\mathbb{R}^D} |p \star \phi(z) - q \star \phi(z)| dz \\
&= \int_{\mathbb{R}^D} \left| \int \phi(z-x)p(x)d\mu(x) \right. \\
&\quad \left. - \int \phi(z-x)p(x)d\mu(x) \right| dz \\
&= \int_{\mathbb{R}^D} \left| \int \phi(z-x)(p(x) \right. \\
&\quad \left. - q(x))d\mu(x) \right| dz \\
&\leq \int_{\mathbb{R}^D} \int |\phi(z-x)(p(x) \\
&\quad - q(x))| d\mu(x) dz \\
&\leq \int \int_{\mathbb{R}^D} \phi(z-x) dz |p(x) - q(x)| d\mu(x) \\
&= \int |(p(x) - q(x))| d\mu(x) \\
&= 2\text{TV}(P, Q).
\end{aligned}$$

□

Chapter 8

Cluster Trees on Manifolds

In this chapter we investigate the problem of estimating the cluster tree for a density f supported on or near a smooth d -dimensional manifold M isometrically embedded in \mathbb{R}^D . We analyze a modified version of a k -nearest neighbor based algorithm recently proposed by Chaudhuri and Dasgupta [44]. The main results of this chapter show that under mild assumptions on f and M , we obtain rates of convergence that depend on d only but not on the ambient dimension D . We also show that similar (albeit non-algorithmic) results can be obtained for kernel density estimators. We sketch a construction of a sample complexity lower bound instance for a natural class of *manifold oblivious* clustering algorithms. We further briefly consider the *known* manifold case and show that in this case a spatially adaptive algorithm achieves better rates.

8.1 Introduction

In this chapter, we study the problem of estimating the cluster tree of a density when the density is supported on or near a manifold. Let $\mathbf{X} := \{X_1, \dots, X_n\}$ be a sample drawn i.i.d. from a distribution P with density f . The connected components $\mathbb{C}_f(\lambda)$ of the upper level set $\{x : f(x) \geq \lambda\}$ are called *density clusters*. The collection $\mathcal{C} = \{\mathbb{C}_f(\lambda) : \lambda \geq 0\}$ of all such clusters is called the *cluster tree* and estimating this cluster tree is referred to as *density clustering*.

The density clustering paradigm is attractive for various reasons. One of the main difficulties of clustering is that often the true goals of clustering are not clear and this makes clusters, and clustering as a task seem poorly defined. Density clustering however is estimating a well defined population quantity, making its goal, consistent recovery of the *population* density clusters, clear. Typically only mild assumptions are made on the density f and this allows extremely general shapes and numbers of clusters at each level. Finally, the *cluster tree* is an inherently hierarchical object and thus density clustering algorithms typically do not require specification of the “right” level, rather they capture a summary of the density across all levels.

The search for a simple, statistically consistent estimator of the cluster tree has a long history. Hartigan [90] showed that the popular single-linkage algorithm is not consistent for a sample

from \mathbb{R}^D , with $D > 1$. Recently, Chaudhuri and Dasgupta [44] analyzed an algorithm which is both simple and consistent. The algorithm finds the connected components of a sequence of carefully constructed neighborhood graphs. They showed that, as long as the parameters of the algorithm are chosen appropriately, the resulting collection of connected components correctly estimates the cluster tree with high probability.

In this chapter, we are concerned with the problem of estimating the cluster tree when the density f is supported on or near a low dimensional manifold. The motivation for this work stems from the problem of devising and analyzing clustering algorithms with provable performance that can be used in high dimensional applications. When data live in high dimensions, clustering (as well as other statistical tasks) generally become prohibitively difficult due to the curse of dimensionality, which demands a very large sample size. In many high dimensional applications however data is not spread uniformly but rather concentrates around a low dimensional set. This so-called manifold hypothesis motivates the study of data generated on or near low dimensional manifolds and the study of procedures that can adapt effectively to the intrinsic dimensionality of this data.

8.1.1 Contributions

Here is a brief summary of the main contributions of this chapter:

1. We show that the simple algorithm studied in the paper [44] is consistent and has fast rates of convergence for data on or near a low dimensional manifold M . The algorithm does not require the user to first estimate M (which is a difficult problem). In other words, the algorithm adapts to the (unknown) manifold.
2. We show that the sample complexity for identifying salient clusters is independent of the ambient dimension.
3. We sketch a construction of a sample complexity lower bound instance for a natural class of clustering algorithms that we study in this chapter.
4. We show that in the *known* manifold case a modified *spatially adaptive* algorithm achieves better rates, similar to the near minimax-optimal rates of Chaudhuri and Dasgupta [44].
5. We introduce a framework for studying consistency of clustering when the distribution is not supported on a manifold but rather, is concentrated near a manifold. The generative model in this case is that the data are first sampled from a distribution on a manifold and then noise is added. The original data are latent (unobserved). We show that for certain noise models we can still efficiently recover the cluster tree on the *latent* samples.
6. We show similar *statistical* results for the level sets of kernel density estimates for an appropriately chosen bandwidth. *Computing* the level sets of the kernel density estimate is however a challenging problem that we do not address in this thesis.
7. We present some simulations to confirm our theoretical results.

8.1.2 Related Work

The idea of using probability density functions for clustering dates back to Wishart [203]. Hartigan [90] expanded on this idea and formalized the notions of high-density clustering, of the cluster tree and of consistency and fractional consistency of clustering algorithms. In particular, Hartigan [90] showed that single linkage clustering is consistent when $D = 1$ but is only fractionally consistent when $D > 1$. Stuetzle and R. [181] and Stuetzle [180] have also proposed procedures for recovering the cluster tree. None of these procedures however, come with the theoretical guarantees given by Chaudhuri and Dasgupta [44], which demonstrated that a generalization of Wishart’s algorithm allows one to estimate parts of the cluster tree for distributions with full-dimensional support near-optimally under rather mild assumptions. This paper forms the starting point for our work and is reviewed in more detail in the next section.

In the last two decades, much of the research effort involving the use of nonparametric density estimators for clustering has focused on the more specialized problems of optimal estimation of the support of the distribution or of a fixed level set. However, consistency of estimators of a fixed level set does not imply cluster tree consistency, and extending the techniques and analyses mentioned above to hold simultaneously over a variety of density levels is non-trivial. See for example the papers [52, 53, 135, 155, 159, 160, 161, 172, 193, 200], and references therein. Estimating the cluster tree has more recently been considered by Kpotufe and von Luxburg [117] who also give a simple pruning procedure for removing spurious clusters. Steinwart [178] and Sriperumbudur and Steinwart [177] propose procedures for determining recursively the lowest split in the cluster tree and give conditions for asymptotic consistency with minimal assumptions on the density.

8.2 Background and Assumptions

Let P be a distribution supported on an unknown d -dimensional manifold M . We assume that the manifold M is a d -dimensional Riemannian manifold without boundary embedded in a compact set $\mathcal{X} \subset \mathbb{R}^D$ with $d < D$. We further assume that the volume of the manifold is bounded from above by a constant, i.e., $\text{vol}_d(M) \leq C$. The main regularity condition we impose on M is that its condition number be not too large. The *condition number* of M is $1/\tau$, where τ is the largest number such that the open normal bundle about M of radius r is imbedded in \mathbb{R}^D for every $r < \tau$. The condition number is discussed in more detail in the previous chapter as well as the paper [146].

The Euclidean norm is denoted by $\|\cdot\|$ and v_d denotes the volume of the d -dimensional unit ball in \mathbb{R}^d . $B(x, r)$ denotes the full-dimensional ball of radius r centered at x and $B_M(x, r) := B(x, r) \cap M$. For $Z \subset \mathbb{R}^d$ and $\sigma > 0$, define $Z_\sigma = Z + B(0, \sigma)$ and $Z_{M,\sigma} = (Z + B(0, \sigma)) \cap M$. Note that Z_σ is full dimensional, while if $Z \subseteq M$ then $Z_{M,\sigma}$ is d -dimensional.

Let f be the density of P with respect to the uniform measure on M . For $\lambda \geq 0$, let $\mathbb{C}_f(\lambda)$ be the collection of connected components of the level set $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ and define the

cluster tree of f to be the hierarchy $\mathcal{C} = \{\mathbb{C}_f(\lambda) : \lambda \geq 0\}$. For a fixed λ , any member of $\mathbb{C}_f(\lambda)$ is a cluster. For a cluster C its restriction to the sample \mathbf{X} is defined to be $C[\mathbf{X}] = C \cap \mathbf{X}$. The restriction of the cluster tree \mathcal{C} to \mathbf{X} is defined to be $\mathcal{C}[\mathbf{X}] = \{C \cap \mathbf{X} : C \in \mathcal{C}\}$. Informally, this restriction is a dendrogram-like hierarchical partition of \mathbf{X} .

To give finite sample results, following Chaudhuri and Dasgupta [44], we define the notion of salient clusters. Our definitions are slight modifications of those in Chaudhuri and Dasgupta [44] to take into account the manifold assumption.

Definition 9. *Clusters A and A' are (σ, ϵ) separated if there exists a nonempty $S \subset M$ such that:*

1. *Any path along M from A to A' intersects S .*
2. $\sup_{x \in S_{M, \sigma}} f(x) < (1 - \epsilon) \inf_{x \in A_{M, \sigma} \cup A'_{M, \sigma}} f(x)$.

Chaudhuri and Dasgupta [44] analyze a robust single linkage (RSL) algorithm (in Figure 8.1). An RSL algorithm estimates the connected components at a level λ in two stages. In the first stage, the sample is *cleaned* by thresholding the k -nearest neighbor distance of the sample points at a radius r and then, in the second stage, the cleaned sample is *connected* at a connection radius R . The connected components of the resulting graph give an estimate of the restriction $\mathbb{C}_f(\lambda)[\mathbf{X}]$. In Section 8.4 we prove a sample complexity lower bound for the *class of RSL algorithms* which we now define.

Definition 10. *The class of RSL algorithms refers to any algorithm that is of the form described in the algorithm in Figure 8.1 and relying on Euclidean balls, with any choice of k , r and R .*

We define two notions of consistency for an estimator $\hat{\mathcal{C}}$ of the cluster tree:

Definition 11 (Hartigan consistency). *For any sets $A, A' \subset \mathcal{X}$, let A_n (resp., A'_n) denote the smallest cluster of $\hat{\mathcal{C}}$ containing $A \cap \mathbf{X}$ (resp., $A' \cap \mathbf{X}$). We say $\hat{\mathcal{C}}$ is consistent if, whenever A and A' are different connected components of $\{x : f(x) \geq \lambda\}$ (for some $\lambda > 0$), the probability that A_n is disconnected from A'_n approaches 1 as $n \rightarrow \infty$.*

Definition 12 ((σ, ϵ) consistency). *For any sets $A, A' \subset \mathcal{X}$ such that A and A' are (σ, ϵ) separated, let A_n (resp., A'_n) denote the smallest cluster of $\hat{\mathcal{C}}$ containing $A \cap \mathbf{X}$ (resp., $A' \cap \mathbf{X}$). We say $\hat{\mathcal{C}}$ is consistent if, whenever A and A' are different connected components of $\{x : f(x) \geq \lambda\}$ (for some $\lambda > 0$), the probability that A_n is disconnected from A'_n approaches 1 as $n \rightarrow \infty$.*

The notion of (σ, ϵ) consistency is similar that of Hartigan consistency except restricted to (σ, ϵ) separated clusters A and A' , and typically associated with a finite sample of size n . Chaudhuri and Dasgupta [44] prove the following theorem, establishing finite sample bounds for a particular RSL algorithm. In this theorem there is no manifold and f is a density with respect to the Lebesgue measure on \mathbb{R}^D .

Theorem 8.2.1. *There is a constant C such that the following holds. Suppose that we run the algorithm in Figure 8.1 with*

$$R = \sqrt{2}r \quad \text{and} \quad k = C \left(\frac{D \log n}{\epsilon^2} \right) \log^2(1/\delta)$$

1. For each X_i , $r_k(X_i) := \inf\{r : B(X_i, r) \text{ contains } k \text{ data points}\}$.
2. As r grows from 0 to ∞ :
 - (a) Construct a graph $G_{r,R}$ with nodes $\{X_i : r_k(X_i) \leq r\}$ and edges (X_i, X_j) if $\|X_i - X_j\| \leq R$.
 - (b) Let $\mathbb{C}(r)$ be the connected components of $G_{r,R}$.
3. Denote $\widehat{\mathcal{C}} = \{\mathbb{C}(r) : r \in [0, \infty)\}$ and return $\widehat{\mathcal{C}}$.

Figure 8.1: Robust Single Linkage (RSL) Algorithm

then with probability at least $1 - \delta$, the algorithm output $\widehat{\mathcal{C}}$ is (σ, ϵ) consistent provided

$$\lambda \geq \frac{1}{v_D(\sigma/2)^D} \frac{k}{n} \left(1 + \frac{\epsilon}{2}\right).$$

The theorem as stated does not explicitly give a sample complexity bound but it is straightforward to obtain one by plugging in the value for k and solving for n in the inequality that restricts λ to be large enough (as a function of n).

In particular, notice that if

$$n \geq O\left(\frac{D}{\lambda \epsilon^2 v_D(\sigma/2)^D} \log \frac{D}{\lambda \epsilon^2 v_D(\sigma/2)^D}\right)$$

then we can resolve any pair of (σ, ϵ) clusters at level at least λ . It is important to note that this theorem does not apply to the setting when distributions are supported on a lower dimensional set for at least two reasons: (1) the density f is singular with respect to the Lebesgue measure on \mathcal{X} and so the cluster tree is trivial, and (2) the definitions of saliency with respect to \mathcal{X} are typically not satisfied when f has a lower dimensional support.

8.3 Clustering on Manifolds

In this section we show that the RSL algorithm can be adapted to recover the cluster tree of a distribution supported on a manifold of dimension $d < D$ with the rates depending only on d . In place of the cluster salience parameter σ , our rates involve a new parameter ρ

$$\rho := \min\left(\frac{3\sigma}{16}, \frac{\epsilon\tau}{72d}, \frac{\tau}{16}\right).$$

The precise reason for this definition of ρ will be clear from the proofs (particularly of Lemma 8.3.3) but for now notice that in addition to σ it is dependent on the condition number $1/\tau$ and deteriorates as the condition number increases. Finally, to succinctly present our results we use $\mu := \log n + d \log(1/\rho)$.

Theorem 8.3.1. *There are universal constants C_1 and C_2 such that the following holds. For any $\delta > 0$, $0 < \epsilon < 1/2$, run the algorithm in Figure 8.1 on a sample \mathbf{X} drawn from f , where the parameters are set according to the equations*

$$R = 4\rho \quad \text{and} \quad k = C_1 \log^2(1/\delta)(\mu/\epsilon^2).$$

Then with probability at least $1 - \delta$, $\widehat{\mathcal{C}}$ is (σ, ϵ) consistent. In particular, the clusters containing $A[\mathbf{X}]$ and $A'[\mathbf{X}]$, where A and A' are (σ, ϵ) separated, are internally connected and mutually disconnected in $\mathbb{C}(r)$ for r defined by

$$v_d r^d \lambda = \frac{1}{1 - \epsilon/6} \left(\frac{k}{n} + \frac{C_2 \log(1/\delta)}{n} \sqrt{k\mu} \right)$$

provided

$$\lambda \geq \frac{2}{v_d \rho^d} \frac{k}{n}.$$

Before we prove this theorem a few remarks are in order:

1. To obtain an explicit sample complexity, as in Theorem 8.2.1, we plug in the value of k and solve for n from the inequality restricting λ . The sample complexity of the RSL algorithm for recovering (σ, ϵ) clusters at level at least λ on a manifold M with condition number at most $1/\tau$ is

$$n = O \left(\frac{d}{\lambda \epsilon^2 v_d \rho^d} \log \frac{d}{\lambda \epsilon^2 v_d \rho^d} \right)$$

where $\rho = C \min(\sigma, \epsilon\tau/d, \tau)$. Ignoring constants that depend on d the main difference between this result and the result of Chaudhuri and Dasgupta [44] (Theorem 8.2.1) is that our results only depend on the manifold dimension d and not the ambient dimension D (typically $D \gg d$). There is also a dependence of our result on $1/(\epsilon\tau)^d$, for $\epsilon\tau \ll \sigma$. In Section 8.4 we sketch the construction of an instance that suggests that this dependence is not an artifact of our analysis and that the sample complexity of the class of RSL algorithms is at least $n \geq 1/(\epsilon\tau)^{\Omega(d)}$.

2. Another aspect is that our choice of the connection radius R depends on the (typically) unknown ρ , while for comparison, the connection radius in Chaudhuri and Dasgupta [44] is chosen to be $\sqrt{2}r$. Under the mild assumption that $\lambda \leq n^{O(1)}$ (which is satisfied for instance, if the density on M is bounded from above), we show in Section 8.9.8 that an identical theorem holds for $R = 4r$. k is the only real tuning parameter of this algorithm whose choice depends on ϵ and an unknown leading constant.
3. It is easy to see that this theorem also establishes consistency for recovering the entire cluster tree by selecting an appropriate schedule on σ_n, ϵ_n and k_n that ensures that *all* clusters are distinguished for n large enough (see Chaudhuri and Dasgupta [44] for a formal proof).

Our proofs structurally mirror those in Chaudhuri and Dasgupta [44]. We begin with a few technical results in 8.3.1. In Section 8.3.2 we establish (σ, ϵ) consistency by showing that the

clusters are mutually disjoint and internally connected. The main technical challenge is that the curvature of the manifold, modulated by its condition number $1/\tau$, limits our ability to resolve the density level sets from a finite sample, by limiting the maximum cleaning and connection radii the algorithm can use. In what follows, we carefully analyze this effect and show that somewhat surprisingly, despite this curvature, essentially the same algorithm is able to adapt to the unknown manifold and produce a consistent estimate of the entire cluster tree. Similar manifold adaptivity results have been shown in classification [54] and in non-parametric regression [30, 116].

8.3.1 Technical results

In our proof, we use the uniform convergence of the empirical mass of Euclidean balls to their true mass. In the full dimensional setting of Chaudhuri and Dasgupta [44], this follows from standard VC inequalities. To the best of our knowledge however sharp (ambient dimension independent) inequalities for manifolds are unknown. We get around this obstacle by using the insight that, in order to analyze the RSL algorithms, uniform convergence for Euclidean balls around the *sample points* and around a *fixed minimum s -net \mathcal{N} of M* (for an appropriately chosen s) suffice to analyze the RSL algorithm.

Recall, an s -net $\mathcal{N} \subseteq M$ is such that every point of M is at a distance at most s from some point in \mathcal{N} . Let

$$\mathcal{B}_{n,\mathcal{N}} := \left\{ B(z, s) : z \in \mathcal{N} \cup \mathbf{X}, s \geq 0 \right\}$$

be the collection of balls whose centers are sample or net points. We are ready to state our uniform convergence lemma. The proof is in Section 8.9.3.

Lemma 8.3.2 (Uniform Convergence). *Assume $k \geq \mu$. Then there exists a constant C_0 such that the following holds. For every $\delta > 0$, with probability $> 1 - \delta$, for all $B \in \mathcal{B}_{n,\mathcal{N}}$, we have:*

$$\begin{aligned} P(B) \geq \frac{C_\delta \mu}{n} &\implies P_n(B) > 0, \\ P(B) \geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu} &\implies P_n(B) \geq \frac{k}{n}, \\ P(B) \leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu} &\implies P_n(B) < \frac{k}{n}, \end{aligned}$$

where $C_\delta := 2C_0 \log(2/\delta)$, and $\mu := 1 + \log n + \log |\mathcal{N}| = Cd + \log n + d \log(1/s)$. Here $P_n(B) = |\mathbf{X} \cap B|/n$ denotes the empirical probability measure of B , and C is a universal constant.

Next we provide a tight estimate of the volume of a small ball intersected with M . This bounds the distortion of the apparent density due to the curvature of the manifold and is central to many of our arguments. Intuitively, the claim states that the volume is approximately that of a d -dimensional Euclidean ball, provided that its radius is small enough compared to τ . The lower bound is based on Lemma 5.3 of Niyogi et al. [146] while the upper bound is based on a modification of the main result of Chazal [46].

Lemma 8.3.3 (Ball volumes). *Assume $r < \tau/2$. Define $S := B(x, r) \cap M$ for a point $x \in M$. Then*

$$\left(1 - \frac{r^2}{4\tau^2}\right)^{d/2} v_d r^d \leq \text{vol}_d(S) \leq v_d \left(\frac{\tau}{\tau - 2r}\right)^d r^d,$$

where $r_1 = \tau - \tau\sqrt{1 - 2r/\tau}$. In particular, if $r \leq \epsilon\tau/72d$ for $0 \leq \epsilon < 1$, then

$$v_d r^d (1 - \epsilon/6) \leq \text{vol}_d(S) \leq v_d r^d (1 + \epsilon/6).$$

8.3.2 Separation and Connectedness

Lemma 8.3.4 (Separation). *Assume that we pick k , r and R to satisfy the conditions:*

$$\begin{aligned} r &\leq \rho \\ R &= 4\rho \\ v_d r^d (1 - \epsilon/6) \lambda &\geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu} \\ v_d r^d (1 + \epsilon/6) \lambda (1 - \epsilon) &\leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}. \end{aligned}$$

Then with probability $1 - \delta$, we have:

1. All points in $A_{\sigma-r}$ and $A'_{\sigma-r}$ are kept, and all points in $S_{\sigma-r}$ are removed.
2. The two point sets $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ are disconnected in $G_{r,R}$.

Proof. The proof is analogous to the separation proof of Chaudhuri and Dasgupta [44] with several modifications. Most importantly, we need to ensure that despite the curvature of the manifold we can still resolve the density well enough to guarantee that we can identify and eliminate points in the region of separation.

Throughout the proof, we will assume that the good event in Lemma 8.3.2 (uniform convergence for $\mathcal{B}_{n,\mathcal{N}}$) occurs. Since $r \leq \epsilon\tau/72d$, by Lemma 8.3.3 $\text{vol}(B_M(x, r))$ is between $v_d r^d (1 - \epsilon/6)$ and $v_d r^d (1 + \epsilon/6)$, for any $x \in M$. So if $X_i \in A \cup A'$, then $B_M(X_i, r)$ has mass at least $v_d r^d (1 - \epsilon/6) \cdot \lambda$. Since this is $\geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}$ by assumption, this ball contains at least k sample points, and hence X_i is kept.

On the other hand, if $X_i \in S_{\sigma-r}$, then the set $B_M(X_i, r)$ contains mass at most $v_d r^d (1 + \epsilon/6) \cdot \lambda (1 - \epsilon)$. This is $\leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}$. Thus by Lemma 8.3.2 $B_M(X_i, r)$ contains fewer than k sample points, and hence X_i is removed.

To prove the graph is disconnected, we first need a bound on the geodesic distance between two points that are at most R apart in Euclidean distance. Such an estimate follows from Proposition 6.3 in Niyogi et al. [146] who show that if $\|p - q\| = R \leq \tau/2$, then the geodesic distance

$$d_M(p, q) \leq \tau - \tau\sqrt{1 - \frac{2R}{\tau}}.$$

In particular, if $R \leq \tau/4$, then $d_M(p, q) < R(1 + \frac{4R}{\tau}) \leq 2R$. Now, notice that if the graph is connected there must be an edge that connects two points that are at a geodesic distance of at least $2(\sigma - r)$. Any path between a point in A and a point in A' along M must pass through $S_{\sigma-r}$ and must have a geodesic length of at least $2(\sigma - r)$. This is impossible if the connection radius satisfies $2R < 2(\sigma - r)$, which follows by the assumptions on r and R . \square

All the conditions in Lemma 8.3.4 can be simultaneously satisfied by setting $k := 16C_\delta^2(\mu/\epsilon^2)$, and

$$v_d r^d (1 - \epsilon/6) \cdot \lambda = \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}. \quad (8.1)$$

The condition on r is satisfied since

$$\lambda \geq \frac{2}{v_d \rho^d} \frac{k}{n}$$

and the condition on R is satisfied by its definition.

Lemma 8.3.5 (Connectedness). *Assume that the parameters k, r and R satisfy the separation conditions (in Lemma 8.3.4). Then, with probability at least $1 - \delta$, $A[\mathbf{X}]$ is connected in $G_{r,R}$.*

Proof. Let us show that any two points in $A \cap \mathbf{X}$ are connected in $G_{r,R}$. Consider $y, y' \in A \cap \mathbf{X}$. Since A is connected, there is a path P between y, y' lying entirely inside A , i.e., a continuous map $P : [0, 1] \rightarrow A$ such that $P(0) = y$ and $P(1) = y'$. We can find a sequence of points $y_0, \dots, y_t \in P$ such that $y_0 = y$, $y_t = y'$, and the geodesic distance on M (and hence the Euclidean distance) between y_{i-1} and y_i is at most η , for an arbitrarily small constant η .

Let \mathcal{N} be minimal $R/4$ -net of M . There exist $z_i \in \mathcal{N}$ such that $\|y_i - z_i\| \leq R/4$. Since $y_i \in A$, we have $z_i \in A_{M, R/4}$, and hence the ball $B_M(z_i, R/4)$ lies completely inside $A_{M, R/2} \subseteq A_{M, \sigma-r}$. In particular, the density inside the ball is at least λ everywhere, and hence the mass inside it is at least

$$v_d (R/4)^d (1 - \epsilon/6) \lambda \geq \frac{C_\delta \mu}{n}.$$

Observe that $R \geq 4r$ and so this condition is satisfied as a consequence of satisfying Equation 8.1. Thus Lemma 8.3.2 guarantees that the ball $B_M(z_i, R/4)$ contains at least one sample point, say x_i . (Without loss of generality, we may assume $x_0 = y$ and $x_t = y'$.) Since the ball lies completely in $A_{M, \sigma-r}$, the sample point x_i is not removed in the cleaning step (Lemma 8.3.4).

Finally, we bound $d(x_{i-1}, x_i)$ by considering the sequence of points $(x_{i-1}, z_{i-1}, y_{i-1}, y_i, z_i, x_i)$. The pair (y_{i-1}, y_i) are at most s apart and the other successive pairs at most $R/4$ apart, hence $d(x_{i-1}, x_i) \leq 4(R/4) + \eta = R + \eta$. The claim follows by letting $\eta \rightarrow 0$. \square

8.4 A lower bound instance for the class of RSL algorithms

Recall that the sample complexity in Theorem 8.3.1 scales as

$$n = O\left(\frac{d}{\lambda \epsilon^2 v_d \rho^d} \log \frac{d}{\lambda \epsilon^2 v_d \rho^d}\right)$$

where $\rho = C \min(\sigma, \epsilon\tau/d, \tau)$. For full dimensional densities, Chaudhuri and Dasgupta [44] showed the information theoretic lower bound

$$n = \Omega\left(\frac{1}{\lambda\epsilon^2 v_D \sigma^D} \log \frac{1}{\lambda\epsilon^2 v_D \sigma^D}\right).$$

Their construction can be straightforwardly modified to a d -dimensional instance on a smooth manifold. Ignoring constants that depend on d , these upper and lower bounds can still differ by a factor of $1/(\epsilon\tau)^d$, for $\epsilon\tau \ll \sigma$. In this section we provide an informal sketch of a hard instance for the class of RSL algorithms (see Definition 10) that suggests a sample complexity lower bound of $n \geq 1/(\epsilon\tau)^{\Omega(d)}$.

We first describe our lower bound instance. The manifold M consists of two disjoint components, C and C' . The component C in turn contains three parts, which we call ‘top’, ‘middle’, and ‘bottom’ respectively. The middle part, denoted M_2 , is the portion of the standard d -dimensional unit sphere $\mathbb{S}^d(0, 1)$ between the planes $x_1 = +\sqrt{1 - 4\tau^2}$ and $x_1 = -\sqrt{1 - 4\tau^2}$. The top part, denoted M_1 , is the upper hemisphere of radius 2τ centered at $(+\sqrt{1 - 4\tau^2}, 0, 0, \dots, 0)$. The bottom part, denoted M_3 , is a symmetric hemisphere centered at $(-\sqrt{1 - 4\tau^2}, 0, 0, \dots, 0)$. Thus C is obtained by gluing a portion of the unit sphere with two (small) hemispherical caps. C as described does not have a condition number at most $1/\tau$ because of the ‘‘corners’’ at the intersection of M_2 and $M_1 \cup M_3$. This can be fixed without affecting the essence of the construction by smoothing this intersection by rolling a ball of radius τ around it (a similar construction is made rigorous in Theorem 6 of Genovese et al. [81]). Finally, the component C' is a sphere far away from C whose function ensure that f integrates to 1.

Let P be the distribution on M whose density over C is λ if $|x_1| > 1/2$, and $\lambda(1-\epsilon)$ if $|x_1| \leq 1/2$, where λ is chosen small enough such that $\lambda \text{vol}_d(C) \leq 1$. The density over C' is chosen such that the total mass of the manifold is 1. Now M_1 and M_3 are (σ, ϵ) separated at level λ for $\sigma = \Omega(1)$. The separator set S is the equator of M_2 in the plane $x_1 = 0$.

We now provide some intuition for why RSL algorithms will require $n \geq 1/(\epsilon\tau)^{\Omega(d)}$ to succeed on this instance. We focus our discussion on RSL algorithms with $k > 2$, i.e. on algorithms that do in fact use a *cleaning* step, ignoring the single linkage algorithm which is known to be inconsistent for full dimensional densities.

Intuitively, because of the curvature of the described instance, the mass of a sufficiently large Euclidean ball in the separator set is *larger* than the mass of a corresponding ball in the true clusters. This means that any algorithm that uses large balls cannot reliably clean the sample and this restricts the size of the balls that can be used. Now if points in the regions of high density are to survive then there must be k sample points in the *small* ball around any point in the true clusters and this gives us a lower bound on the necessary sample size.

The RSL algorithms work by counting the number of sample points inside the balls $B(x, r)$ centered at the sample points x , for some radius r . In order for the algorithm to reliably resolve (σ, ϵ) clusters, it should distinguish points in the separator set $S \subset M_2$ from those in the level λ clusters $M_1 \cup M_3$. A necessary condition for this is that the mass of a ball $B(x, r)$ for $x \in S_{\sigma-r}$ should be strictly smaller than the mass inside $B(y, r)$ for $y \in M_1 \cup M_3$. In Section 8.9.4, we show that this condition restricts the radius r to be at most $O(\tau\sqrt{\epsilon/d})$.

1. For each X_i , $r_k(X_i) := \inf\{r : B(X_i, r) \text{ contains } k \text{ data points}\}$.
2. As r grows from 0 to ∞ :
 - (a) Construct a graph $G_{r,R}$ with nodes $\{X_i : r_k(X_i) \leq r\}$, where r_{X_i} is the V -ball radius of X_i for $V = v_d r^d$, and edges (X_i, X_j) if $\|X_i - X_j\| \leq R$.
 - (b) Let $\mathbb{C}(r)$ be the connected components of $G_{r,R}$.
3. Denote $\widehat{\mathcal{C}} = \{\mathbb{C}(r) : r \in [0, \infty)\}$ and return $\widehat{\mathcal{C}}$.

Figure 8.2: Spatially Adaptive Robust Single Linkage Algorithm

Now, consider any sample point x_0 in $M_1 \cup M_3$ (such an x exists with high probability). Since x_0 should not be removed during the cleaning step, the ball $B(x_0, r)$ must contain some other sample point (indeed, it must contain at least $k - 1$ more sample points). By a union bound, this happens with probability at most

$$(n - 1)v_d r^d \lambda \leq O(d^{-d/2} n \tau^d \epsilon^{d/2} \lambda).$$

If we want the algorithm to succeed with probability at least $1/2$ (say) then

$$n \geq \Omega\left(\frac{d^{d/2}}{\tau^d \lambda \epsilon^{d/2}}\right).$$

8.5 A modified algorithm for the known manifold case

In this section we consider the case when the manifold is *known*. In particular, we assume that we have an oracle that given as input a point $x \in M$ and a number V returns us a radius r_x such that $\text{vol}_d(B_M(x, r_x)) = V$. We call the ball $B(x, r_x)$ the V -ball around x , and the oracle a V -ball oracle.

Given access to the V -ball oracle we show that a modified *spatially adaptive* RSL algorithm achieves the rate

$$n \geq O\left(\frac{1}{\lambda v_d \rho^d \epsilon^2} \log \frac{1}{\lambda v_d \rho^d \epsilon^2}\right)$$

where

$$\rho := \min\left\{\frac{\sigma}{10}, \frac{\tau}{16}\right\}.$$

In particular, ρ no longer depends on $\epsilon\tau$ and for the case of τ fixed (ignoring constants depending on d) the algorithm achieves the near minimax optimal rates of Chaudhuri and Dasgupta [44], in the manifold setting with d replacing D .

The modified algorithm is in Figure 8.2 and it uses two parameters, k and V , to be specified shortly.

We begin with a preliminary lemma which is a straightforward consequence of Lemma 8.3.3.

Lemma 8.5.1. *If $V = v_d r^d$, then $r_l \leq r_x \leq r_u$, where*

$$r_l := r \left(1 - \frac{6r}{\tau} \right) \text{ and } r_u := r \left(1 + \frac{6r}{\tau} \right).$$

Theorem 8.5.2. *There are universal constants C_1 and C_2 such that the following holds. For any $\delta > 0$, $0 < \epsilon < 1/2$, run the algorithm in Figure 8.2 on a sample \mathbf{X} drawn from f , where the parameters are set according to the equations*

$$R = 4r_u = r \left(1 + \frac{6r}{\tau} \right) \text{ and } k = C_1 \log^2(1/\delta)(\mu/\epsilon^2)$$

for r defined by

$$v_d r^d \lambda = \frac{k}{n} + \frac{C_2 \log(1/\delta)}{n} \sqrt{k\mu}.$$

Then with probability at least $1 - \delta$, $\widehat{\mathcal{C}}$ is (σ, ϵ) consistent. In particular, the clusters containing $A[\mathbf{X}]$ and $A'[\mathbf{X}]$, where A and A' are (σ, ϵ) separated, are internally connected and mutually disconnected in $\mathbb{C}(r)$ provided

$$\lambda \geq \frac{2}{v_d \rho^d} \frac{k}{n}.$$

Proof. The theorem is a straightforward consequence of the following lemma.

Lemma 8.5.3 (Separation and Connectedness). *For the parameter choices prescribed in the theorem, provided we satisfy the following*

$$\begin{aligned} 5r_u &\leq \sigma & \text{and} & & R &\leq \tau/2 \\ V\lambda &\geq & \frac{k}{n} &+ & \frac{C_\delta}{n} &\sqrt{k\mu} \\ V\lambda(1 - \epsilon) &\leq & \frac{k}{n} &- & \frac{C_\delta}{n} &\sqrt{k\mu} \end{aligned}$$

the following properties hold w.p. at least $1 - \delta$:

1. All points in $A_{\sigma-r_u}$ and $A'_{\sigma-r_u}$ are kept, and all points in $S_{\sigma-r_u}$ are removed.
2. The two point sets $A[\mathbf{X}]$ and $A'[\mathbf{X}]$ are disconnected in the graph $G_{r,R}$.
3. $A[\mathbf{X}]$ and $A'[\mathbf{X}]$ are internally connected.

Proof. The proof is similar to that of Theorem 8.3.1 and we only highlight the differences.

1. The V -ball around any point x in the manifold has volume *exactly* V by definition, and hence part (1) is true under the good event described in Lemma 8.3.2. In particular notice that using V -balls removes the necessity for estimating the ball volumes.
2. We show part (2) by contradiction. Assume that the graph connects a pair of points from A and A' . Then the connection step guarantees that every edge of the path from A to A' is of Euclidean distance $\leq R \leq \tau/2$, and hence geodesic distance $\leq 2R$. Therefore, by part (1), there must be an edge of (geodesic) length $2(\sigma - r_u)$. This gives us a contradiction, provided $2R \leq 2(\sigma - r_u)$.

3. For part (3) note that $R = 4r_u \geq 4r_x$, and hence an $R/4$ -ball around any net point in $A_{M,R/4}$ contains at least one sample point. The rest of the proof is unchanged. \square

As in the proof of Theorem 8.3.1, we set the parameters according to $k = C_\delta^2(\mu/\epsilon^2)$, and

$$v_d r^d \lambda = \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}.$$

By our assumption on ρ and λ , we can see that $r \leq \rho$, and that

$$r_u = r \left(1 + \frac{6r}{\tau}\right) \leq \rho \left(1 + \frac{6\rho}{\tau}\right) \leq 2\rho.$$

Now, setting $R = 4r_u$, we find that the requirements $R \leq \tau/2$ and $R + r \leq \sigma$ are automatically satisfied. Similarly, the final requirement

$$v_d r^d \lambda (1 - \epsilon) \leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}$$

is also satisfied because of our choices of r and k . \square

8.6 Cluster tree recovery in the presence of noise

So far we have considered the problem of recovering the cluster tree given samples from a density supported *on* a lower dimensional manifold. In this section we extend these results to the more general situation when we have *noisy* samples concentrated *near* a lower dimensional manifold. Indeed it can be argued that the manifold + noise model is a natural and general model for high-dimensional data.

In the noisy setting, it is clear that we can infer the cluster tree of the *noisy* density in a straightforward way. A stronger requirement would be consistency with respect to the underlying *latent* sample. Following the literature on manifold estimation ([21, 81]) we consider two main noise models. For both of them, we specify a distribution Q for the noisy sample.

1. Clutter Noise: We observe data Y_1, \dots, Y_n from the mixture

$$Q := (1 - \pi)U + \pi P$$

where $0 < \pi \leq 1$ and U is a uniform distribution on \mathcal{X} .

Denote the samples drawn from P in this mixture

$$\mathbf{X} = \{X_1, \dots, X_m\}.$$

The points drawn from U are called background clutter. In this case, we can show:

Theorem 8.6.1. *There are universal constants C_1 and C_2 such that the following holds. For any $\delta > 0$, $0 < \epsilon < 1/2$, run the algorithm in Figure 8.1 on a sample $\{Y_1, \dots, Y_n\}$, with parameters*

$$R := 4\rho \quad k := C_1 \log^2(1/\delta)(\mu/\epsilon^2).$$

Then with probability at least $1 - \delta$, $\widehat{\mathcal{C}}$ is (σ, ϵ) consistent. In particular, the clusters containing $A[\mathbf{X}]$ and $A'[\mathbf{X}]$ are internally connected and mutually disconnected in $\mathbb{C}(r)$ for r defined by

$$\pi v_d r^d \lambda = \frac{1}{1 - \epsilon/6} \left(\frac{k}{n} + \frac{C_2 \log(1/\delta)}{n} \sqrt{k\mu} \right)$$

provided

$$\lambda \geq \max \left\{ \frac{2}{v_d \rho^d} \frac{k}{n}, \frac{2v_D^{d/D} (1 - \pi)^{d/D}}{v_d \epsilon^{d/D} \pi} \left(\frac{k}{n} \right)^{1-d/D} \right\}$$

where ρ is now slightly modified (in constants), i.e., $\rho := \min\left(\frac{\sigma}{7}, \frac{\epsilon\tau}{72d}, \frac{\tau}{24}\right)$.

2. Additive Noise: The data are of the form $Y_i = X_i + \eta_i$ where $X_1, \dots, X_n \sim P$, and η_1, \dots, η_n are a sample from any bounded noise distribution Φ , with $\eta_i \in B(0, \theta)$. Note that Q is the convolution of P and Φ , $Q = P \star \Phi$.

Theorem 8.6.2. *There are universal constants C_1 and C_2 such that the following holds. For any $\delta > 0$, $0 < \epsilon < 1/2$, run the algorithm in Figure 8.1 on the sample $\{Y_1, \dots, Y_n\}$ with parameters*

$$R := 5\rho \quad k := C_1 \log^2(1/\delta)(\mu/\epsilon^2).$$

Then with probability at least $1 - \delta$, $\widehat{\mathcal{C}}$ is (σ, ϵ) consistent for $\theta \leq \rho\epsilon/24d$. In particular, the clusters containing $\{Y_i : X_i \in A\}$ and $\{Y_i : X_i \in A'\}$ are internally connected and mutually disconnected in $\mathbb{C}(r)$ for r defined by

$$v_d r^d (1 - \epsilon/12)(1 - \epsilon/6) \lambda = \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}$$

if

$$\lambda \geq \frac{2}{v_d \rho^d} \frac{k}{n}$$

and $\theta \leq \rho\epsilon/24d$, where

$$\rho := \min\left(\frac{\sigma}{7}, \frac{\tau}{24}, \frac{\epsilon\tau}{144d}\right).$$

The proofs for both Theorems 8.6.1 and 8.6.2 appear in Section 8.9.5. Notice that in each case we receive samples from a *full* D -dimensional distribution but are still able to achieve rates independent of D because these distributions are concentrated around the lower dimensional M . For the clutter noise case we produce a tree that is consistent for samples drawn from P (which are *exactly* on M), while in the additive noise case we produce a tree on the observed Y_i s which is (σ, ϵ) consistent for the *latent* X_i s (for θ small enough). It is worth noting that in the case of clutter noise we can still consistently recover the *entire* cluster tree. Intuitively, this is because the k -NN distances for points on M are much smaller than for clutter points that are far away

from M . As a result the clutter noise only affects a vanishingly low level set of the cluster tree. In the case of additive noise with small variance, it is possible to recover well-separated clusters at ambient dimension independent rates. It is also possible to recover the cluster tree in the presence of general additive noise distributions via deconvolution [21, 114] but we do not pursue this approach here.

8.7 Kernel Density Estimators

The results of the previous sections have used k -nearest neighbors based density estimators. However, similar (albeit non-algorithmic) results can be obtained for kernel density estimators.

For the full dimensional cases we consider the usual kernel density estimators

$$\hat{f}_h(x) = \frac{1}{nh^D} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

For the manifold case we consider the following estimator (notice that unlike the usual kernel density estimate it does not integrate to 1),

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

In each case, $K : \mathbb{R}^D \rightarrow \mathbb{R}$ is a kernel. In each case, there is an associated population quantity that will be useful. In the full dimensional case

$$f_h(x) = \frac{1}{h^D} \mathbb{E}_{X \sim f} K\left(\frac{x - X}{h}\right)$$

and in the manifold case

$$f_h(x) = \frac{1}{h^d} \mathbb{E}_{X \sim f} K\left(\frac{x - X}{h}\right).$$

As before $\mathcal{C}(\hat{f}_h)$ denotes the cluster tree of the kernel density estimate.

8.7.1 Assumptions and preliminaries

We will make one of the following assumptions on the kernel:

Assumption 5 (Bounded support).

[5A] *For the case of full-dimensional densities we will assume the kernel has bounded support and integrates to 1, i.e.*

$$\{x : K(x) > 0\} \subseteq B(0, 1)$$

and

$$\int_{x \in \mathbb{R}^D} K(\|x\|) = 1.$$

Following Giné and Guillou [83], we will further assume that the class of functions

$$\mathcal{F} = \left\{ K \left(\frac{x - \circ}{h} \right), x \in \mathbb{R}^D, h > 0 \right\}$$

satisfies, for some positive number A and v

$$\sup_P \mathcal{N}(\mathcal{F}_h, L_2(P), \epsilon \|F\|_{L_2(P)}) \leq \left(\frac{A}{\epsilon} \right)^v$$

where $\mathcal{N}(T, d, \epsilon)$ denotes the ϵ -covering number of the metric space (T, d) , F is the envelope function of \mathcal{F} and the supremum is taken over the set of all probability measures on \mathbb{R}^D . A and v are called the VC characteristics of the kernel.

[5B] For the case of densities supported on lower-dimensional manifolds we will assume a particular form for the kernel

$$K(x) = \frac{\mathbb{I}(x \leq 1)}{v_d}.$$

Observe that this kernel also satisfies the VC assumption above.

The first assumption is quite mild and can be further relaxed to include kernels with an appropriate tail decay, albeit at the cost of more complicated proofs. The second assumption allows us to avoid dealing with integrals over the manifold but can also be similarly relaxed.

Assumption 6 (Bandwidth regularity: BR(m)). For some $c > 0$,

$$h_n \searrow 0, \quad \frac{nh_n^m}{|\log h_n|} \rightarrow \infty \quad \frac{|\log h_n|}{\log \log n} \rightarrow \infty \quad \text{and} \quad h_n^m \leq ch_{2n}^m.$$

We will first state two preliminary results showing the uniform consistency of the kernel density estimate.

The first Lemma appears in a similar form in the paper of Rinaldo and Wasserman [161] (Proposition 9) and is a modification of a result of Giné and Guillou [83] (Corollary 2.2). The proof is omitted.

Lemma 8.7.1 (Full dimensional density). Given n samples from a distribution which has a bounded density f with respect to the Lebesgue measure on \mathbb{R}^D

1. For $n \geq n_0$, where n_0 is a constant depending only on the VC characteristics of K , $\|K\|_\infty$, $\|K\|_2$ and f_{\max} , and fixed $h \leq h_0$ depending only on $\|K\|_\infty$ and f_{\max} there is a constant C depending on K such that

$$P \left(\|\hat{f}_h - f_h\|_\infty \geq C' \cdot C \sqrt{\frac{f_{\max} \log(1/h)}{nh^D}} \right) \leq \left(\frac{1}{h} \right)^{C'}$$

for any large enough constant C' depending on K and f_{\max} of our choice.

2. For any sequence $h_n \leq h_0$ as before, satisfying Assumption 6, $BR(D)$, for all $n \geq n_0$ as before

$$P \left(\|\hat{f}_{h_n} - f_{h_n}\|_\infty \geq C' \cdot C \sqrt{\frac{f_{\max} \log(1/h_n)}{nh_n^D}} \right) \leq \left(\frac{1}{h}\right)^{C'}.$$

For the ball kernel of Assumption 5 a similar result holds for densities supported on a lower dimensional manifold.

Lemma 8.7.2 (Manifold case). *Given n samples from a distribution supported on a smooth Riemannian manifold M with condition number at most $1/\tau$ with bounded density f with respect to the uniform measure on M*

1. For $n \geq n_0$, where n_0 is a constant depending only on the VC characteristics of K , $\|K\|_\infty$, $\|K\|_2$ and $\|f\|_\infty$, and fixed $h \leq \min(\frac{\tau}{8}, h_0)$ where h_0 depends only on $\|K\|_\infty$ and $\|f\|_\infty$ there is a constant C_δ depending on δ and n_0 such that

$$P \left(\|\hat{f}_h - f_h\|_\infty \geq C' \cdot C \sqrt{\frac{f_{\max} \log(1/h)}{nh^d}} \right) \leq \left(\frac{1}{h}\right)^{C'}.$$

2. For any sequence $h_n \leq \min(\frac{\tau}{8}, h_0)$ as before, satisfying Assumption 6, $BR(d)$, for all $n \geq n_0$ as before

$$P \left(\|\hat{f}_{h_n} - f_{h_n}\|_\infty \geq C' \cdot C \sqrt{\frac{f_{\max} \log(1/h_n)}{nh_n^d}} \right) \leq \left(\frac{1}{h}\right)^{C'}.$$

Proof. The proof follows along the lines of those in the papers of Giné and Guillou [83], Rinaldo and Wasserman [161]. The main modification to achieve d rates involves a more careful calculation of the variance.

To apply Talagrand's inequality in the proof of Giné and Guillou [83] we need to bound

$$\sup_{g \in \mathcal{F}} \text{Var}_f g.$$

\mathcal{F} is the set of kernel functions with various bandwidths, and centers anywhere on M .

Let us show how to bound $\sup_{g \in \mathcal{F}_h} \text{Var}_f g$ for a single bandwidth h .

$$\begin{aligned} \text{Var}_{X \sim p} \left(K \left(\frac{x - X}{h} \right) \right) &= \mathbb{E}_X \left[K \left(\frac{x - X}{h} \right) - \mathbb{E}_X K \left(\frac{x - X}{h} \right) \right]^2 \\ &\leq \left[\mathbb{E}_X K^2 \left(\frac{x - X}{h} \right) \right] \\ &= \int_{X \in M} K^2 \left(\frac{x - X}{h} \right) f(X) dX \\ &\leq \|K\|_\infty^2 \int I(X \in B(x, h)) f(X) dX \\ &\leq h^d C_d \|K\|_\infty^2 \|f\|_\infty. \end{aligned}$$

The last step follows if $h \leq \frac{\tau}{8}$ by the ball volume Lemma 8.3.3. Notice that the variance does not depend on x and so the bound holds uniformly over all x on M .

Replacing this bound on the variance in the proof of Giné and Guillaou [83] we obtain the desired result. \square

8.7.2 Rates of convergence for the cluster tree

Our first result mirrors the main result of Chaudhuri and Dasgupta [44].

Theorem 8.7.3 (Full dimensional cluster tree). *There is a constant C_δ depending on the VC characteristics of the kernel, $\|K\|_\infty, \|K\|_2, \|f\|_\infty$ and δ such that the following holds with probability at least $1 - \delta$, $\mathcal{C}(\hat{p}_\sigma)$ is (σ, ϵ) consistent for any pair of clusters A, A' at level at least λ for*

$$n \geq \frac{C_\delta}{\sigma^D \lambda^2 \epsilon^2} \log \left(\frac{1}{\sigma} \right).$$

Notice, in particular that while for the k -nearest neighbor based algorithm the choice of k depends on ϵ for the kernel density estimate the optimal choice of bandwidth depends on σ . Also notice unlike the result of Chaudhuri and Dasgupta [44] this result requires the density to be uniformly upper bounded.

Proof. To prove this theorem it suffices to show that the regions A and A' are internally connected and mutually separated.

Let us first show that σ -clusters A and A' (for any $\lambda, \epsilon > 0$) are connected and separated in $\mathcal{C}(f_\sigma)$. Consider any point $x \in A \cup A'$,

$$f_\sigma(x) = \int_{y \in B(x, \sigma)} K \left(\frac{y-x}{h} \right) f(y) dy \geq \lambda \int_{y \in B(x, \sigma)} K \left(\frac{y-x}{h} \right) dy \geq \lambda.$$

Similarly, we can see that for any point in the separator S , $f_\sigma(x) < \lambda(1 - \epsilon)$. In particular, σ -clusters A and A' are distinguished in $\mathcal{C}(f_\sigma)$ at level λ as desired.

Now, we use Lemma 8.7.1. Notice for a constant C_δ

$$n \geq \frac{C_\delta}{\sigma^D \lambda^2 \epsilon^2} \log \left(\frac{1}{\sigma} \right)$$

we have

$$\|\hat{f}_\sigma - f_\sigma\|_\infty \leq \frac{\lambda \epsilon}{2}$$

with probability $1 - \delta$. Let \mathcal{E}_1 denote the event $\{\|\hat{f}_\sigma - f_\sigma\|_\infty \leq \frac{\lambda \epsilon}{2}\}$.

Now, let us consider the cluster tree of \hat{f}_σ at level $\lambda - \frac{\lambda \epsilon}{2}$. On \mathcal{E}_1 , for any point $x \in A \cup A'$ we know $f_\sigma \geq \lambda$ and thus $\hat{f}_\sigma \geq \lambda - \frac{\lambda \epsilon}{2}$. Similarly for $x \in S$ we have $\hat{f}_\sigma < \lambda - \frac{\lambda \epsilon}{2}$. These together show that on \mathcal{E}_1 A and A' are distinguished in $\mathcal{C}(\hat{f}_\sigma)$ at level $\lambda - \frac{\lambda \epsilon}{2}$. This establishes the theorem. \square

To establish Hartigan consistency we select a schedule h_n satisfying Assumption 6. Under mild conditions connected components of any level set at λ , are (σ, ϵ) separated for some $\sigma, \epsilon > 0$ and are distinguished for n large enough.

We can similarly give a manifold version of this result. Define

$$\rho = \min \left(\sigma, \frac{\tau}{8}, \frac{\epsilon\tau}{72d} \right).$$

Theorem 8.7.4 (Cluster tree on manifolds). *There is a constant C_δ depending on the VC characteristics of the kernel, $\|K\|_\infty, \|K\|_2, \|f\|_\infty$ and δ such that the following holds with probability at least $1 - \delta$, for all $\epsilon \leq 1/2$ $\mathcal{C}(\hat{p}_\rho)$ is (σ, ϵ) consistent for any pair of clusters A, A' at level at least λ for*

$$n \geq \frac{C_\delta}{\rho^D \lambda^2 \epsilon^2} \log \left(\frac{1}{\rho} \right).$$

Proof. Let us again consider f_ρ . For any point $x \in A \cup A'$,

$$f_\rho(x) = \frac{1}{h^d} \mathbb{E}_{X \sim f} K \left(\frac{x - X}{h} \right) = \frac{1}{v_d \rho^d} \int_{X \in B_M(x, h)} dX \geq \lambda \left(1 - \frac{\epsilon}{6} \right)$$

where the second equality follows from the assumed form of the kernel, and the inequality follows from Lemma 8.3.3 under the assumption on ρ . Similarly, for any point in S we have

$$f_\rho(x) < \lambda (1 - \epsilon) \left(1 + \frac{\epsilon}{6} \right).$$

The gap between these is at least $\lambda\epsilon/2$, and hence A and $'$ are distinguished in f_ρ at level $\lambda(1 - \epsilon/6)$.

The proof that these clusters are distinguished in \hat{f}_ρ follows from an identical argument to the one in the proof of Theorem 8.7.3, replacing the use of Lemma 8.7.1 with Lemma 8.7.2. \square

8.8 Simulations

Figure 8.3 depicts the results of simulations we performed to test our main theoretical predictions. For Figure 8.3(B) we sample data from a mixture distribution on a unit d -sphere. The mixture has 10 salient clusters (with a total mixture weight of 0.7) mixed with uniform samples on the sphere with mixture weight 0.3. Finally, we mix samples from this density with D -dimensional clutter noise with $\pi = 0.8$. A sample is shown in Figure 8.3(A) for $d = 2, D = 3$ and $n = 1000$. For Figures 8.3(C)-(H) we simulate data from the lower bound instance described in Section 8.4.

In Figure 8.3(B), we plot the probability of successfully recovering the 10 clusters in the cluster tree as a function of sample size. The figure confirms that the sample size is independent of the ambient dimension D but (typically) gets worse with the manifold dimension d . In particular,

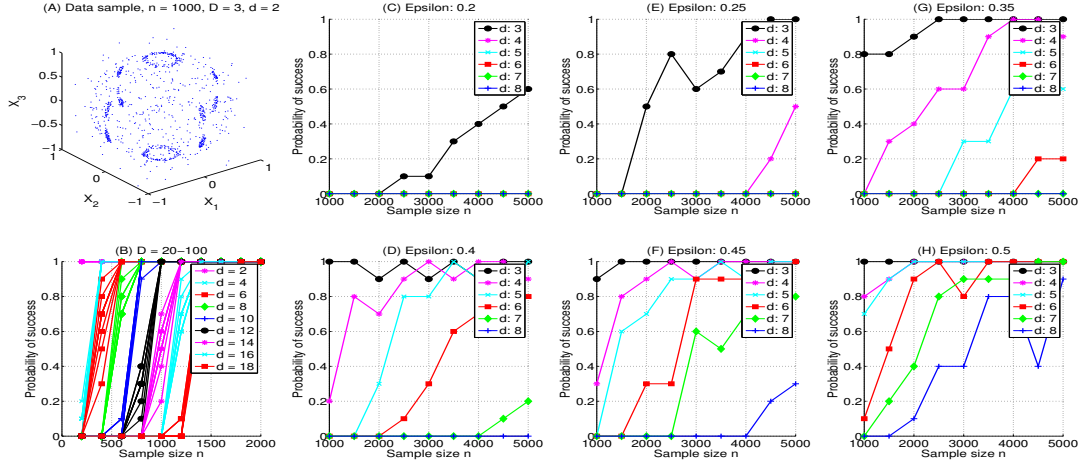


Figure 8.3: Figures show the average probability of success across 10 trials for different (n, d, D, ϵ) .

the figure shows that for $D = \{20, 40, 60, 80, 100\}$ (in the same color) sample complexities are nearly unchanged. Figures 8.3(C)-(H), shows the effect on sample size of (ϵ, d) for the lower bound instance. Notice, that for a fixed ϵ and n the probability of success decays rapidly with increasing d and that for a fixed d and n the probability of success grows with ϵ , in agreement with our $1/\epsilon^{\Omega(d)}$ prediction and in contrast to the $1/\epsilon^2$ scaling predicted by Chaudhuri and Dasgupta [44] for recovering a full-dimensional cluster tree.

8.9 Additional proofs

In this section we first prove some technical lemmas before giving full proofs of various claims made in the chapter.

8.9.1 Volume estimates for small balls on manifolds

Theorem 8.9.1. *If*

$$r \leq \frac{\epsilon\tau}{12d}$$

for $0 \leq \epsilon < 1$ then

$$v_d r^d (1 - \epsilon) \leq \text{vol}(S) \leq v_d r^d (1 + \epsilon).$$

Proof. The lower bound follows from Niyogi et al. [146] (Lemma 5.3) who show that

Lemma 8.9.2. *For $r < \frac{\tau}{2}$*

$$\text{vol}(S) \geq \left(1 - \frac{r^2}{4\tau^2}\right)^{d/2} v_d r^d.$$

The upper bound follows from Chazal [46] who shows that

Lemma 8.9.3. For $r < \frac{\tau}{2}$

$$\text{vol}(S) \leq v_d \left(\frac{\tau}{\tau - 2\alpha} \right)^d \alpha^d$$

where

$$\alpha = \tau - \tau \sqrt{1 - \frac{2r}{\tau}}.$$

To produce the result of the theorem we will need some careful manipulation of these two lemmas. In particular, we need the following estimates

Lemma 8.9.4.

$$f(x) = (1 - x)^{1/2} \geq 1 - \frac{x}{2} - x^2$$

if $0 \leq x \leq \frac{1}{2}$.

$$f(x) = (1 + x)^n \leq 1 + 2nx$$

if $0 \leq x \leq \frac{1}{2n}$.

$$f(x) = (1 - x)^{-1} \leq 1 + 2x$$

if $0 \leq x \leq 1/2$.

$$f(x) = (1 - x)^n \geq 1 - 2nx$$

if $0 \leq x \leq \frac{1}{2n}$.

The proof of this lemma is straightforward based on approximations via Taylor's series and we omit them.

Using Lemma 8.9.4 we have

$$\alpha \leq r \left(1 + \frac{4r}{\tau} \right)$$

if $r \leq \frac{\tau}{4}$. Now, using this also notice that

$$\frac{\tau}{\tau - 2\alpha} \leq \frac{1}{1 - \frac{2r}{\tau} \left(1 + \frac{4r}{\tau} \right)} \leq 1 + \frac{4r}{\tau} \left(1 + \frac{4r}{\tau} \right)$$

where the second inequality follows from Lemma 8.9.4 if $r \leq \tau/8$.

Combining these we have the following:

for all $r \leq \frac{\tau}{8}$

$$v_d r^d \left(1 - \frac{r^2}{4\tau^2} \right)^{d/2} \leq \text{vol}(S) \leq v_d r^d \left(1 + \frac{6r}{\tau} \right)^d$$

The final result now follows another application of Lemma 8.9.4 on each side of this inequality. \square

8.9.2 Bound on covering number

We need the following bound on the covering number of a manifold. See the paper [146] (p. 16) for a proof.

Lemma 8.9.5. *For $s \leq 2\tau$, the s -covering number of M is at most*

$$\frac{\text{vol}_d(M)}{\cos^d(\arcsin(s/4\tau))v_d(s/2)^d} \leq O\left(\frac{\text{vol}_d(M)c^d}{v_d s^d}\right)$$

for an absolute constant c . In particular, if $\text{vol}_d(M)$ is bounded above by a constant, the s -covering number of M is at most $O(c^d/(v_d s^d))$.

Proof. We prove only the second claim. For $s \leq 2\tau$, we have $\arcsin(s/4\tau) \leq \pi/6$, and hence $\cos(\arcsin(s/4\tau)) \geq \sqrt{3}/2$. Plugging this in the bound, we get

$$|\mathcal{N}| \leq \frac{\text{vol}_d(M)(2/\sqrt{3})^d}{v_d(s/2)^d},$$

which gives the claim with $c = 4/\sqrt{3}$. □

8.9.3 Uniform convergence

In this subsection, we prove uniform convergence for balls centered on sample and net points (Lemma 8.3.2). Consider the family of balls centered at a fixed point z , $\mathcal{B}_z := \{B(z, s) : s \geq 0\}$. This collection has VC dimension 1. Thus with probability $1 - \delta'$, it holds that for every $B \in \mathcal{B}_z$, we have

$$\max\left\{\frac{P(B) - P_n(B)}{\sqrt{P(B)}}, \frac{P(B) - P_n(B)}{\sqrt{P_n(B)}}\right\} \leq 2\sqrt{\frac{\log(2n) + \log(4/\delta')}{n}},$$

where $P(B)$ is the true mass of B , and $P_n(B) = |\mathbf{X} \cap B|/n$ is its empirical measure. By a union bound over all $z \in \mathcal{N}$, setting $\delta' := \delta/(2|\mathcal{N}|)$, the following holds uniformly for every $z \in \mathcal{N}$ and every $B \in \mathcal{B}_z$ with probability $1 - \delta/2$:

$$\max\left\{\frac{P(B) - P_n(B)}{\sqrt{P(B)}}, \frac{P(B) - P_n(B)}{\sqrt{P_n(B)}}\right\} \leq 2\sqrt{\frac{\log(2n) + \log(8|\mathcal{N}|/\delta)}{n}}.$$

To provide a similar uniform convergence result for balls centered at a sample point X_i , we consider the $(n-1)$ -subsample X_i^{n-1} of \mathbf{X} obtained by deleting X_i from the sample. Let P_i^{n-1} be the empirical probability measure of this subsample:

$$P_{n-1}(B) := \frac{1}{n-1} \sum_{j \neq i} \mathbb{I}[X_j \in B].$$

It is easy to check that P_{n-1} is uniformly close to P_n . In particular, for every set B containing X_i , we have

$$P_{n-1}(B) \leq P_n(B) \leq P_{n-1}(B) + \frac{1}{n}. \quad (8.2)$$

Now, with probability at least $1 - \delta/(2n)$, for any ball B centered at X_i ,

$$\begin{aligned} P(B) - P_{n-1}(B) &\leq 2\sqrt{\frac{\log(2n-2) + \log 8n/\delta}{n-1}} \cdot \sqrt{P(B)}, \\ P_{n-1}(B) - P(B) &\leq 2\sqrt{\frac{\log(2n-2) + \log 8n/\delta}{n-1}} \cdot \sqrt{P_{n-1}(B)}. \end{aligned}$$

Using (8.2), we get

$$\begin{aligned} P(B) - P_n(B) &\leq 2\sqrt{\frac{\log(2n-2) + \log 8n/\delta}{n-1}} \cdot \sqrt{P(B)}, \\ P_n(B) - P(B) &\leq 2\sqrt{\frac{\log(2n-2) + \log 8n/\delta}{n-1}} \cdot \sqrt{P_n(B)} + \frac{1}{n}. \end{aligned}$$

By a union bound over all $X_i \in \mathbf{X}$, we get the claimed inequalities for all sample points with probability $1 - \delta/2$.

Putting together our bounds for balls around sample and net points, with probability at least $1 - \delta$, it holds that for all $B \in \mathcal{B}_{n,\mathcal{N}}$, we have

$$\begin{aligned} P(B) - P_n(B) &\leq O\left(\sqrt{\frac{\mu + \log(1/\delta)}{n}}\right) \cdot \sqrt{P(B)}, \\ P_n(B) - P(B) &\leq O\left(\sqrt{\frac{\mu + \log(1/\delta)}{n}}\right) \cdot \sqrt{P_n(B)} + \frac{1}{n}. \end{aligned}$$

for $\mu = 1 + \log n + \log |\mathcal{N}| = O(d) + \log n + d \log(1/s)$ (using Lemma 8.9.5). The lemma now follows using simple manipulations of these inequalities (see [44] for details).

8.9.4 Sketch of the lower bound instance

The following lemma gives an estimate of the volume of the intersection of a small ball with a sphere.

Lemma 8.9.6 (Volume of a spherical cap). *Suppose \mathbb{S}^d is a d -dimensional sphere of radius τ (embedded in \mathbb{R}^{d+1}), and let $x \in \mathbb{S}^d$. Then, for small enough r , it holds that*

$$\text{vol}_d(B(x, r) \cap \mathbb{S}^d) = v_d r^d \left(1 - c_d \frac{r^2}{\tau^2} + O_d\left(\frac{r^4}{\tau^4}\right) \right)$$

where $c_d := \frac{d(d-2)}{8(d+2)}$. Note that $c_1 < 0$, $c_2 = 0$, and $c_d > 0$ for all $d \geq 3$.

In this section, we prove Lemma 8.9.6. The height h of the cap can be easily checked to be equal to $h = r^2/2\tau$. Now, the volume of the cap is given by the formula

$$v_{cap} = \frac{\pi^{(d+1)/2}\tau^d}{\Gamma((d+1)/2)} I_\alpha(d/2, 1/2)$$

where the parameter α is defined by

$$\alpha := \frac{2\tau h - h^2}{\tau} = \frac{r^2}{\tau^2} \left(1 - \frac{r^2}{4\tau^2}\right).$$

Further $I_\alpha(\cdot, \cdot)$ represents the incomplete beta function:

$$\begin{aligned} I_\alpha(z, w) &= \frac{B(\alpha; z, w)}{B(z, w)} \\ &= \frac{\int_0^\alpha u^{z-1}(1-u)^{w-1} du}{B(z, w)} \\ &= \frac{\Gamma(z+w)}{\Gamma(z)\Gamma(w)} \int_0^\alpha u^{z-1}(1-u)^{w-1} du. \end{aligned}$$

Thus,

$$\begin{aligned} v_{cap} &= \frac{\pi^{(d+1)/2}\tau^d}{\Gamma((d+1)/2)} \cdot \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\Gamma(1/2)} \cdot \int_0^\alpha u^{d/2-1}(1-u)^{-1/2} du \\ &= \frac{\pi^{d/2}\tau^d}{\Gamma(d/2)} \int_0^\alpha u^{d/2-1}(1-u)^{-1/2} du \\ &= \frac{dv_d\tau^d}{2} \int_0^\alpha u^{d/2-1}(1-u)^{-1/2} du. \end{aligned}$$

Since $\alpha \rightarrow 0$ as $r \rightarrow 0$, we can approximate the integral by expanding the integrand as a Taylor series around 0:

$$\begin{aligned} v_{cap} &= \frac{dv_d\tau^d}{2} \int_0^\alpha u^{d/2-1} \left(1 + u/2 + O(u^2)\right) du \\ &= \frac{dv_d\tau^d}{2} \left(\frac{\alpha^{d/2}}{d/2} + \frac{1}{2} \frac{\alpha^{d/2+1}}{d/2+1} + O(\alpha^{d/2+2}) \right) \\ &= v_d\tau^d \alpha^{d/2} \left(1 + \frac{d}{2(d+2)}\alpha + O(\alpha^2) \right). \end{aligned}$$

Finally, using $\alpha := \frac{r^2}{\tau^2} \left(1 - \frac{r^2}{4\tau^2}\right)$, we get

$$\begin{aligned} v_{cap} &= v_d r^d \left(1 - \frac{r^2}{4\tau^2}\right)^{d/2} \left(1 + \frac{dr^2}{2(d+2)\tau^2} + O\left(\frac{r^4}{\tau^4}\right)\right) \\ &= v_d r^d \cdot \left(1 - \frac{dr^2}{8\tau^2} + \frac{dr^2}{2(d+2)\tau^2} + O_d\left(\frac{r^4}{\tau^4}\right)\right), \end{aligned}$$

which simplifies to the claimed estimate.

We now show that it must be the case that $r \leq O(\tau\sqrt{\epsilon/d})$. We argued that for the algorithm to reliably resolve the (σ, ϵ) separated clusters M_1 and M_3 , an r -ball around a sample point in $S_{\sigma-r}$ must have mass appreciably smaller than those around points in M_1 . By the previous lemma, the two kinds of balls have volumes

$$v_d r^d \left(1 - c_d \frac{r^2}{1^2} + O_d\left(\frac{r^4}{1^4}\right) \right) = v_d r^d (1 - c_d r^2 + O_d(r^4))$$

and

$$v_d r^d \left(1 - c_d \frac{r^2}{4\tau^2} + O_d\left(\frac{r^4}{16\tau^4}\right) \right) = v_d r^d \left(1 - c_d \frac{r^2}{4\tau^2} + O_d\left(\frac{r^4}{\tau^4}\right) \right).$$

Thus we must have

$$v_d r^d v_d r^d (1 - c_d r^2 + O_d(r^4)) \cdot \lambda(1 - \epsilon) \leq v_d r^d \left(1 - c_d \frac{r^2}{4\tau^2} + O_d\left(\frac{r^4}{\tau^4}\right) \right) \cdot \lambda.$$

This implies that $r^2 \leq O\left(\frac{4\tau^2\epsilon}{(1-4\tau^2)c_d}\right)$. Hence if $\tau \leq 1/4$, we have $r \leq \tau\sqrt{\epsilon/c_d}$. Plugging in $c_d = \Omega(d)$ gives us the claim.

8.9.5 Clustering with noisy samples

8.9.6 Proof of Theorem 8.6.1

As before we begin by showing separation followed by a proof of connectivity. Recall that $\rho := \min\left(\frac{\sigma}{7}, \frac{\epsilon\tau}{72d}, \frac{\tau}{24}\right)$.

Lemma 8.9.7 (Separation). *Assume that we pick k , r and R to satisfy the conditions:*

$$\begin{aligned} r &\leq \rho, & R &= 4\rho \\ \pi \cdot v_d r^d (1 - \epsilon/6) \cdot \lambda &\geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}, \\ \pi \cdot v_d r^d (1 + \epsilon/6) \cdot \lambda(1 - \epsilon) + (1 - \pi) \cdot v_D r^D &\leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}. \end{aligned}$$

Then with probability $1 - \delta$, it holds that:

1. All points in $A_{M, \sigma-r}$ and $A'_{M, \sigma-r}$ are kept, and all points in $\mathcal{X} \setminus M_r$ and $S_{\sigma-r}$ are removed. Here, M_r is the tubular region around M of width r .
2. The two point sets $A[\mathbf{X}]$ and $A'[\mathbf{X}]$ are disconnected in the graph $G_{r,R}$.

Proof. The proof of the first claim is similar to the noiseless setting, except that the probability mass inside a ball now has contributions from both the manifold and the background clutter. For $x \in S_{\sigma-r}$, the probability mass of the ball $B(x, r)$ under Q is at most $\pi v_d r^d (1 + \epsilon/6) \cdot \lambda(1 -$

$\epsilon) + (1 - \pi)v_D r^D$, which is at most $\frac{k}{n} - \frac{C_\delta}{n}\sqrt{k\mu}$. Thus x is removed during the cleaning step. Similarly, if $x \notin M_r$, the ball $B(x, r)$ does not intersect the manifold, and hence its mass is at most $(1 - \pi)v_D r^D$. Hence all points outside M_r are removed. Finally, if $x \in (A_{M, \sigma-r} \cup A'_{M, \sigma-r}) \cap \mathbf{X}$, then the mass of the ball $B_M(x, r)$ is at least $v_d r^d (1 - \epsilon/6)\lambda$ (ignoring the contribution of the noise). This is at least $\frac{k}{n} + \frac{C_\delta}{n}\sqrt{k\mu}$, and hence x is kept.

To prove the second claim, suppose that sets $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ are connected in $G_{r, R}$. Then there exists a sequence of sample points y_0, y_1, \dots, y_t such that $y_0 \in A$, $y_t \in A'$ and $d(y_{i-1}, y_i) \leq R$ for all $1 \leq i \leq t$. Let x_i be the projection of y_i on M , i.e., x_i is the point of M closest to y_i . We have already showed that each y_i lies inside the tube M_r , so $d(x_i, y_i) \leq r$, and hence by triangle inequality, we have $d(x_{i-1}, x_i) \leq R + 2r \leq \tau/4$. Hence, the geodesic distance between x_{i-1} and x_i is $< 2(R + 2r)$. Now, by an argument analogous to the noiseless setting, there exists a pair (x_{i-1}, x_i) which are at a (geodesic) distance at least $2(\sigma - r)$. This is a contradiction since our parameter setting implies that $2(\sigma - r) \geq 2(R + 2r)$. \square

Lemma 8.9.8 (Connectedness). *Assume that the parameters k, r and R satisfy the separation conditions (in Lemma 8.9.7). Then, with probability at least $1 - \delta$, $A \cap \mathbf{Y}$ is connected in $G_{r, R}$.*

Proof. The proof of this lemma is identical to Lemma 8.3.5 and is omitted. \square

We now show how to pick the parameters to satisfy the conditions in Lemma 8.9.7. Set $k := 144C_\delta^2(\mu/\epsilon^2)$, and define r by

$$\pi v_d r^d (1 - \epsilon/6) \cdot \lambda = \frac{k}{n} + \frac{C_\delta}{n}\sqrt{k\mu}.$$

It is easy to check that this setting satisfies all our requirements, provided that the term $(1 - \pi)v_D r^D$ arising from the clutter noise satisfies the additional constraint

$$(1 - \pi)v_D r^D \leq (\epsilon/2) \times \pi v_d r^d \lambda.$$

The definition of r implies that r is upper bounded by $\left(\frac{2k}{n\lambda\pi v_d}\right)^{1/d}$. Thus it suffices to ensure that

$$(1 - \pi)v_D \left(\frac{2k}{n\lambda\pi v_d}\right)^{D/d} \leq (\epsilon/2) \cdot \frac{2k}{n} = \frac{k\epsilon}{n}.$$

This is equivalent to the condition

$$\lambda \geq \frac{2v_D^{d/D}}{v_d \epsilon^{d/D}} \cdot \frac{(1 - \pi)^{d/D}}{\pi} \cdot \left(\frac{k}{n}\right)^{1-d/D},$$

which is assumed by Theorem 8.6.1.

8.9.7 Proof of Theorem 8.6.2

Let P be a distribution on a manifold M with density f . Let $\mathbf{X} = (X_1, \dots, X_n)$ be the latent sample from P , and let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be the observed sample. The only fact that we use about the observed sample is that it is close to the corresponding latent sample point: $d(Y_i, X_i) \leq \theta$, where θ is the *noise radius*. We show that we can adapt the RSL algorithm to resolve (σ, ϵ) separated clusters (A, A') , provided that θ is sufficiently small compared to both σ and ϵ .

Again, we will pick values for k, r, R based on a parameter ρ , defined as $\rho := \min(\frac{\sigma}{7}, \frac{\tau}{24}, \frac{\epsilon\tau}{144d})$.

Lemma 8.9.9 (Separation). *Suppose k, r, R are chosen to satisfy*

$$\begin{aligned} \theta &\leq r/2 & r &\leq \rho & R &:= 5\rho, \\ v_d(r - 2\theta)^d(1 - \epsilon/6) \cdot \lambda &\geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}, \\ v_d(r + 2\theta)^d(1 + \epsilon/6) \cdot \lambda(1 - \epsilon) &\leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}, \end{aligned}$$

then, with probability $1 - \delta$, the following holds uniformly over all (σ, ϵ) separated clusters (A, A') :

1. *If a latent sample point $X_i \in A_{M, \sigma-r+2\theta} \cup A'_{M, \sigma-r+2\theta}$, then the corresponding sample point Y_i is kept during the cleaning step. If $X_i \in S_{M, \sigma-r-2\theta}$, then Y_i is removed.*
2. *The sets $\{Y_i : X_i \in A\}$ and $\{Y_i : X_i \in A'\}$ are disconnected in the graph $G_{r,R}$.*

Proof. To prove the first claim, suppose $X_i \in A_{\sigma-r+2\theta} \cup A'_{\sigma-r+2\theta}$. Consider the ball $B_M(X_i, r - 2\theta)$. It is completely inside $A_{M, \sigma} \cup A'_{M, \sigma}$, hence the density f inside it is at least λ . Moreover, if X_j is in $B_M(X_i, r - 2\theta)$, then by triangle inequality, we have

$$d(Y_j, Y_i) \leq d(X_j, Y_j) + d(X_j, X_i) + d(Y_i, X_i) \leq r.$$

Hence the ball $B(X_i, r)$ contains at least k sample points, provided $B_M(X_i, r - 2\theta)$ contains at least k points from \mathbf{X} . Finally, the true mass of the set $B_M(X_i, r - 2\theta)$ is at least

$$v_d(r - 2\theta)^d(1 - \epsilon/6) \cdot \lambda \geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}.$$

Hence it contains at least k latent sample points, and we are done.

Similarly, suppose $X_i \in S_{\sigma-r-2\theta}$, and consider the ball $B_M(X_i, r + 2\theta)$. It is completely contained inside $S_{M, \sigma}$ and hence the density inside the ball is at most $\lambda(1 - \epsilon)$. Moreover, if X_j is outside the set, then

$$d(Y_j, Y_i) \geq d(X_j, X_j) - d(X_i, Y_i) - d(X_j, Y_j) > r.$$

Hence the ball $B(Y_i, r)$ contains fewer than k sample points, provided $B_M(X_i, r + 2\theta)$ contains fewer than k points from \mathbf{X} . The true mass of the ball $B_M(X_i, r + 2\theta)$ is at most

$$v_d(r + 2\theta)^d(1 + \epsilon/6) \cdot \lambda(1 - \epsilon) \leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu}.$$

Hence the ball contains fewer than k latent sample points, and we are done.

We now prove that the graph $G_{r,R}$ is disconnected. Suppose not. Then there must exist a sequence of latent sample points $x_0, x_1, \dots, x_t \in \mathbf{Y}$ and a corresponding sequence of noisy sample points $y_0, \dots, y_t \in \mathbf{X}$ such that $x_0 \in A$, $x_t \in A'$, and $d(y_{i-1}, y_i) \leq R$. Clearly $d(x_{i-1}, x_i) \leq R + 2\theta \leq \tau/4$. Thus the geodesic distance between x_{i-1} and x_i is less than $2(R + 2\theta)$. However, by the (σ, ϵ) separation condition, we must have a successive pair (x_{i-1}, x_i) whose geodesic distance is at least $2(\sigma - r)$. This is a contradiction since we have set our parameters such that $2(\sigma - r) \geq 2(R + 2\theta)$. \square

Lemma 8.9.10 (Connectedness). *Assume that the conditions of Lemma 8.9.9 are satisfied. Then, with probability at least $1 - \delta$, the following holds uniformly over all A : if $\inf_{x \in A_{M,\sigma}} f(x) \geq \lambda$, then $\{Y_i : X_i \in A\}$ is connected in $G_{r,R}$.*

Proof. The proof is similar to that of Lemma 8.3.5, so we indicate only the necessary modifications, omitting the details. We now use a net of radius $(R - 2\theta)/4$, and the condition that $R \geq 4r$ is replaced by $R - 2\theta \geq 4r$. Finally, the x_i 's defined in the proof are latent sample points, whereas the algorithm observes an arbitrary point y_i in a θ -ball around the x_i . Thus, the distance between y_{i-1} and y_i is at most

$$4 \cdot \frac{R - 2\theta}{4} + d(y_i, x_i) + d(y_{i-1}, x_{i-1}) \leq R.$$

\square

In order to satisfy the conditions stated in Lemma 8.9.9, we need the assumption that θ is small compared to r . More precisely, we will assume that $\theta \leq r\epsilon/24d$. Under this assumption, we can satisfy the above conditions by ensuring that

$$\begin{aligned} v_d r^d (1 - \epsilon/12)(1 - \epsilon/6) \cdot \lambda &\geq \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}, \\ v_d r^d (1 + \epsilon/6)(1 + \epsilon/6) \cdot \lambda(1 - \epsilon) &\leq \frac{k}{n} - \frac{C_\delta}{n} \sqrt{k\mu} \end{aligned}$$

As before, we can satisfy these equations by setting $k := O(C_\delta^2 \mu / \epsilon^2)$, and r according to

$$v_d r^d (1 - \epsilon/12)(1 - \epsilon/6) \cdot \lambda = \frac{k}{n} + \frac{C_\delta}{n} \sqrt{k\mu}.$$

8.9.8 Connection radius for polynomially bounded densities

In this section, we prove that in our algorithm (Figure 8.1), we can pick the connection radius R to be $R := 4r$, independent of the other parameters, provided that the density level satisfies $\lambda \leq n^A$ for some absolute constant A . (Our original setting picked $R = 4\rho$ and $r \leq \rho$.)

More precisely, we will argue that the parameter μ in the algorithm can be safely replaced by a related parameter $\tilde{\mu} := 2A \log n$ without affecting the performance of the algorithm. Pick $k = O(C_\delta^2 \tilde{\mu} / \epsilon^2)$, and set r, R by the equations

$$v_d r^d \lambda = \frac{1}{1 - \epsilon/6} \left(\frac{k}{n} + \frac{C_2 \log(1/\delta)}{n} \sqrt{k \tilde{\mu}} \right),$$

$$R = 4r.$$

The crucial ingredient in the analysis of our algorithm is the uniform convergence property of balls centered at the sample points and net points (Lemma 8.3.2), so we first verify that this statement remains true. Note that by our choice of r , we have

$$v_d r^d \lambda \geq \frac{k}{n} \geq \frac{1}{n},$$

so that $1/r^d \leq v_d n \lambda \leq v_d n^{A+1} \leq n^{A+1}$ (since $v_d < 1$ for sufficiently large d). As before, we consider a net \mathcal{N} of radius $R/4$ (i.e., r); by Lemma 8.9.5, size of this net is at most c^d / r^d for some absolute constant $c > 0$. Thus by Lemma 8.3.2, we have the uniform convergence property, provided the parameter μ is replaced by

$$\log n + \log |\mathcal{N}| = \log n + \log(1/r^d) + O(1) = (A + 2) \log n + O(1).$$

Notice that $\tilde{\mu}$ is picked to be a safe upper bound on this quantity, hence the lemma holds when μ is replaced by $\tilde{\mu}$.

Finally, it is easy to check that our choice of parameters satisfies all the conditions given in the separation lemma. Hence the separation and connectedness guarantees (Lemmas 8.3.4 and 8.3.5), together with their proofs, remain unaffected.

8.10 Discussion

In this chapter we have shown that simple non-parametric estimators based on k nearest neighbors and kernel density estimates are manifold adaptive estimators of the cluster tree. We have also introduced the problem of cluster tree recovery in the presence of noise. Many open questions remain, particularly regarding the minimax optimal rates of convergence and rates of convergence in the tubular noise case which we hope to address in future work.

One of the main advantages of the k nearest neighbors based estimator is its easy computability. In the case of *known* manifolds we have shown a more general *spatially adaptive* algorithm achieves better rates and in current work we are trying to understand the extent to which spatially adaptive estimators can help when the manifold is unknown.

Finally, simple modifications of these simple non-parametric estimators can also be used as estimators of various geometric properties of the level sets of the density. We are currently working on these extensions.

Chapter 9

Conclusions and Future Work

Much of this thesis characterizes how structure helps avoid the curse of dimensionality in a variety of problems. A genuinely unified and comprehensive understanding of various notions of intrinsic low-dimensionality and their effect on our ability to learn from noisy and high-dimensional data is still not available.

Statistical performance guarantees are only one half of the story. Often our ability to make inferences from large datasets is limited by the limited computational resources we have available. Understanding and characterizing tradeoffs between statistical and computational complexities is an important future direction. Recently, for instance Shender and Lafferty [170] have attempted a partial characterization for linear regression. Developing a framework, akin to the minimax framework that we have used throughout this thesis, but one that takes computational complexity into account would be extremely interesting.

While these are broad goals that we hope to address in the near future, we conclude this chapter with several concrete problems that we are addressing in our current work.

9.1 Sparse high-dimensional inference

9.1.1 Sparse Maximum Mean Discrepancy

The two sample testing problem is the following hypothesis testing problem: given two sets of samples, one from a distribution \mathbb{P} and the other from a distribution \mathbb{Q} , distinguish if $\mathbb{P} = \mathbb{Q}$ or not.

One natural approach to this problem, is an RKHS embedding based test, which uses a test statistic known as the Maximum Mean Discrepancy (MMD) [86].

$$\text{MMD}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \|\mathbb{P}f - \mathbb{Q}f\|_{\mathcal{H}}$$

where \mathcal{F} is an RKHS with kernel k .

The MMD is similar in form to the maximum kernel CCA coefficient and similar to kernel CCA suffers from the curse of dimensionality and is unsuited to the high-dimensional two sample testing problem.

Motivated by similar considerations to the one in Chapter 3 we can define a sparse additive MMD which computes the MMD statistic over additive RKHSs. Formally the population statistic is,

$$\text{MMD}_s(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \|\mathbb{P}f - \mathbb{Q}f\|_{\mathcal{H}} - \lambda_1 \|f\|_{\mathcal{D}} - \lambda_2 \|f\|_1$$

where \mathcal{F} is now an additive RKHS,

$$\mathcal{F} = \left\{ f : f(X) = \sum_{j=1}^p f_j(X_j), f_j \in \mathcal{H}_j \right\}$$

In current work we are investigating this statistic in an attempt to precisely characterize when it outperforms the vanilla MMD.

9.1.2 Convex relaxations and sparse additive kernel PCA

There are two other interesting directions in which the work of Chapter 3 could be extended. One is to consider other matrix factorization problems like principal components analysis (PCA) and the other is to consider convex relaxations. We describe both of these together.

Kernel PCA is often motivated as PCA in feature space. One can take a slightly different perspective. One possible proposal for kernel PCA is:

$$\begin{aligned} \max_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n f^2(X_i) \\ \text{subject to } \|f\|_{\mathcal{H}}^2 \leq 1 \end{aligned}$$

For a gram matrix K , using the representer theorem we obtain

$$\begin{aligned} \max_{\alpha} \frac{1}{n} \alpha^T K^2 \alpha \\ \text{subject to } \alpha^T K \alpha \leq 1 \end{aligned}$$

which is just a generalized eigenvalue problem. If K is invertible then it is equivalent to the eigenvalue problem

$$\begin{aligned} \max_{\alpha} \frac{1}{n} \alpha^T K \alpha \\ \text{subject to } \alpha^T \alpha \leq 1 \end{aligned}$$

The natural semidefinite lift of the PCA problem from d'Aspremont et al. [55] is

$$\begin{aligned} \max_{V \succeq 0} \quad & \text{tr}(V X^T X) \\ \text{subject to} \quad & \text{tr}(V) = 1 \end{aligned}$$

Now, the sparse version is to just

$$\begin{aligned} \max_{V \succeq 0} \quad & \text{tr}(V X^T X) \\ \text{subject to} \quad & \text{tr}(V) = 1 \\ & \|V\|_1 \leq c_1 \end{aligned}$$

In additive kernel PCA we focus on additively decomposed Hilbert spaces. So we have

$$f(X) = \sum_{i=1}^p f_j(X_j)$$

We would like to now induce sparsity at the level of functions. Let us focus on the convex relaxation approach.

We would like to solve

$$\begin{aligned} \max_{\alpha} \quad & \frac{1}{n} \left(\sum_j K_{X_j} \alpha_j \right)^T \left(\sum_j K_{X_j} \alpha_j \right) \\ \text{subject to} \quad & \sum_{j=1}^p \sqrt{\alpha_j^T K_{X_j} \alpha_j} \leq 1 \\ & \sum_{j=1}^p \sqrt{\alpha_j^T K_{X_j}^2 \alpha_j} \leq c_1 \end{aligned}$$

Ignoring cross terms in the objective we arrive at

$$\begin{aligned} \max_{\alpha} \quad & \frac{1}{n} \sum_j \alpha_j^T K_{X_j}^2 \alpha_j \\ \text{subject to} \quad & \sum_{j=1}^p \sqrt{\alpha_j^T K_{X_j} \alpha_j} \leq 1 \\ & \sum_{j=1}^p \sqrt{\alpha_j^T K_{X_j}^2 \alpha_j} \leq c_1 \end{aligned}$$

The natural convex lift of this problem would be to

$$\begin{aligned} & \max_{A \succeq 0} \frac{1}{n} \sum_j \text{tr}(A_j K_{X_j}^2) \\ & \text{subject to } \sum_{j=1}^p \sqrt{\text{tr}(A_j K_{X_j})} \leq 1 \\ & \sum_{j=1}^p \sqrt{\text{tr}(A_j K_{X_j}^2)} \leq c_1 \end{aligned}$$

In current work we are investigating the statistical properties of this approach to sparse additive kernel PCA.

9.1.3 Fast algorithms for additive kernel problems

Despite the impressive theoretical guarantees for the additive kernel formulations for regression [115, 156] and CCA [20], they are not widely used because of the computational difficulty in solving these problems. These problems are typically second order cone programs for which off-the-shelf solutions are not yet scalable. The backfitting algorithms for the functional versions [20, 157] are often much more tractable and preferred in practice.

It would be interesting to investigate the scalability of new first order optimization methods like the Alternating Directions Method of Multipliers (ADMM) algorithm of Boyd et al. [32].

Consider the following additive kernel regression formulation from Raskutti et al. [156]

$$\begin{aligned} (\hat{\alpha}_1, \dots, \hat{\alpha}_p) = \arg \min_{\alpha_j \in \mathbb{R}^n, \alpha_j^T K_{X_j} \alpha_j \leq 1} & \left\{ \frac{1}{2n} \|y - \sum_{j=1}^p K_{X_j} \alpha_j\|_2^2 + \right. \\ & \left. \lambda_n \sum_{j=1}^p \sqrt{\frac{1}{n} \|K_{X_j} \alpha_j\|_2^2} + \rho_n \sum_{j=1}^p \sqrt{\alpha_j^T K_{X_j} \alpha_j} \right\} \end{aligned} \quad (9.1)$$

One possible ADMM procedure solves the following equivalent program

$$\begin{aligned} (\hat{\alpha}_1, \dots, \hat{\alpha}_p) = \arg \min_{\alpha_j \in \mathbb{R}^n} & \left\{ \frac{1}{2n} \|y - \sum_{j=1}^p K_{X_j} \alpha_j\|_2^2 + \right. \\ & \left. \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^p \|x_j\|_2 + \rho_n \sum_{j=1}^p \|z_j\|_2 + \sum_{j=1}^p g_j(w_j) \right\} \end{aligned}$$

$$\begin{aligned} \text{subject to } w_j &= \alpha_j \\ x_j &= K_{X_j} \alpha_j \\ z_j &= K_{X_j}^{1/2} \alpha_j \end{aligned}$$

where $g_j(x)$ is the convex indicator function of the set $x^T K_{X_j} x \leq 1$, i.e. $g_j(x) = 0$ if $x^T K_{X_j} x \leq 1$ and ∞ otherwise.

Now, we form the augmented Lagrangian. For a given penalty parameter ρ .

$$\begin{aligned} L(\alpha, w, x, z, a, b, c) = & \left\{ \frac{1}{2n} \left\| y - \sum_{j=1}^p K_{X_j} \alpha_j \right\|_2^2 + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^p \|x_j\|_2 + \rho \sum_{j=1}^p \|z_j\|_2 + \sum_{j=1}^p g_j(w_j) \right. \\ & + \sum_{j=1}^p a_j^T (w_j - \alpha_j) + \sum_{j=1}^p b_j^T (z_j - K_{X_j}^{1/2} \alpha_j) + \sum_{j=1}^p c_j^T (x_j - K_{X_j} \alpha_j) \\ & \left. + \frac{\rho}{2} \left(\sum_{j=1}^p \|w_j - \alpha_j\|_2^2 + \|z_j - K_{X_j}^{1/2} \alpha_j\|_2^2 + \|x_j - K_{X_j} \alpha_j\|_2^2 \right) \right\} \end{aligned}$$

The ADMM algorithm involves minimizing this expression in α, w, x, z and then performing dual ascent on a, b, c . Each of these steps is reasonably simple (most are closed form). After some calculus we arrive at the following algorithm:

1.

$$\alpha_j \leftarrow \underbrace{\left(K_{X_j}^2/n + \rho(I + K_{X_j} + K_{X_j}^2) \right)^{-1}}_{\text{cache this}} \left(K_{X_j} y/n + a \right. \\ \left. + K_{X_j}^{1/2} b + K_{X_j} c + \rho(w + K_{X_j}^{1/2} z + K_{X_j} x) \right)$$

$$2. \quad z_j \leftarrow S_{\rho n/\rho} \left(\|K_{X_j}^{1/2} \alpha_j - \frac{b_j}{\rho}\|_2 \right) \frac{K_{X_j}^{1/2} \alpha_j - \frac{b_j}{\rho}}{\|K_{X_j}^{1/2} \alpha_j - \frac{b_j}{\rho}\|_2}$$

$$3. \quad x_j \leftarrow S_{\lambda_n/(\rho\sqrt{n})} \left(\|K_{X_j} \alpha_j - \frac{c_j}{\rho}\|_2 \right) \frac{K_{X_j} \alpha_j - \frac{c_j}{\rho}}{\|K_{X_j} \alpha_j - \frac{c_j}{\rho}\|_2}$$

$$4. \quad w_j \leftarrow \Pi_{C_j}(\alpha_j)$$

where Π_{C_j} is the Euclidean projection onto the set $\alpha_j^T K_{X_j} \alpha_j \leq 1$. The projection $\Pi(x)$ can be computed in two steps:

(a) If $x^T K_{X_j} x \leq 1$ return x .

(b) Else return $u = (\lambda K_{X_j} + I)^{-1} \alpha_j$ where λ is selected so that $u^T K_{X_j} u = 1$.

$$5. \quad a_j \leftarrow a_j + \rho(w_j - \alpha_j)$$

$$6. b_j \leftarrow b_j + \rho(z_j - K_{X_j}^{1/2} \alpha_j)$$

$$7. c_j \leftarrow c_j + \rho(x_j - K_{X_j} \alpha_j)$$

It would be interesting to comprehensively compare this algorithm with the backfitting procedures in terms of their computational complexity.

9.2 Other statistical problems in topological data analysis

In this thesis (in particular Chapter 7) we have focussed on homology inference from random samples. Homology inference requires the selection of a tuning parameter to select the “scale”, i.e. the radius parameter in the union of balls. *Persistent* homology is a method for probing topological properties of point clouds and functions. The method involves tracking the birth and death of topological features as one varies this tuning parameter. Features with short lifetimes are informally considered to be “topological noise.” Many of these fascinating ideas however do not yet have a rigorous statistical backing. In recent work [16], we derived confidence intervals on topological features. This allows us to distinguish between significant features and topological noise. There are several other interesting questions that we hope to investigate in future work including for instance understanding the power (ability to control the Type II error) of the confidence intervals we have proposed.

9.2.1 Machine learning with topological features

TDA provides the user with an extensive toolbox of interesting topological summaries of point clouds. Recently the papers [142, 152, 153] have considered various supervised learning problems on distributions (i.e. where each data point is a distribution or a sample from a distribution).

More generally one could consider both supervised and unsupervised learning problems on general point clouds. In this context, TDA could be useful to generate various powerful features from these point clouds on which standard machine learning algorithms could be applied.

9.3 Clustering with noisy and high-dimensional data

There are several natural extensions to the work on clustering described in this thesis. In the paper [118], we considered the extension of our work on spectral clustering to selectively sampled similarities. The algorithm we consider in this work selectively samples entire rows of the similarity matrix, which is natural for instance in network tomography applications. Characterizing the minimax rate here remains open, as does the problem of precisely understanding spectral clustering with randomly sampled similarities.

In this thesis we considered the problem of block structured activations, which arise for instance when a natural ordering of objects (and features) is known. A natural extension would be to consider the problem of clustering with side information, where the side information is used to infer this ordering (perhaps partially). In the worst case of course this problem reduces to the bi-clustering problem, also considered in this thesis.

Bibliography

- [1] D. Achlioptas and F. Mcsherry. On spectral learning of mixtures of distributions. In *Computational Learning Theory*, pages 458–469. 2005. 68
- [2] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *Ann. Statist.*, 38(5):3063–3092, 2010. 10, 114, 118
- [3] R. J. Adler, O. Bobrowski, M. S. Borman, E. Subag, and S. Weinberger. Persistent homology for random fields and complexes. In J. O. Berger, T. Cai, and I. M. Johnstone, editors, *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, pages 124–143. Institute of Mathematical Statistics, 2010. 168
- [4] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 2003. 39
- [5] D. Altschuh, T. Vernet, P. Berti, D. Moras, and K. Nagai. Coordinated amino acid changes in homologous protein families. *Protein Eng.*, 2(3):193–199, 1988. 34
- [6] A. Amini and M. Wainwright. High-Dimensional Analysis Of Semidefinite Relaxations For Sparse Principal Components. *The Annals of Statistics*, 37(5B):2877–2921, 2009. 117, 124, 128
- [7] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007. 34
- [8] E. Arias-Castro. Detecting a vector based on linear measurements. *Electronic Journal of Statistics*, 6:547–558, 2012. 139, 141, 142, 143, 145, 146
- [9] E. Arias-Castro, E. Candès, and M. Davenport. On the fundamental limits of adaptive sensing. *arXiv:1111.4646*, 2011. 140, 141, 142, 152
- [10] E. Arias-Castro, E. Candès, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011. 139
- [11] E. Arias-Castro, E. J. Candès, and A. Durand. Detection of an anomalous cluster in a network. *Ann. Stat.*, 39(1):278–304, 2011. 10, 114, 118
- [12] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *Ann. Statist.*, 36(4):1726–1757, 2008. 10, 114, 117
- [13] E. Arias-Castro, D. L. Donoho, and X. Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.*, 34(1):326–349, 2006.

- [14] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2003. 43, 44, 46, 52
- [15] J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):pp. 135–171, 2003. 117
- [16] S. Balakrishnan, B. Fasy, F. Lecci, A. Rinaldo, A. Singh, and L. Wasserman. Statistical inference for persistent homology. 2013. [arXiv:1303.7117](#). 233
- [17] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011. 12
- [18] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh. Recovering block-structured activations using compressive measurements. 2012. [arXiv:1209.3431](#). 12
- [19] S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. 2013. [arXiv:1307.6515](#). 12
- [20] S. Balakrishnan, K. Puniyani, and J. D. Lafferty. Sparse additive functional and kernel cca. *ICML*, 2012. 9, 12, 231
- [21] S. Balakrishnan, A. Rinaldo, D. Sheehy, A. Singh, and L. Wasserman. Minimax rates for homology inference. *AISTATS*, 2012. 12, 211, 213
- [22] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 954–962. MIT Press, 2011. 12, 68
- [23] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010. 140, 142
- [24] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002. 50
- [25] U. Bayer, P. Milani Comparetti, C. Hlauscheck, C. Kruegel, and E. Kirda. Scalable, Behavior-Based Malware Clustering. In *16th Symposium on Network and Distributed System Security (NDSS)*. 2009. 115
- [26] F. Benaych-Georges and R. Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *ArXiv e-prints*, 2011. [1103.2221](#). 122
- [27] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000. viii, 26
- [28] J. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3):616–618, 1977. 18
- [29] S. Bhamidi, P. S. Dey, and A. B. Nobel. Energy Landscape for large average submatrix detection problems in Gaussian random matrices. *ArXiv e-prints*, 2012. [1211.2284](#). 140, 143

- [30] P. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Technical report, Department of Statistics, UC Berkeley*. 2006. 205
- [31] L. Birgé. An alternative point of view on Lepski’s method. *Lecture Notes-Monograph Series*, 36:113–133, 2001. 165
- [32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011. 231
- [33] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 21
- [34] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *JASA*, 80(391):pp. 580–598, 1985. 44
- [35] S. C. Brubaker and S. Vempala. Isotropic pca and affine-invariant clustering. In *FOCS*, pages 551–560. 2008. 68
- [36] S. Busygin, O. Prokopyev, and P. Pardalos. Biclustering in data mining. *Computers & Operations Research*, 35(9):2964–2987, 2008. 117
- [37] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *ArXiv e-prints*, 2011. 1109.0898. 140, 143
- [38] C. Butucea, Y. I. Ingster, and I. Suslina. Sharp Variable Selection of a Sparse Submatrix in a High-Dimensional Noisy Matrix. *ArXiv e-prints*, 2013. 1303.5647. 140, 143
- [39] E. Candès and M. Davenport. How well can we estimate a sparse vector? *arXiv:1104.5246*, 2011. 140, 144
- [40] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. 139
- [41] E. Candès and T. Tao. The dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics*, 35(6):2313–2351, 2007. 139
- [42] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 21, 2008. 139
- [43] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009. 136
- [44] K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 343–351. 2010. 12, 199, 200, 201, 202, 204, 205, 206, 208, 209, 216, 218, 221
- [45] K. Chaudhuri, F. C. Graham, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research - Proceedings Track*, 23:35.1–35.23, 2012. 68, 87
- [46] F. Chazal. An upper bound for the volume of geodesic balls in submanifolds of euclidean spaces. *Personal Communication, available at*

- <http://geometrica.saclay.inria.fr/team/Fred.Chazal/BallVolumeJan2013.pdf>, 2013. 205, 219
- [47] F. Chazal, D. Cohen-Steiner, and Q. Merigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 2011. To appear. 168
- [48] A. Chen, A. A. Amini, P. J. Bickel, and E. Levina. Fitting community models to large sparse networks. *CoRR*, abs/1207.2340, 2012. 68
- [49] X. Chen and H. Liu. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Biosciences*, pages 1–24, 2012. 43, 44, 55
- [50] M. K. Chung, P. Bubenik, and P. T. Kim. Persistence diagrams of cortical surface data. In *Proceedings of the 21st International Conference on Information Processing in Medical Imaging*, IPMI '09, pages 386–397. Springer-Verlag, Berlin, Heidelberg, 2009. 168
- [51] I. Csiszar and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006. 39
- [52] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Annals of Statistics*, 25(6):2300–2312, 1997. 201
- [53] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Aust. N. Z. J. Stat.*, 48(1):7–19, 2006. 201
- [54] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *STOC*, pages 537–546. 2008. 205
- [55] A. d’Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49:434–448, 2007. 117, 122, 123, 124, 230
- [56] M. Davenport and E. Arias-Castro. Compressive binary search. *arXiv:1202.0937*, 2012. 140, 141, 142, 144, 149, 161
- [57] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1:317–366, 2001. 135
- [58] V. de Silva. *PLEX: Simplicial complexes in MATLAB*, 2013. 167, 177
- [59] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In M. Alexa and S. Rusinkiewicz, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2004. 167, 177
- [60] V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007. 168
- [61] M. Dequeant et al. Comparison of pattern detection methods in microarray time series of the segmentation clock. *PLoS ONE*, 3(8):e2856, 2008. 168
- [62] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278(5338):680–686, 1997. 78
- [63] A. Dhulesia, J. Gsponer, and M. Vendruscolo. Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a pdz domain protein. *Journal of the American Chemical Society*, 130(28):8931–8939, 2008. 27

- [64] D. Donoho. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 139
- [65] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004. 148
- [66] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk. Sparse signal detection from incoherent projections. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2006. 139
- [67] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998. 14, 21, 25
- [68] H. Edelsbrunner and J. Harer. *Computational topology*. American mathematical society, 2009. 168
- [69] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak. Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities. *CoRR*, abs/1102.3887, 2011. 78
- [70] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991. 169, 176, 183
- [71] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604, 2010. 44
- [72] S. N. Fatakia, S. Costanzi, and C. C. Chow. Computing highly correlated positions using mutual information and graph theory for g protein-coupled receptors. *PLoS ONE*, 4(3):e4681, 2009. 34
- [73] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2003. 35
- [74] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam Protein Families Database. *Nucleic Acids Research*, 2010. 8, 14, 21, 25, 29, 78
- [75] R. Fletcher. Semi-definite matrix constraints in optimization. *SIAM Journal on Control and Optimization*, 23:493, 1985. 136
- [76] A. A. Fodor and R. W. Aldrich. On evolutionary conservation of thermodynamic coupling in proteins. *Journal of Biological Chemistry*, 279(18):19046–19050, 2004. 34
- [77] R. Foygel, M. Horrell, M. Drton, and J. Lafferty. Nonparametric reduced rank regression. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1637–1645. MIT Press, 2012. 9
- [78] A. Fuchs, A. J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, and D. Frishman. Co-evolving residues in membrane proteins. *Bioinformatics*, 23(24):3312–3319, 2007. 34
- [79] E. Fuentes, C. Der, and A. Lee. Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *Journal of molecular biology*, 335(4):1105–1115, 2004. 27, 29
- [80] J. Gamble and G. Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, 101(9):2184–2199, 2010. 168
- [81] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold

- estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012. 208, 211
- [82] B. Gidas. Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs Distributions. *Institute for Mathematics and Its Applications*, 10:129–+, 1988. 19
- [83] E. Giné and A. Guillaou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l’Institut Henri Poincaré (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002. 214, 215, 216
- [84] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, 18(4):309–317, 1994. 34
- [85] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004. 51
- [86] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel method for the two-sample problem. *JMLR*, 2012. 228
- [87] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, , B. Schoelkopf, and N. Logothetis. Behaviour and convergence of the constrained covariance. Technical Report 130, MPI for Biological Cybernetics, 2004. 50, 59
- [88] J. Hartigan. *Clustering Algorithms*. Wiley, 1975. 87
- [89] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):pp. 123–129, 1972. 114
- [90] J. A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):pp. 388–394, 1981. 87, 199, 201
- [91] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):pp. 297–310, 1986. 43
- [92] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. 167, 171
- [93] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Proceedings of the 43rd Asilomar conference on Signals, systems and computers*, pages 1551 –1555. 2009. 140, 142
- [94] J. Haupt and R. Nowak. Compressive sampling for signal detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1509–1512. 2007. 139
- [95] M. Hein and M. Maier. Manifold denoising. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 561–568. MIT Press, 2006. 168
- [96] H. Hoefling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10:883–906, 2009. 19, 21
- [97] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):pp. 321–377, 1936. 8, 42
- [98] L. Huang, D. Yan, M. I. Jordan, and N. Taft. Spectral Clustering with Perturbed Data. In *Advances in Neural Inforation Processing Systems*. 2009. 68

- [99] Y. I. Ingster, A. B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010. [139](#)
- [100] J. Jang, D. Brumley, and S. Venkataraman. Bitshred: feature hashing malware for scalable triage and semantic analysis. In *Proceedings of the 18th ACM conference on Computer and communications security, CCS '11*, pages 309–320. 2011. [140](#)
- [101] J. Jin. Fast network community detection by SCORE. *ArXiv e-prints*, 2012. [1211.5803](#). [68](#)
- [102] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. [115](#), [117](#)
- [103] I. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. [117](#), [165](#)
- [104] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *JASA*, 104(486):682–693, 2009. [44](#)
- [105] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000. [51](#)
- [106] M. Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658 – 1671, 2009. [168](#)
- [107] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *18th Annual Conference on Learning Theory (COLT)*, pages 444–457. 2005. [68](#)
- [108] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14:846–856, 1998. [14](#)
- [109] K. Karplus, K. Sjlander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm, C. Sander, E. England, and E. England. Predicting protein structure using hidden markov models. In *Proteins: Structure, Function, and Genetics*, pages 134–139. 1997. [14](#)
- [110] P. M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. J. Guibas, and V. S. Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007. [168](#)
- [111] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *NIPS*, pages 909–917. 2011. [12](#), [140](#), [143](#), [164](#)
- [112] V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011. [140](#), [143](#)
- [113] V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Ann. Statist.*, 28(2):591–629, 2000. [169](#), [176](#), [183](#), [193](#)
- [114] V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *Ann. Statist.*, 28(2):591–629, 2000. [213](#)
- [115] V. I. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*,

- 38(6):3660–3695, 2010. 43, 55, 231
- [116] S. Kpotufe and S. Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *J. Comput. Syst. Sci.*, 78(5):1496–1515, 2012. 205
- [117] S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 225–232. ACM, New York, NY, USA, 2011. 201
- [118] A. Krishnamurthy, S. Balakrishnan, M. Xu, and A. Singh. Efficient active algorithms for hierarchical clustering. *ICML*, 2012. 233
- [119] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994. 14
- [120] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 299–308. IEEE Computer Society, Washington, DC, USA, 2010. 87
- [121] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. 165
- [122] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica sinica*, 12:61–86, 2002. 115, 124
- [123] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010. 115, 124
- [124] S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l_1 -regularization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 817–824. MIT Press, Cambridge, MA, 2007. 8, 35
- [125] S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l_1 -regularization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 817–824. MIT Press, Cambridge, MA, 2007. 18, 19
- [126] E. Lehmann and J. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 2005. 175
- [127] G. e. a. Lenz. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. U.S.A.*, 105:13520–13525, 2008. 55
- [128] J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned dag model: Investigation of a bayesian and a frequentist approach. *23rd annual conference on Uncertainty in Artificial Intelligence*, 2007. 24
- [129] D. C. Liu, J. Nocedal, D. C. Liu, and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989. 21
- [130] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric

- Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012. 9
- [131] J. Liu and W. Wang. Op-cluster: Clustering by tendency in high dimensional space. In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM '03*, pages 187–. IEEE Computer Society, Washington, DC, USA, 2003. 115
- [132] Y. Liu, J. G. Carbonell, P. Weigele, and V. Gopalakrishnan. Protein fold recognition using segmentation conditional random fields. *Journal of Computational Biology*, 13(2):394–406, 2006. 14
- [133] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:295–299, 1999. 14, 21, 27, 29, 34
- [134] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on computational Biology and Bioinformatics*, pages 24–45, 2004. 117
- [135] M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. *Theor. Comput. Sci.*, 410(19):1749–1764, 2009. 201
- [136] M. Malloy and R. Nowak. Near-optimal compressive binary search. *arXiv:1203.1804*, 2012. 140, 141, 142, 161, 162
- [137] K. Mardia, J. Kent, and J. Bibby. *Multivariate analysis*. Academic press, 1980. 160
- [138] F. McSherry. Spectral partitioning of random graphs. In *IEEE Symposium on Foundations of Computer Science*, page 529. 2001. ix, 67, 68, 70, 75, 77, 112
- [139] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009. 43
- [140] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, J. R. D’Agostino, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, B. Gaston, N. N. Jarjour, R. Sorkness, W. J. Calhoun, K. F. Chung, S. A. A. Comhair, R. A. Dweik, E. Israel, S. P. Peters, W. W. Busse, et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med*, 181(4):315–323, 2010. 140
- [141] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995. 184
- [142] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 10–18. MIT Press, 2012. 233
- [143] H. Narayanan, M. Belkin, and P. Niyogi. On the relation between low density separation, spectral clustering and graph cuts. In *Advances in Neural Information Processing Systems (NIPS) 19*. 2006. 87
- [144] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011. 140, 143
- [145] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001. 68, 71, 72

- [146] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008. [168](#), [169](#), [178](#), [185](#), [186](#), [191](#), [201](#), [205](#), [206](#), [218](#), [220](#)
- [147] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning and clustering. *SIAM J. Comput.*, 40(3):646–663, 2011. [168](#), [171](#), [188](#), [192](#), [193](#), [195](#)
- [148] A. Onatski. Asymptotics of the principal components estimator of large factor models with weak factors. *Economics Department, Columbia University*, 2009. [117](#)
- [149] E. Parkhomenko, D. Tritchler, and J. Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc*, 1 Suppl 1, 2007. [43](#), [44](#)
- [150] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004. [117](#)
- [151] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny. *Topological methods in Data Analysis and Visualization: Theory, Algorithms and Applications*. Springer, 2001. [168](#)
- [152] B. Poczos, A. Rinaldo, A. Singh, and L. Wasserman. Distribution-free distribution regression. 2013. [arXiv:1302.0082](#). [233](#)
- [153] B. Poczos, L. Xiong, D. J. Sutherland, and J. Schneider. Support distribution machines. Technical report, Carnegie Mellon University, 2012. [233](#)
- [154] D. D. Pollock and W. R. Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.*, 10(6):647–657, 1997. [35](#)
- [155] W. Polonik. Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Annals of Statistics*, 23(3):855–882, 1995. [201](#)
- [156] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *JMLR*, 2010. [43](#), [231](#)
- [157] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *JRSSB (Statistical Methodology)*, 71(5):1009–1030, 2009. [9](#), [43](#), [45](#), [50](#), [59](#), [231](#)
- [158] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1 regularized logistic regression. *Annals of Statistics*, to appear, 2009. [8](#)
- [159] P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009. [201](#)
- [160] A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *Journal of Machine Learning Research*, 13:905–948, 2012. [201](#)
- [161] A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010. [0907.3454](#). [201](#), [214](#), [215](#)
- [162] R. Rockafellar. *The theory of subgradients and its applications to problems of optimization. Convex and nonconvex functions*. Heldermann, 1981. [136](#)
- [163] K. Rohe, S. Chatterjee, and B. Yu. Spectral Clustering and the High-Dimensional Stochastic Block Model. *Technical Report 791, Statistics Department, UC Berkeley*, 2010. [68](#), [70](#), [71](#)
- [164] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme

- singular values. In *International Congress of Mathematicians*. 2010. 92
- [165] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437:579–583, 2005. 14, 25, 39
- [166] A. Sacan, O. Ozturk, H. Ferhatosmanoglu, and Y. Wang. Lfm-pro: a tool for detecting significant local structural sites in proteins. *Bioinformatics*, 23:709–716, 2007. 168
- [167] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. In *CVPR*. IEEE Computer Society, 2008. 18, 19, 21, 35
- [168] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. 39
- [169] H. Shen and J. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008. 117
- [170] D. Shender and J. Lafferty. Computation-risk tradeoffs for covariance-thresholded regression. In *ICML '13: Proceedings of the 30th Annual International Conference on Machine Learning*. 2013. 228
- [171] V. D. Silva and R. Ghrist. Homological sensor networks. *Notices of the American Mathematical Society*, 54:2007, 2007. 168
- [172] A. Singh, C. Scott, and R. Nowak. Adaptive $\{H\}$ ausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782, 2009. 201
- [173] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, and D. L. Ringach. Topological analysis of population activity in visual cortex. *J. Vis.*, 8(8):1–18, 2008. <http://journalofvision.org/8/8/11/Singh-2008-jov-8-8-11.pdf>. 168
- [174] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437:512–518, 2005. 14, 25, 34
- [175] A. Soni and J. Haupt. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. *arXiv:1111.6923*, 2011. 140, 142
- [176] D. Spielman. *Lecture Notes on Spectral Graph Theory*, 2009. 90
- [177] B. K. Sriperumbudur and I. Steinwart. Consistency and rates for clustering with dbscan. *Journal of Machine Learning Research - Proceedings Track*, 22:1090–1098, 2012. 201
- [178] I. Steinwart. Adaptive density level set clustering. *Journal of Machine Learning Research - Proceedings Track*, 19:703–738, 2011. 201
- [179] G. Stewart. Perturbation theory for the singular value decomposition. *Computer Science Technical Report Series; Vol. CS-TR-2539*, page 13, 1990. 131
- [180] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):025–047, 2003. 201
- [181] W. Stuetzle and N. R. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010. 201

- [182] X. Sun and A. B. Nobel. On the maximal size of Large-Average and ANOVA-fit Submatrices in a Gaussian Random Matrix. *ArXiv e-prints*, 2010. [1009.0562](#). [115](#), [117](#), [122](#), [140](#), [143](#)
- [183] G. M. Sel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10:59–69, 2003. [8](#)
- [184] A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. *Handbook of computational molecular biology*, 2004. [117](#)
- [185] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. In *BIOKDD '05: Proceedings of the 5th international workshop on Bioinformatics*, pages 12–20. ACM, New York, NY, USA, 2005. [14](#), [21](#)
- [186] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(2):183–197, 2008. [viii](#), [14](#), [21](#), [22](#), [23](#), [25](#), [34](#)
- [187] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins: Structure, Function, and Bioinformatics*, 76(4):911–29, 2009. [14](#), [34](#)
- [188] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):506–516, 2009. [14](#)
- [189] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Protein Design by Sampling an Undirected Graphical Model of Residue Constraints. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(3):506–516, 2009. [14](#), [34](#)
- [190] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994. [34](#)
- [191] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006. [19](#)
- [192] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009. [85](#), [107](#), [118](#), [125](#), [126](#), [145](#), [147](#), [154](#), [158](#)
- [193] A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997. [201](#)
- [194] L. Ungar and D. P. Foster. A formal statistical approach to collaborative filtering. In *CONALD*. 98. [115](#)
- [195] U. von Luxburg. A Tutorial on Spectral Clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics, 2006. [67](#)
- [196] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of Spectral Clustering. In *The Annals of Statistics*, pages 857–864. MIT Press, 2004. [67](#)
- [197] M. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

140, 141, 142, 148

- [198] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009. 140, 148
- [199] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1 -regularized logistic regression. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1465–1472. MIT Press, Cambridge, MA, 2007. 18, 19, 34
- [200] G. Walther. Granulometric smoothing. *Annals of Statistics*, 25(6):2273–2299, 1997. 201
- [201] S. Wang, R. R. Gutell, and D. P. Miranker. Biclustering as a method for RNA local multiple sequence alignment. *Bioinformatics*, 23:3289–3296, 2007. 115
- [202] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, 106:67–72, 2009. 34
- [203] D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining. In *Proceedings of the Colloquium on Numerical Taxonomy held in the University of St. Andrews*, pages 282–308. 1969. 201
- [204] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515, 2009. 117
- [205] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. 9, 42, 43, 44, 46, 47, 52, 54, 55
- [206] D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8:Article28, 2009. 9, 42, 43
- [207] Y. Yang. Can the strengths of aic and bic be shared? *BIOMETRICA*, 92:2003, 2003. 39
- [208] S. Yoon, C. Nardini, L. Benini, and G. De Micheli. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4):339–354, 2005. 140
- [209] B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997. 173
- [210] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006. 34
- [211] A. Zomorodian. *Topology for Computing*. Cambridge University Press, 2005. 168
- [212] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 117