Graph Structured Normal Means Inference

James Sharpnack

May 2013
CMU-ML-13-102

Carnegie Mellon

Graph Structured Normal Means Inference

James Sharpnack

May 2013
CMU-ML-13-102

School of Computer Science
Machine Learning Department

Dietrich College of Humanities and Social Sciences
Department of Statistics

Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aarti Singh, Co-chair
Alessandro Rinaldo, Co-chair
Larry Wasserman
Gary Miller
Ery Arias-Castro

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was funded in part by the Air Force Office of Scientific Research under grant number FA95501010382. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Normal means, Gaussian goodness-of-fit, graph structure, estimation, detection, localization, anomaly detection, statistical inference, network, Laplacian, spanning tree wavelets, spectral scan

To my parents, Jan and Doug.

Abstract

This thesis addresses statistical estimation and testing of signals over a graph when measurements are noisy and high-dimensional. Graph structured patterns appear in applications as diverse as sensor networks, virology in human networks, congestion in internet routers, and advertising in social networks. We will develop asymptotic guarantees of the performance of statistical estimators and tests, by stating conditions for consistency by properties of the graph (e.g. graph spectra). The goal of this thesis is to demonstrate theoretically that by exploiting the graph structure one can achieve statistical consistency in extremely noisy conditions.

We begin with the study of a projection estimator called *Laplacian eigenmaps*, and find that eigenvalue concentration plays a central role in the ability to estimate graph structured patterns. We continue with the study of the *edge lasso*, a least squares procedure with total variation penalty, and determine combinatorial conditions under which changepoints (edges across which the underlying signal changes) on the graph are recovered. We will shift focus to testing for anomalous activations in the graph, using the *scan statistic relaxations*, the spectral scan statistic and the graph ellipsoid scan statistic. We will also show how one can form a decomposition of the graph from a spanning tree which will lead to a test for activity in the graph. This will lead to the construction of a *spanning tree wavelet basis*, which can be used to localize activations on the graph.

Acknowledgments

First and foremost I would like to thank my parents, Janyce and Douglas Sharpnack. Without their love and support I would not have any of the opportunities that I have now. Aarti Singh has been my advisor since I entered into the Machine Learning portion of my Ph.D. program. If it were not for Aarti's patience, dedication to my intellectual development, and unwavering support, I would certainly not be the researcher I have become. I would like to also thank my co-advisor, Alessandro Rinaldo, who has been very supportive intellectually and personally. Each member of my committee (Larry Wasserman, Gary Miller, and Ery Arias-Castro) has had a formative effect on my career, either through courses or discussions. Also, this research would not be possible without my collaborators (other than my advisors): Akshay Krishnamurthy and Mladen Kolar. I would like to thank professors Chad Schafer, Ann Lee, and Steve Fienberg for their immense help early on in the Ph.D. program. Lastly, I would like to thank my great group of friends and family for their love and support: Orville and Ruth Egbert, Reatha and Michael Sharpnack, Becky and Zane Tarumoto, Victoria Werderitch, Darren Homrighausen, Dan McDonald, and Anne-Sophie Charest.

Contents

Chapter 1

Introduction

Statistical inference is inherently difficult when there are few samples and the parameter space is large. The only way to avoid this problem, given a limited amount of data, is to impose constraints on the parameter space. This thesis focuses on the problem of detecting, localizing and estimating patterns over a graph when observations are corrupted by noise. Hence, we consider the case when parameter constraints derive from a graph structure, generally graphs given by real world networks.

The problem of estimating graph-structured activations is relevant to many applications including identifying congestion in router and road networks, eliciting preferences in social networks, and localizing viruses in human and computer networks. While several machine learning algorithms are designed to estimate graph-structured patterns[12, 55, 63] very few statistical guarantees are known. Much less work addresses the detection of anomalous patterns in graphs from a statistical testing perspective. This is despite a variety of real-world applications such as activity detection in social networks, network surveillance, disease outbreak detection, biomedical imaging, sensor network detection, gene network analysis, environmental monitoring and malware detection. Recent theoretical contributions in the statistical literature[1, 3] have detailed the inherent difficulty of such a testing problem but have positive results only under restrictive conditions on the graph topology. By combining knowledge from high-dimensional statistics, graph theory and mathematical programming, the characterization of detection algorithms over any graph topology by their statistical properties is possible.

Aside from the statistical challenges, the computational complexity of these algorithms must be addressed. Due to the combinatorial nature of graph based methods, problems can easily shift from having polynomial-time algorithms to having running times exponential in the size of the graph. The applications of graph structured inference require that any method be scalable to large graphs. As we will see some proposed statistical procedures will be intractable, suggesting that approximation algorithms and relaxations are necessary. Luckily, computer science boasts a plethora of efficient graph based algorithms that are adaptable to these statistical problems. Before we elaborate on the statistical setup, we will examine two real-world examples in which one detects and localizes graph structured signals.

1.0.1 Disease Detection in Human Networks

Most people think of tests for contagious diseases as reporting a positive or negative indicating the existence of a disease. In fact, many common experimental techniques in virology report various indicators of a virus, such as antibody protein concentrations (western blot, enzyme-linked immunosorbent assay) or measuring virus concentrations directly (the plaque assay). One popular method, the western blot, reports concentrations by the shade of bands from an x-ray film darkened by a luminescent compound. Infectious diseases diffuse within human networks, e.g. I interact with my officemates and roommates quite often, they would be my neighbors in this network, and can communicate any diseases to them more easily then other individuals living in Pittsburgh. If we can exploit this network structure in the detection of infectious diseases, then we may be able to detect and localize an incipient infection under low signal-to-noise ratios (very light bands in the western blot).

1.0.2 Sensor Networks

One of the advantages of representing structure with a graph, and not through a Euclidean embedding, is its versatility when representing unexpected constraints and topological artifacts. Sensor networks might be deployed for detecting nuclear substances, water contaminants, or activity in video surveillance. When sensors are placed in an environment for the sake of detecting and localizing some activity, the structure inherent in that environment should influence the statistical inference. For example, the spread of a contaminant in water changes greatly when the environment is the Gulf of Mexico (as in the Deepwater Horizon oil spill) or a water supply network. Water supply contamination is a common cause for outbreaks of cholera, gastroenteritis, E. coli, and polio. Because of the potential for large scale health problems, it is of interest to detect contaminated water under low signal-to-noise regimes. As we will see, by exploiting the graph structure (in this case, the pipe network for the water supply), one can detect activity in networks when the activity is very faint. Furthermore, the graph structure provides a versatile framework for modeling environmental constraints.

1.1 Problem Setup

We will be studying the setting in which a graph that provides the structure to our inference is known a priori, and we are tasked with identifying latent parameters over the graph. An **undirected graph** is a set of **vertices** V which can be taken to be the natural numbers $[p] = \{1, \ldots, p\}$ for some $p > 0$ and pairs of vertices $E \subseteq V \times V$ called the **edges** (with $|E| = m$). To each edge is possibly associated a weight, which we will denote W_e for some $e \in E$. In this case we have a **weighted graph**, and we let the graph be the triplet $G = (V, E, W)$. We are now ready to define the graph-structured normal means problem, which will be the main focus of this thesis.

1.1.1 Structured Normal Means Problem

In the graph-structured normal means problem, we observe one realization of the random vector

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^p$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_p)$ is Gaussian white noise with *known* variance σ^2. The goal is to make inferences regarding the unknown $\mathbf{x}$, when it is believed to be smooth with respect to a graph. Before we define what it means for $\mathbf{x}$ to be smooth over a graph, we must first introduce two differential operators: the incidence matrix and the graph Laplacian.

We begin by constructing (arbitrarily) an **orientation** of G by defining a head $e^+ \in e$ and tail $e^- \in e$ for each edge $e \in E$. The **incidence matrix** $\nabla \in \mathbb{R}^{m \times p}$ for the oriented graph is the matrix whose $\nabla_{e,v}$ entry is 1 if $v = e^+$, -1 if $v = e^-$ and 0 otherwise (it is $\sqrt{W_e}$ or 0 in a weighted graph). The incidence matrix is indeed the discrete analogue of the gradient operator, which is a comparison to which we will frequently adhere. Another commonly studied discrete differential operator is the **Laplacian** of G which is defined as $\Delta = \nabla^\top \nabla$. Let $d_v = \sum_{w \in V} W_{v,w}$ and $\mathbf{D} = \mathrm{diag}(\{d_v\}_{v \in V})$ be the diagonal degree matrix. Then $\Delta = \mathbf{D} - \mathbf{W}$ is positive semi-definite and for $\mathbf{z} \in \mathbb{R}^p$, $\mathbf{z}^\top \Delta \mathbf{z} = \sum_{v,w} W_{v,w}(z_v - z_w)^2$.

Much of the study of the Laplacian revolves around its spectrum, which will play a central role in this thesis. Let us denote the decomposition $\Delta = \mathbf{U}\Lambda\mathbf{U}^\top$ for $\mathbf{U} \in \mathbb{R}^{p \times p}$ and diagonal $\Lambda = \mathrm{diag}(\{\lambda_i\}_{i=1}^p)$. Furthermore, let $\lambda_i \leq \lambda_{i+1}$ without loss of generality. Then it is known that $\lambda_1 = 0$ and $|\{i : \lambda_i = 0\}|$ is the number of connected components of the graph G. Furthermore, because of the invariance of trace under rotations, $\sum_{i=1}^p \lambda_i = \sum_{i=1}^p d_i$, hence the average eigenvalue is equal to the average degree.

For a vector, $\mathbf{z} \in \mathbb{R}^p$, define $\mathrm{supp}(\mathbf{z}) = \{i \in [p] : z_i \neq 0\}$ ($[p] = \{1, \ldots, p\}$), $\bar{z} = \frac{1}{n}\sum_{i=1}^p z_i$, $\bar{\mathbf{z}} = \bar{z}\mathbf{1}$, and $\|\mathbf{w}\|_0 = |\mathrm{supp}(\mathbf{w})|$. Furthermore, define the ℓ_k norms for $k > 0$ to be $\|\mathbf{z}\|_k = (\sum_{i \in [p]} z_i^k)^{1/k}$, with $\|\mathbf{z}\|_\infty = \max_{i \in [n]} |z_i|$. We will also be considering the induced norms of matrices, specifically for a matrix $\mathbf{M}$ let $|\!|\!|\mathbf{M}|\!|\!|_{k,l} = \sup_{\|\mathbf{z}\|_k \leq 1} \|\mathbf{M}\mathbf{z}\|_l$. Immediately, we see that $\|\nabla \mathbf{z}\|_2^2 = \mathbf{z}^\top \Delta \mathbf{z}$ and $\mathrm{supp}(\nabla \mathbf{z}) = \{e \in E : z_{e^+} \neq z_{e^-}\}$. Furthermore, if $A \subset V$ then let $\bar{A} = V \backslash A$ and let ∂A denote the edges leaving A (the boundary of A).

We now have the machinery in place to define what we mean by graph structured signals. This thesis studies estimation and detection primarily with respect several distinct function classes, with parameter $\rho > 0$:

1. ℓ_2 **graph-structure**: the class $\mathcal{X}_2(\rho) = \{\mathbf{x} \in \mathbb{R}^p : \|\nabla \mathbf{x}\|_2^2 \leq \rho\}$
2. ℓ_0 **graph-structure**: the class $\mathcal{X}_0(\rho) = \{\mathbf{x} \in \mathbb{R}^p : \|\nabla \mathbf{x}\|_0 \leq \rho\}$
3. **balanced graph-structure**: for each $\mathbf{x} \in \mathcal{X}_{PC}(\rho)$ there exists $A \subset V$, $\mathbf{x} = b_0 \mathbf{1}_A + b_1 \mathbf{1}_{\bar{A}}$ with $p|\partial A|/(|A||\bar{A}|) \leq \rho$ for some $b_0, b_1 \in \mathbb{R}$ (here PC stands for piece-wise constant)

We will also consider two separate prior distributions that are induced by the graph G, the Ising prior and the Gaussian graphical model (GGM). Below $p(\mathbf{x})$ denotes the density with respect to either the Lebesgue measure (in the GGM) or the discrete measure over the hypercube $\{0,1\}^p$ (for Ising).

$$\begin{array}{ll} \text{Gaussian graphical model:} & p(\mathbf{x}) \propto \exp(-\mathbf{x}^\top \Sigma^{-1} \mathbf{x}) \\ \text{Ising model:} & p(\mathbf{x}) \propto \exp(-\mathbf{x}^\top \Delta \mathbf{x}) \end{array} \tag{1.2}$$

In the Gaussian graphical model, $\Delta = \Sigma^{-1}$ denotes the inverse covariance matrix whose zero entries indicate the absence of an edge between the corresponding nodes in the graph. With these function classes and priors in mind, we study the statistical performance of estimation, localization, and detection.

1.1.2 Estimation and Changepoint Localization

There are two standards that we will ask of an estimator $\hat{\mathbf{x}}$ in the normal means problem: recovery with an ℓ_2 loss and the recovery of its structure through the changepoints, $\text{supp}(\nabla \mathbf{x})$. In Chapter 2 we will consider a Laplacian eigenmaps projection estimator $\hat{\mathbf{x}}$. We will say that the estimator is ℓ_2 *consistent* if

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \xrightarrow{\mathbb{P}} 0.$$

The risk $\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|$ is known as the mean square error (MSE). We will highlight conditions under which we can achieve ℓ_2 consistency for the ℓ_2 graph structure, the Ising, and the GGM in Chapter 2. *Localization* refers to the ability to obtain the elements of $C = \text{supp}(\mathbf{x})$, while tolerating some loss in the recovery. This setting will be studied in detail in Chapter 5. The following natural notion of distance between clusters ($C, C' \subseteq [p]$) will be important for both theoretical and validation purposes.

$$d(C, C') = 2\left(1 - \frac{|C \cap C'|}{\sqrt{|C||C'|}}\right)$$

The idea is that $d(C, C')$ is small if and only if there is large overlap between C and C' relative to their size. We will say that an estimator of $\hat{C}$ of C is localization consistent if

$$\mathbb{E}d(C, \hat{C}) \to 0 \Rightarrow d(C, \hat{C}) \xrightarrow{\mathbb{P}} 0$$

We will prove in Chapter 2 (Proposition 9) is that estimation consistency implies localization consistency.

Changepoint sparsistency is when we hope to recover exactly $\text{supp}(\nabla \mathbf{x})$, in fact we require that the signs of the changes be correctly recovered, i.e. ,

$$\lim_{p \to \infty} \mathbb{P}\{\text{sign}(\nabla \hat{\mathbf{x}}) = \text{sign}(\nabla \mathbf{x})\} = 1.$$

In Chapter 3, we study the changepoint sparsistency of the edge lasso, an instantiation of the generalized lasso.[78] We will see that while changepoint consistency may be a strong criteria, it directly leads to nearly-oracle rates of convergence for the ℓ_2 loss.

1.1.3 Detection

The detection of graph-structured signals may refer to one of two things: detecting an anomalous cluster from zero background activation, or detecting an anomalous cluster from constant

background activation. In Chapter 4, we study the case of *constant background activation* with balanced alternative, in which we assume the following testing hypotheses:

$$H_0 \colon \nabla \mathbf{x} = \mathbf{0} \quad \text{vs} \quad H_1 \colon \mathbf{x} \in \mathcal{X}_{PC}(\rho), \|\mathbf{x} - \bar{\mathbf{x}}\| \geq \mu \tag{1.3}$$

In Chapter 5, we study *zero background activation* in which we assume the null and alternative hypotheses:

$$H_0 \colon \mathbf{x} = \mathbf{0} \quad \text{vs} \quad H_1 \colon \mathbf{x} \in \mathcal{X}_0(\rho), \|\mathbf{x}\|_2 \geq \mu \tag{1.4}$$

In both cases, H_0 represents business as usual while H_1 encompasses all of the foreseeable anomalous activity. It is the composite nature of H_1 that causes theoretical difficulties.

It is imperative that we control both the probability of false alarm, and the false acceptance of the null. Let a test be a mapping $T(\mathbf{y}) \in \{0, 1\}$, where 1 indicates that we reject the null. To this end, we define our measure of risk to be

$$R(T) = \sup_{\mathbf{x} \in H_0} \mathbb{E}_{\boldsymbol{\epsilon}}[T] + \sup_{\mathbf{x} \in H_1} \mathbb{E}_{\boldsymbol{\epsilon}}[1 - T]$$

where $\mathbb{E}_{\boldsymbol{\epsilon}}$ denotes the expectation with respect to $\mathbf{y} \sim N(\mathbf{x}, \sigma^2 \mathbf{I}_p)$. The test T may be randomized, in which case the risk is $\mathbb{E}_T R(T)$. Notice that if the distribution of the random test T is independent of $\mathbf{x}$, then $\mathbb{E}_T \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathbf{x}}[1 - T] = \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{T,\mathbf{x}}[1 - T]$. This is the setting of [3] which we should contrast to the Bayesian setup in [1]. We will say that H_0 and H_1 are *asymptotically distinguished* by a test, T, if $\lim_{n \to \infty} R(T) = 0$. If such a test exists then H_0 and H_1 are asymptotically distinguished, otherwise they are asymptotically indistinguishable. If it is the case that there is an SNR scaling (μ/σ) for which there is a test that asymptotically distinguishes H_0 from H_1 and for any SNR sequence of lower order, H_0 and H_1 are asymptotically indistinguishable by that test, then we call this the *critical SNR* for the test.

1.2 Related Work

In Chapter 2, we will analyze the asymptotic statistical performance of projection estimators on graphs, with both a frequentist result and Bayesian corollary for the Ising prior. Much work has been devoted to the use of shrinkage estimators for ellipsoid constraints, resulting in the asymptotic optimality for the Pinsker estimator for low-noise asymptotics [8, 16, 40]. Invariably efficient estimators with ellipsoid constraints shrink components of the observed vector that correspond to minor axes of the ellipsoid. Other shrinkage procedures have been studied extensively, such as the projection estimator and Tikhonov regularization[79]. For simplicity we study the performance of projection estimators, but characterize their statistical consistency through spectral graph theory.

Markov random fields (MRF) provide a succinct framework in which the underlying signal is modeled as a draw from an Ising or Potts model [12, 63]. Most work on MRFs suggest the use of the maximum a posteriori (MAP) estimator which is the Bayes rule under $0/1$ loss. Less is known about the Bayes rule under Hamming distance loss, in which the estimator is the posterior centroid [11], a procedure known to be computationally intractable. A similar line of research is the use of kernels over graphs. The study of kernels over graphs began with the development

of diffusion kernels [44], and was extended through Green's functions on graphs [74]. A related body of work extends marginalized kernels to graphs [42, 56], while recently it has been shown that this and the aforementioned definitions are members of an overarching framework with computationally efficient constructions [80]. While kernels on graphs provide computationally efficient procedures for inference on graphs, much less is known about their asymptotic statistical efficiency.

We find that statistical consistency of specific shrinkage estimators imply that the Laplacian eigenbasis gives statistically efficient representations of graph structured patterns. There have been several attempts at constructing multi-scale basis for graphs that can efficiently represent localized activation patterns, notably diffusion wavelets [14] and treelets [47], however their approximation capabilities are not well understood. [73] and [30] independently proposed unbalanced Haar wavelets and characterized their approximation properties for tree-structured binary patterns.

In Chapter 3, we study a total variation denoising procedure, called the edge lasso, over general graph structures. The edge lasso is a generalization of the fused lasso originally proposed in [77] to enable recovery of high-dimensional patterns which are smooth (piece-wise constant) on a graph. The key idea is to penalize the ℓ_1-norm of differences of measurements at vertices that share an edge to encourage sparsity of the edges which connect vertices that have different signal values. While there have been some attempts [34, 52, 78] at coming up with efficient algorithms for solving the fused lasso optimization, a theoretical analysis of its performance is mostly lacking. The only exception, to the best of our knowledge, are [33, 64] which analyze the linear graph topology.

In Chapters 4 and 5, we consider statistical tests for the graph structured normal means problem. Normal means testing in high-dimensions is a well established and fundamental problem in statistics. Much is known when H_1 derives from a smooth function space such as Besov spaces or Sobolev spaces[35, 36]. Only recently has combinatorial structures such as graphs been proposed as the underlying structure of H_1. A significant portion of the recent work in this area [1, 3, 4, 5] has focused on incorporating structural assumptions on the signal, as a way to mitigate the effect of high-dimensionality and also because many real-life problems can be represented as instances of the normal means problem with graph-structured signals (see, for an example, [37]).

Another line of research relevant to the problem posed in Chapter 4 is optimal fail detection with nuisance parameters and matched subspace detection in the signal processing literature: see, e.g. [7, 25, 26, 67]. Though our problem can be cast as a special case of the more general problem of optimal testing of a linear subspace with nuisance parameters, the focus on a graph-structured signal, as well as the type of analysis based on the interplay between the scan statistics and the spectral properties of the graph contained in this work, is novel.

Chapter 2

Laplacian Eigenmaps for Estimation and Localization

This chapter is devoted to the analysis of an estimator based on the graph Laplacian eigenbasis, and establish conditions for ℓ_2 consistency when latent patterns arise from the Ising and GGM models in (1.2) or when the pattern is ℓ_2 graph structured. For both deterministic and probabilistic network models, the results indicate that by leveraging the network interaction structure, it is possible to consistently recover high-dimensional patterns even when the noise variance relative to the dimension with network size. Below is a summary of the contributions that can be found in the subsections. A more detailed analysis with proofs can be found in [71].

1. **Main result:** under the Ising, GGM priors and ℓ_2 graph structure the ℓ_2 risk of Laplacian eigenmaps projection is bounded by a function of the Laplacian eigenvalues.
2. **Hierarchical block structure:** under the hierarchical block structure, the Laplacian eigenvectors gives the Haar wavelet basis, and we can achieve ℓ_2 consistency with polynomially growing noise variance.
3. **Regular torus structure:** for a torus of increasing dimension, Laplacian eigenmaps achieves ℓ_2 consistency with polynomially growing noise variance.
4. **Random graph structure:** for the supercritical Erdös-Rényi, we can achieve ℓ_2 consistency with polynomially growing noise variance.
5. **Estimation implies localization:** Given any estimator that is ℓ_2 consistent, we can show that it generates an estimator of $\mathrm{supp}(\mathbf{x})$ that is localization consistent.

2.1 Main Result for Laplacian Eigenmaps

If the network activation patterns are generated by a Gaussian graphical model, it is easy to see that the eigenvalues of the Laplacian (inverse covariance) determine the MSE decay. Consider the GGM prior as in (1.2), then the posterior distribution is

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left((2\sigma^2\Delta + \mathbf{I})^{-1}\mathbf{y}, \left(2\Delta + \sigma^{-2}\mathbf{I}\right)^{-1}\right), \tag{2.1}$$

where $\mathbf{I}$ is the identity matrix. The posterior mean is the Bayes optimal estimator with Bayes MSE, $\sum_{i\in[p]}(2\lambda_i + \sigma^{-2})^{-1}$, where $\{\lambda_i\}_{i\in[p]}$ are the ordered eigenvalues of Δ. The binary Ising model is essentially a discrete version of the GGM, however, the Bayes rule and risk for the Ising model have no known closed form. For binary graph structured patterns drawn from an Ising prior, we suggest a different estimator based on projections onto the graph Laplacian eigenbasis. Recall that the graph Laplacian Δ has spectral decomposition, $\Delta = \mathbf{U}\Lambda\mathbf{U}^\top$, and denote the first k eigenvectors (corresponding to the smallest eigenvalues) of Δ by $\mathbf{U}_{[k]}$. Define the estimator

$$\widehat{\mathbf{x}}_k = \mathbf{U}_{[k]}\mathbf{U}_{[k]}^\top \mathbf{y}, \tag{2.2}$$

which is a hard thresholding of the projection of network measurements onto the graph Laplacian eigenbasis. The following theorem bounds the MSE of this estimator.

Theorem 1. *1. The maximum MSE of the estimator in* (2.2) *for the observation model in* (1.1)*, when the activation patterns satisfy $\mathbf{x}^\top\Delta\mathbf{x} \le \rho$ (i.e. lie within $\mathcal{X}_2(\rho)$) is bounded as*

$$R := \sup_{\mathbf{x}:\mathbf{x}^\top\Delta\mathbf{x}\le\rho} \mathbb{E}_{\boldsymbol{\epsilon}}\|\widehat{\mathbf{x}}_k - \mathbf{x}\|^2 \le \min\left(p, \frac{\rho}{\lambda_{k+1}}\right) + k\sigma^2$$

2. *The Bayes MSE of the estimator in* (2.2) *for the observation model in* (1.1)*, when the activation patterns are drawn from the GGM prior is bounded as*

$$R_B := \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}\|\widehat{\mathbf{x}}_k - \mathbf{x}\|^2 = \sum_{i=k+1}^{p} \frac{1}{2\lambda_i} + k\sigma^2 \le \frac{p}{2\lambda_{k+1}} + k\sigma^2$$

3. *The Bayes MSE when the binary activation patterns are drawn from the Ising prior is bounded as*

$$R_B := \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}\|\widehat{\mathbf{x}}_k - \mathbf{x}\|^2 \le \min\left(p, \frac{p\delta}{\lambda_{k+1}}\right) + k\sigma^2 + pe^{-p}$$

where $0 < \delta < 2$ is a constant and λ_{k+1} is the $(k+1)^{th}$ smallest eigenvalue of Δ.

Proof. We will prove (3) as (1) and (2) follow from similar considerations. First, we argue that with high probability, $\mathbf{x}^T\Delta\mathbf{x} \le \delta p$, where $0 < \delta < 2$ is a constant. Let $\Omega = \{\mathbf{x} : \mathbf{x}^T\Delta\mathbf{x} \le \delta p\}$ and $\bar{\Omega}$ denotes its complement. By Markov's inequality, for $t > 0$,

$$\mathbb{P}\{\mathbf{x}^T\Delta\mathbf{x} > K\} = \mathbb{P}\{e^{t\mathbf{x}^T\Delta\mathbf{x}} > e^{tK}\} \le e^{-tK}\,\mathbb{E}e^{t\mathbf{x}^T\Delta\mathbf{x}}$$

Let ν denote the uniform distribution over $\{0,1\}^p$ and $N(\Delta) = \int \nu(dx)e^{-\mathbf{x}^T\Delta\mathbf{x}}$. Then,

$$\mathbb{E}e^{\mathbf{x}^T(t\Delta)\mathbf{x}} = \int \nu(dx)N(\Delta)^{-1}e^{-\mathbf{x}^T\Delta\mathbf{x}}e^{\mathbf{x}^T(t\Delta)\mathbf{x}} = \frac{\int \nu(dx)e^{-\mathbf{x}^T(1-t)\Delta\mathbf{x}}}{N(\Delta)} = \frac{N((1-t)\Delta)}{N(\Delta)} \le 2^p$$

where the last step follows since $N(\Delta) = \sum_{x\in\{0,1\}^p} e^{-x^T\Delta x}$ and $\Delta\mathbf{1} = 0$ implying that $1 \le N(\Delta), N((1-t)\Delta) \le 2^p, \forall t \in (0,1)$. This gives us the Chernoff-type bound,

$$\mathbb{P}(\bar{\Omega}) \le \mathbb{P}\{\mathbf{x}^T\Delta\mathbf{x} > K\} \le e^{-tK}2^p = e^{(\log 2 - tK/p)p} \le e^{-p}$$

by setting $K = \delta p$ and $\delta = \frac{1+\log 2}{t}$. If we choose $t < \frac{1+\log 2}{2}$ then $\delta < 2$.

Let $\mathbf{u}_i$ denote the i^{th} eigenvector of the graph Laplacian Δ, then under this orthonormal basis,

$$\mathbb{E}[\|\widehat{\mathbf{x}}_k - \mathbf{x}\|^2] \leq \mathbb{E}[\sum_{i=k+1}^{p} \mathbf{u}_i^T \mathbf{x}^2 |\, \Omega] + pP(\bar{\Omega}) + k\sigma^2 \leq \sup_{\mathbf{x}:\mathbf{x}^T\Delta\mathbf{x}\leq\delta p} \sum_{i=k+1}^{p} \mathbf{u}_i^T \mathbf{x}^2 + p\, e^{-p} + k\sigma^2.$$

We now establish that $\sup_{\mathbf{x}:\mathbf{x}^T\Delta\mathbf{x}\leq\delta p} \sum_{i=k+1}^{p} (\mathbf{u}_i^T\mathbf{x})^2 \leq p\min(1, \delta/\lambda_{k+1})$, and the result follows. Let $\tilde{\mathbf{x}}_i = \mathbf{u}_{i+k}^T\mathbf{x}, i \in [p-k]$ and note that $\mathbf{x}^T\Delta\mathbf{x} = \sum_{i=1}^{p} \lambda_i (\mathbf{u}_i^T\mathbf{x})^2 \geq \sum_{i=k+1}^{p} \lambda_i \tilde{\mathbf{x}}_i^2$, for λ_i the i^{th} eigenvalue of Δ. Consider the primal problem,

$$\max \sum_{j=1}^{p-k} \tilde{\mathbf{x}}_j^2 \text{ such that } \sum_{j=1}^{p-k} \lambda_j \tilde{\mathbf{x}}_j^2 \leq \delta p,\ \tilde{\mathbf{x}} \in \mathbb{R}^{p-k}$$

Note that $\mathbf{x}$ contained within the ellipsoid $\mathbf{x}^T\Delta\mathbf{x} \leq \delta p$, $\mathbf{x} \in \{0,1\}^p$ implies that $\tilde{\mathbf{x}}$ is feasible, so a solution to the optimization upper bounds $\sup_{\mathbf{x}:\mathbf{x}^T\Delta\mathbf{x}\leq\delta p} \sum_{i=k+1}^{p} (\mathbf{u}_i^T\mathbf{x})^2$. By forming the dual problem, we find that the solution, $\mathbf{x}^*$, to the primal problem attains a bound of $||\tilde{\mathbf{x}}||^2 \leq ||\tilde{\mathbf{x}}^*||^2 = \delta p/\lambda_{k+1}$. Also, $||\tilde{\mathbf{x}}||^2 \leq ||\mathbf{x}||^2 \leq p$, so we obtain the desired bound. □

Through this bias-variance decomposition, we see the eigenspectrum of the graph Laplacian determines a bound on the MSE for binary graph-structured activations.

Remark 2. *Consider the binarized estimator $\widehat{\mathbf{x}}_i' = \mathbf{1}_{\widehat{\mathbf{x}}_i > 1/2}, i \in [p]$. Then the results of Theorem 1 (3) also provide an upper bound on the expected Hamming distance of this new estimator since $\mathbb{E}[d_H(\widehat{\mathbf{x}}', \mathbf{x})] = MSE(\widehat{\mathbf{x}}') \leq 4MSE(\widehat{\mathbf{x}})$, by the triangle inequality (this is similar to Proposition 9).*

2.2 Asymptotic performance under specific graph models

We now discuss the eigenspectrum of some simple graphs and use the MSE bounds derived in the previous section to analyze the amount of noise that can be tolerated while ensuring consistent MSE recovery of high-dimensional patterns. In all these examples, we find that the tolerable noise level scales as $\sigma^2 = o(p^{\gamma-1})$, where $\gamma \in (0,1)$ characterizes the strength of network interactions. Notice that without any structural assumptions naive estimators can only tolerate noise levels that scale like $\sigma^2 \asymp 1/p$ for ℓ_2-consistency. To compare these results to those in the later chapters the signal size $\|\mathbf{x}\|^2 \asymp p$, and so the signal-to-noise ratio is $\sqrt{p}/\sigma$.

2.2.1 Hierarchical structure

Consider that, under an appropriate permutation of rows and columns, the weight matrix $\mathbf{W}$ has the hierarchical block form shown in Figure 2.1. This corresponds to hierarchical graph structured dependencies between node variables, where $\epsilon_\ell > \epsilon_{\ell+1}$ denote the strength of interactions between nodes that are in the same block at level $\ell = 0, 1, \ldots, L$. We find that in this case the eigenvectors $\mathbf{U}$ of the graph Laplacian correspond to unbalanced Haar wavelet basis (proposed in [30, 73]). Using the bound on MSE as given in Theorem 1, we can now derive the noise threshold that allows for consistent MSE recovery of high-dimensional patterns as the network size $p \to \infty$.

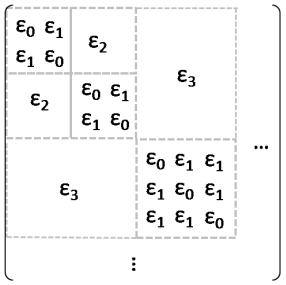


Figure 2.1: Weight matrices corresponding to hierarchical dependencies between node variables.

Corollary 3. *Consider a graph-structured pattern under any of the conditions of Theorem 1 (1)-(3) under the hierarchical block graph. If* $\epsilon_\ell = 2^{-\ell(1-\beta)}\ \forall \ell \leq \gamma \log_2 p + 1$*, for constants* $\gamma, \beta \in (0,1)$*, and* $\epsilon_\ell = 0$ *otherwise, then the SNR threshold for consistent MSE recovery* $(R, R_B = o(1))$ *is*

$$\frac{\sqrt{p}}{\sigma} = \omega(p^{\frac{1-\gamma}{2}}).$$

Proof. Let $\ell^* = (1-\gamma)\log_2 p$. Since $\epsilon_i = 2^{-i(1-\beta)}\ \forall i < L - \ell^* + 1$ and $\epsilon_i = 0$ otherwise, we have for $\ell \geq \ell^*$ and since $L = \log_2 p$, $\lambda_\ell \geq 2^{\beta(L-\ell^*)}2^{\beta-1} = p^{\beta\gamma}2^{\beta-1}$, which is increasing in p. Therefore, we can pick $k = 2^{\ell^*}$ and since $2^{\ell^*}/p = p^{-\gamma}$, the result follows. □

What we find then is that under the heirarchical structure we are able to tolerate an SNR scaling like $p^{\frac{1-\gamma}{2}}$ where in the naive setting we can only tolerate $\sqrt{p}$, thus gaining by a factor of $p^{\gamma/2}$.

2.2.2 Regular torus structure

Now consider the torus graph which is constructed by placing vertices in a regular grid on a ℓ dimensional torus and adding edges of weight 1 to adjacent points. Let $p = r^\ell$. For $\ell = 1$ this is a cycle which has a circulant weight matrix w, with eigenvalues $\{2\cos(\frac{2\pi k}{p}) : k \in [p]\}$ and eigenvectors correspond to the discrete Fourier transform [27]. Let $i = (i_1, ..., i_\ell), j = (j_1, ..., j_\ell) \in [r]^\ell$. Then the weight matrix of the torus in ℓ dimensions is

$$W_{i,j} = w_{i_1,j_1}\delta_{i_2,j_2}...\delta_{i_\ell,j_\ell} + ... + w_{i_\ell,j_\ell}\delta_{i_1,j_1}...\delta_{i_{\ell-1},j_{\ell-1}} \tag{2.3}$$

where δ is the Kronecker delta function. Through concentration of the eigenspectrum, we can choose k such that $\lambda_k \geq \ell$ and $k = \lceil pe^{-\ell/8} \rceil$. So, the risk bound becomes $\mathcal{O}(2p/\ell + p\sigma^2 e^{-\ell/8} + pe^{-p})$, and as we increase dimensions of the torus the MSE decays linearly.

Corollary 4. *Consider a graph-structured pattern under any of the conditions of Theorem 1 (1)-(3) based on a torus graph in* ℓ *dimensions with* $p = r^\ell$ *vertices. If* r *is a constant and* $\ell = 8\gamma \ln r$*, for some constant* $\gamma \in (0,1)$*, then the SNR threshold for consistent MSE recovery* $(R, R_B = o(1))$ *is given as:*

$$\frac{\sqrt{p}}{\sigma} = \omega(p^{\frac{1-\gamma}{2}}).$$

Again, the noise variance can increase with the network size p, and larger γ implies stronger network interactions as each variables interacts with more number of neighbors (ℓ is larger).

Proof. The primary mechanism that will allow us to use Theorem 1 is a concentration result for the eigenvalue distribution of Δ. This result is highly dependent on the fact that the dimension of the torus ℓ must be increasing.

Lemma 5. *Let $\lambda_\bullet^\Delta$ be an eigenvalue of the Laplacian, Δ, of the torus graph in ℓ dimensions with $p = r^\ell$ vertices, chosen uniformly at random. Then*

$$\mathbb{P}\{\lambda_\bullet^\Delta \leq \ell\} \leq \exp\{-\ell/8\}. \tag{2.4}$$

What follows is the proof sketch of Lemma 5. If $v_1, ..., v_\ell$ are a subset of the eigenvectors of w with eigenvalues $\lambda_1, ..., \lambda_\ell$, then $W(v_1 \otimes ... \otimes v_\ell) = (\lambda_1 + ... + \lambda_\ell)(v_1 \otimes ... \otimes v_\ell)$ where $\otimes$ denotes tensor product. Noting that the $D_{ii} = 2\ell, \forall i \in [r]^\ell$ then we see that the Laplacian Δ has eigenvalues $\lambda_i^\Delta = 2\ell - \lambda_i^W = \sum_j^{[\ell]}(2 - \lambda_{i_j}^w)$ for all $i \in [r]^\ell$. Recall $\lambda_k^w = 2\cos(\frac{2\pi k}{n})$ for some $k \in [r]$. Let i be distributed uniformly over $[r]^\ell$. Then $\mathbb{E}[\lambda_{i_j}^w] = 0$, and by Hoeffding's inequality,

$$\mathbb{P}\{\sum_{j=1}^{\ell}(2 - \lambda_{i_j}^w) - 2\ell \leq -t\} \leq \exp\{-2t^2/16\ell\}$$

So, using $t = \ell$ we get that $\mathbb{P}\{\sum_{j=1}^{\ell}(2 - \lambda_{i_j}^w) \leq \ell\} \leq \exp\{\frac{-\ell}{8}\}$ and the result follows. □

2.2.3 Erdös-Rényi random graph structure

Erdös-Rényi (ER) random graphs are generated by adding edges with weight 1 between any two vertices within V with probability q_p. It is known that the probability of edge inclusion (q_p) determines large geometric properties of the graph [17]. Real world networks are generally sparse, so we set $q_p = p^{-(1-\gamma)}$, where $\gamma \in (0, 1)$. Larger γ implies higher probability of edge inclusion and stronger network interaction structure. Using the degree distribution, and a result from perturbation theory, we bound the quantiles of the eigenspectrum of Δ. This enables us to set the sequence of quantiles for the eigenvalue distribution $k = \lceil \alpha_p p^{1-\gamma} \rceil$ such that $\mathbb{P}_G\{\lambda_k \leq p^\gamma/2 - p^{\gamma-1}\} = \mathcal{O}(1/\alpha_p)$. So, we obtain a bound for the expected Bayes MSE (with respect to the graph) $\mathbb{E}_G[R] \leq \mathcal{O}(p^{1-\gamma}) + \sigma^2\mathcal{O}(\alpha_p p^{1-\gamma}) + \mathcal{O}(p/\alpha_p)$.

Corollary 6. *Consider a graph G drawn from an Erdös-Rényi random graph model with p vertices and probability of edge inclusion $q_p = p^{-(1-\gamma)}$ for some constant $\gamma \in (0, 1)$. For a graph-structured pattern under any of the conditions of Theorem 1 (1)-(3), the noise variance that can be tolerated while ensuring consistent MSE recovery ($R, R_B = o_{\mathbb{P}_G}(1)$) is given as:*

$$\frac{\sqrt{p}}{\sigma} = o(p^{\frac{1-\gamma}{2}}).$$

Proof. Using the degree distribution [9], and a result from perturbation theory, we bound the quantiles of the eigenspectrum of Δ.

Lemma 7. *Let $\lambda_\bullet$ denote an eigenvalue of Δ chosen uniformly at random. Let $\mathbb{P}_G$ be the probability measure induced by the ER random graph and $\mathbb{P}_\bullet$ be the uniform distribution over eigenvalues conditional on the graph. Then, for any α_p increasing in p,*

$$\mathbb{P}_G\{\mathbb{P}_\bullet\{\lambda_\bullet \leq p^\gamma/2 - p^{\gamma-1}\} \geq \alpha_p p^{-\gamma}\} = \mathcal{O}(1/\alpha_p) \tag{2.5}$$

Proof of Lemma 7: We introduce a random variable $\bullet$ that is uniform over $[p]$. Note that, conditioned on this random variable, $d_\bullet \sim \text{Binomial}(p-1, q_p)$ and $\text{Var}(d_\bullet) \leq pq_p$. We decompose the Laplacian, $\Delta = \mathbf{D} - \mathbf{W} = (\bar{d}\mathbf{I} - \mathbf{W}) + (\mathbf{D} - \bar{d}\mathbf{I})$, into the expected degree of each vertex ($\bar{d} = (p-1)q_p$), $\mathbf{W}$ and the deviations from the expected degree and use the following lemma.

Lemma 8 (Wielandt-Hoffman Theorem)**.** *[38, 85] Suppose $A = B + C$ are symmetric $p \times p$ matrices and denote the ordered eigenvalues by $\{\lambda_i^A, \lambda_i^B\}_{i=1}^p$. If $||.||_F$ denotes the Frobenius norm,*

$$\sum_{i=1}^{p} (\lambda_i^A - \lambda_i^B)^2 \leq ||C||_F^2 \tag{2.6}$$

Notice that $\mathbb{E}_G||\mathbf{D} - \bar{d}\mathbf{I}||_F^2/p = \text{Var}(d_\bullet)$ and so $\mathbb{E}_G||\lambda^{\bar{d}\mathbf{I}-\mathbf{W}} - \lambda^L||^2/p \leq pq_p = p^\gamma$ *(i)*. Also, it is known that for $\gamma \in (0,1)$ the eigenvalues converge to a semicircular distribution[20] such that $\mathbb{P}_G\{|\lambda_\bullet^W| \leq 2\sqrt{pq_p(1-q_p)}\} \to 1$. Since $2\sqrt{pq_p(1-q_p)} \leq 2p^{\gamma/2}$, we have $\mathbb{E}_G[(\lambda_\bullet^W)^2] \leq 4p^\gamma$ for large enough p *(ii)*. Using triangle inequality,

$$\mathbb{E}_G[(\lambda_\bullet^L - (p-1)q_p)^2] \leq \mathbb{E}_G[(\lambda_\bullet^L - ((p-1)q_p - \lambda_\bullet^W))^2] + \mathbb{E}_G[(\lambda_\bullet^W)^2] \leq 5p^\gamma, \tag{2.7}$$

where the last step follows using *(i)*, *(ii)* and $\lambda_i^{\bar{d}\mathbf{I}-\mathbf{W}} = (p-1)q_p - \lambda_i^W$. By Markov's inequality,

$$\mathbb{P}_G\{\mathbb{P}_\bullet\{\lambda_\bullet^L \leq \frac{p^\gamma}{2} - p^{\gamma-1}\} \geq \alpha_p p^{-\gamma}\} \leq \frac{p^\gamma}{\alpha_p}\mathbb{E}_G[\mathbb{P}_\bullet\{\lambda_\bullet^L \leq \frac{p^\gamma}{2} - p^{\gamma-1}\}] \tag{2.8}$$

for any α_p which is an increasing positive function in p. We now analyze the right hand side.

$$\mathbb{P}_\bullet\{|\lambda_\bullet^L - (p-1)q_p| \geq \epsilon\} \leq \epsilon^{-2}\mathbb{E}_\bullet[(\lambda_\bullet^L - (p-1)q_p)^2]$$

Note that $\mathbb{P}_\bullet\{\lambda_\bullet^L \leq pq_p - q_p - \epsilon\} \leq \mathbb{P}_\bullet\{|\lambda_\bullet^L - (p-1)q_p| \geq \epsilon\}$ and setting $\epsilon = pq_p/2 = p^\gamma/2$,

$$\mathbb{P}_\bullet\{\lambda_\bullet^L \leq p^\gamma/2 - p^{\gamma-1}\} \leq 4p^{-2\gamma}\mathbb{E}_\bullet[(\lambda_\bullet^L - (p-1)q_p)^2].$$

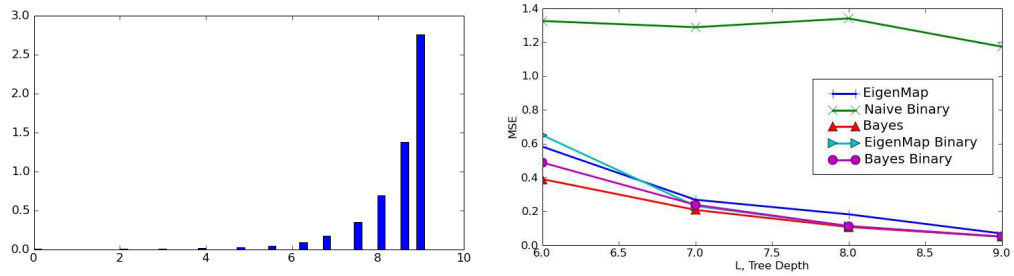
Hence, we are able to complete the lemma, such that for p large enough, using Eqs. (2.8) and (2.7)

$$\mathbb{P}_G\{\mathbb{P}_\bullet\{\lambda_\bullet^L \leq \frac{p^\gamma}{2} - p^{\gamma-1}\} \geq \alpha_p p^{-\gamma}\} \leq \frac{4}{\alpha_p p^\gamma}\mathbb{E}_G[\mathbb{E}_\bullet[(\lambda_\bullet^L - (p-1)q_p)^2]] \leq \frac{20}{\alpha_p}. \tag{2.9}$$

Proof sketch of Corollary 6: By lemma 7 and appropriately specifying the quantiles,

$$\mathbb{E}_G R_B/p \leq \mathbb{E}_G\left[\frac{2}{\lambda_{k_p}} + \sigma^2\frac{k_p}{p} + e^{-p}\right] \leq \left(\frac{2}{p^\gamma/2 - p^{\gamma-1}} + \sigma^2\mathcal{O}(\alpha_p p^{-\gamma}) + e^{-p}\right) + \mathcal{O}(\frac{1}{\alpha_p}) \tag{2.10}$$

Note that we have the freedom to choose $\alpha_p = \sqrt{p^\gamma/\sigma^2}$ making $\sigma^2\mathcal{O}(\alpha_p p^{-\gamma}) = \mathcal{O}(\sqrt{\sigma^2/p^\gamma}) = o(1)$ and $\mathcal{O}(1/\alpha_p) = o(1)$ if $\sigma^2 = o(p^\gamma)$. □

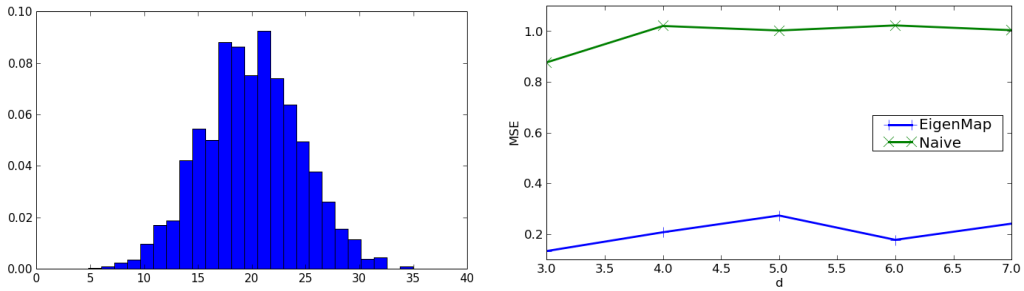


(a) Eigenvalue Histogram for hierarchical tree. (b) Estimator Performance

Figure 2.2: The eigenvalue histogram for the binary tree, $L = 11$, $\beta = .1$ (left) and the performance of various estimators (right) with $\beta = 0.05$ and $\sigma^2 = 4$, both with $\gamma = 1$.

2.3 Experiments

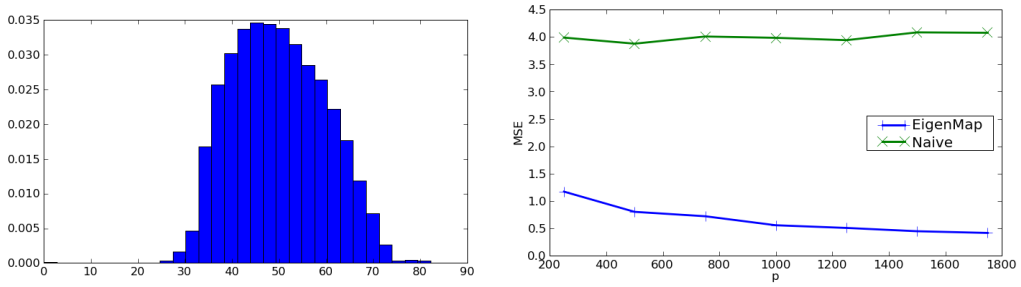
We simulate patterns from the Ising model defined on hierarchical, torus and ER graphs. Since the Ising distribution admits a closed form for the distribution of one node conditional on the rest of the nodes, a Gibbs sampler can be employed. Histograms of the eigenspectrum for the hierarchical tree graph with a large depth, the torus graph in high dimensions, and a draw from the ER graph with many nodes is shown in figures 2.2(a), 2.3(a), 2.4(a) respectively. The eigenspectrum of the torus and ER graphs illustrate the concentration of the eigenvalues about the expected degree of each node.



(a) Eigenvalue Histogram for Torus. (b) Estimator Performance

Figure 2.3: The eigenvalue histogram for the torus with $\ell = 10$ and $p = 5^{10}$ (left) and estimator performances (right) with $p = 3^\ell$ and $\sigma^2 = 1$. Notice that the eigenvalues concentrate around 2ℓ.

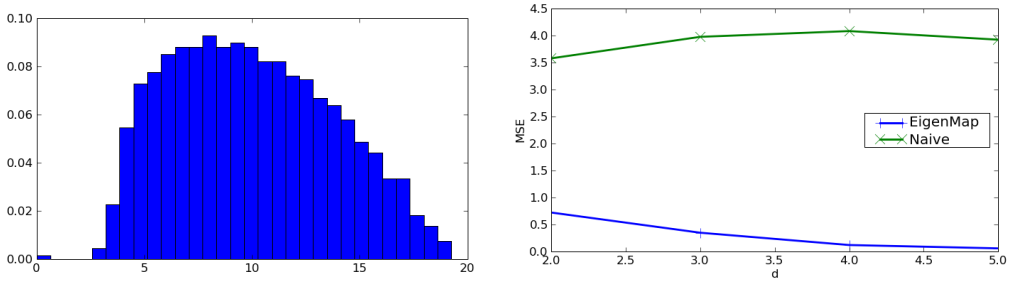
We use iterative eigenvalue solvers to form our estimator and choose the quantile k by minimizing the bound in Theorem 34. We compute the average Bayes MSE, R_B/p, (by taking multiple draws) of our estimator for a noisy sample of node measurements. We observe in all of the models that the eigenmap estimator is a substantial improvement over Naive (the Bayes estimator that ignores the structure). See Figures 2.2(b), 2.3(b), 2.4(b). For the hierarchical model, we also sample from the posterior using a Gibbs sampler and estimate the posterior mean (Bayes rule under MSE). We find that the posterior mean is only a slight improvement over the eigenmap estimator (Figure 2.2(b)), despite it's difficulty to compute. Also, a binarized version of these estimators does not substantially change the MSE.



(a) Eigenvalue Histogram for Erdös-Rényi.

(b) Estimator Performance

Figure 2.4: The eigenvalue histogram for a draw from the ER graph with $p = 2500$ and $q_p = p^{-.5}$ (left) and the estimator performances (right) with $q_p = p^{-.75}$ and $\sigma^2 = 4$. Notice that the eigenvalues are concentrated around p^γ where $q_p = p^{-(1-\gamma)}$.



(a) Eigenvalue Histogram for Watts-Strogatz.

(b) Estimator Performance

Figure 2.5: The eigenvalue histogram for a draw from the Watts-Strogatz graph with $d = 5$ and $p = 4^5$ with 0.25 probability of rewiring (left) and estimator performances (right) with 4^ℓ vertices and $\sigma^2 = 4$. Notice that the eigenvalues are concentrated around 2ℓ.

We also simulate graphs from the Watts-Strogatz 'small world' model [84], which is known to be an appropriate model for self-organizing systems such as biological systems and human networks. The 'small world' graph is generated by forming the torus graph, then rewiring each edge with some constant probability to another vertex uniformly at random such that loops are never created. We observe that the eigenvalues concentrate (more tightly than the torus graph) around the expected degree 2ℓ (Figure 2.5(a)) and note that, like the ER model, the eigenspectrum converges to a nearly semi-circular distribution [20]. Similarly, the MSE decays in a fashion similar to the ER model (Figure 2.5(b)).

2.4 Estimation Implies Localization

ℓ_2 consistency may seem superficial, as in some situations $\mathbf{0}$, may have diminishing mean square error. We will now see that if the mean vector has the form $\mathbf{x} \propto \mathbf{1}_C$ then ℓ_2 consistency (estimation) implies localization consistency of C. Recall that localization means that we have an estimate $\hat{C}$ such that $|C \cap \hat{C}|$ is small relative to $\sqrt{|\hat{C}||C|}$. Suppose that we obtain an estimator that has a low MSE ($\|\mathbf{x} - \hat{\mathbf{x}}\|$) then we immediately can construct an estimate of C.

$$\hat{C} = \arg\max_{C \subseteq [n]} \hat{\mathbf{x}}^\top \frac{\mathbf{1}_C}{\sqrt{|C|}} \tag{2.11}$$

and define $\hat{\mathbf{x}}_T = \|\hat{\mathbf{x}}\|\mathbf{1}_{\hat{C}}/\sqrt{|\hat{C}|}$. This estimate is easily computed by greedily including the largest components of $\hat{\mathbf{x}}$ in $\hat{C}$ until the objective is maximized. Then we know that

$$\hat{\mathbf{x}}^\top \frac{\mathbf{1}_C}{\sqrt{|C|}} \leq \hat{\mathbf{x}}^\top \frac{\mathbf{1}_{\hat{C}}}{\sqrt{|\hat{C}|}} \Rightarrow \|\hat{\mathbf{x}}_T - \hat{\mathbf{x}}\| \leq \|\hat{\mathbf{x}} - \mathbf{x}\|$$

Hence, we have the following MSE

$$\|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 \leq (\|\hat{\mathbf{x}}_T - \hat{\mathbf{x}}\| + \|\hat{\mathbf{x}} - \mathbf{x}\|)^2 \leq 4\|\hat{\mathbf{x}} - \mathbf{x}\|^2$$

$$\|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 = \mu^2 \left(\frac{\|\hat{\mathbf{x}}_T\|^2}{\mu^2} + 1 - 2\frac{|\hat{C} \cap C|}{\sqrt{|C||\hat{C}|}} \right)$$

$$\mu^2 d(\hat{C}, C) = \|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 + \mu^2(1 - \frac{\|\hat{\mathbf{x}}_T\|^2}{\mu^2}) \leq \|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 + |\|\mathbf{x}\|^2 - \|\hat{\mathbf{x}}_T\|^2|$$

$$\leq \|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 + \|\hat{\mathbf{x}}_T - \mathbf{x}\|(\|\mathbf{x}\| + \|\hat{\mathbf{x}}_T\|) \leq 2(\|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 + \mu\|\hat{\mathbf{x}}_T - \mathbf{x}\|)$$

We summarize our result in the following,

Proposition 9. *Consider $\hat{C}$ constructed from $\hat{\mathbf{x}}$ according to* (2.11).

$$d(\hat{C}, C) \leq \frac{1}{\mu^2}(\|\hat{\mathbf{x}}_T - \mathbf{x}\|^2 + |\|\mathbf{x}\|^2 - \|\hat{\mathbf{x}}_T\|^2|) \leq 2(\frac{4}{\mu^2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 + \frac{2}{\mu}\|\hat{\mathbf{x}} - \mathbf{x}\|) \tag{2.12}$$

Of course this means that if the conditions, particularly the SNR scaling, are such that $\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\| \to 0$ then $\mathbb{E}d(C, C') \to 0$.

This has obvious implications for the preceeding Sections, because any $\mathbf{x}$ drawn from the Ising model can be written as $\mathbf{1}_C$. We will see that this can directly port some classical estimators to apply to our localization techniques in Chapter 5.

2.5 Discussion

In this chapter, we have characterized the improvement in SNR threshold, above which ℓ_2 consistency of high-dimensional network activation patterns embedded in heavy noise is possible, as a function of the network size and parameters governing the statistical dependencies in the activation process. Our results indicate that by leveraging the network interaction structure, it is possible to tolerate noise with variance that increases with the size of the network whereas without exploiting dependencies in the node measurements, the noise variance needs to decrease as the network size grows to accommodate for multiple hypothesis testing effects.

Recall, that we have studied the estimation of a Gaussian mean under an ellipsoid constraint. There is some work in statistics along these lines, most notably the Pinsker estimator has been shown to be asymptically minimax [41]. The major difference, between our work and the aforementioned is that the setting in which they prove asymptotic minimaxity is when the ellipsoid is fixed in possibly infinite dimensions and the noise variance is decreasing. This is not applicable to the setting in which the graph may change as $p \to \infty$. In Chapter 4, we will compare our statistical test to one closely related to Pinsker's estimator (Ermakov's test).

One may notice that all of the examples above require that the degree of the vertices is increasing in p. This excludes such important cases as the two dimensional lattice and the balanced binary tree. We will set out to perform a more refined analysis that will show that such cases result in improved performance, under the balanced graph structure (Chapter 4). We will see that there are fundamental limits to what is detectable under the ℓ_0 graph structure (Chapter 5). In this case, we will see that improved SNR regimes require that the degree must be increasing.

While we have only considered MSE recovery, it is often possible to detect the presence of patterns in much heavier noise, even though the activation values may not be accurately recovered [39]. SNR thresholds for the detection of graph structured signals will be addressed in Chapters 4 and 5. In addition, the thresholding estimator based on the graph Laplacian eigenbasis can also be used in high-dimensional linear regression or compressed sensing framework to incorporate structure, in addition to sparsity, of the relevant variables. Another popular estimator is based on total-variation denoising, which we call the Edge lasso. We will be addressing this estimator in the following chapter.

Chapter 3

Edge Lasso for Changepoint Localization

The fused lasso was proposed recently to enable recovery of high-dimensional patterns which are piece-wise constant on a graph, by penalizing the ℓ_1-norm of differences of measurements at vertices that share an edge. While there have been some attempts at coming up with efficient algorithms for solving the fused lasso optimization, a theoretical analysis of its performance is mostly lacking except for the simple linear graph topology. In this section, we investigate changepoint sparsistency of fused lasso for general graph structures, i.e. its ability to correctly recover the exact support of piece-wise constant graph-structured patterns asymptotically (for large-scale graphs). To emphasize this distinction over previous work, we will refer to it as Edge Lasso.

We focus on the (structured) normal means setting, and our results provide necessary and sufficient conditions on the graph properties as well as the signal-to-noise ratio needed to ensure changepoint sparsistency. Let $A \in \mathcal{A}$ denote the maximal sets of vertices with constant activation (viz. $x_v = x_w, \forall v, w \in A$). We exemplify our results using simple graph-structured patterns, and demonstrate that in some cases fused lasso is changepoint sparsistent at very weak signal-to-noise ratios, which may scale as $\sqrt{(\log p)/|A|}$, where p is the number of vertices in the graph and A is the smallest element of $\mathcal{A}$ (see figure 3.1). In other cases, it performs no better than thresholding the difference of measurements at vertices which share an edge (which requires signal-to-noise ratio that scales as $\sqrt{\log p}$). The standard of recovery that we will study is the exact recovery of the boundary set, $\mathcal{B} = \text{supp}(\nabla \mathbf{x})$. All results and detailed proofs can be found in [70]. We summarize the current results regarding the edge lasso:

1. **Edge thresholding:** Generic chaining provides upper and lower bounds on the changepoint sparsistency of edge thresholding, a simple estimator, that provides a natural comparison for the edge lasso.
2. **Noiseless changepoint sparsistency:** There are combinatorial conditions under which the edge lasso is changepoint sparsistent given the noiseless model ($\boldsymbol{\epsilon} = \mathbf{0}$).
3. **Noisy changepoint sparsistency:** By combining the conditions for the noiseless model with concentration of measure and spectral graph theory, one can obtain conditions for changepoint sparsistency in the noisy model.
4. **Asymptotics for specific graphs:** The edge lasso fails to be changepoint sparsistent for the 1 and 2 dimensional lattice, while it obtains nearly oracle rates for the nested complete

graph.

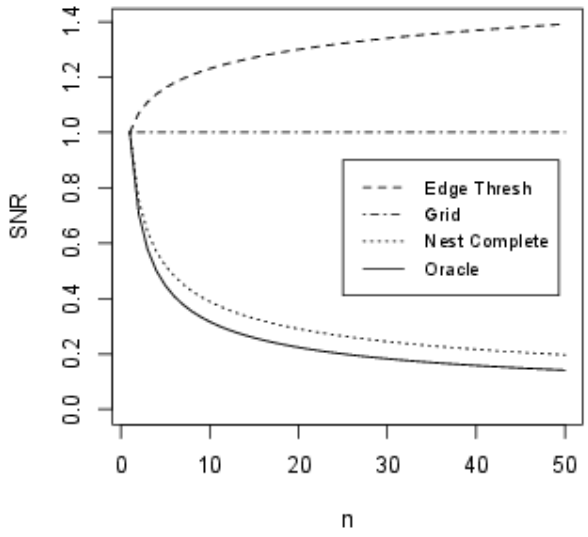


Figure 3.1: A qualitative summary of our changepoint sparsistency results for the SNR required by Edge thresholding, Edge lasso (for the 1-d, 2-d Grid and Nested Complete Graph), and an Oracle that has a priori knowledge of $\mathcal{A}$. (In the figure it is assumed that $|A|$ scales like p for all $A \in \mathcal{A}$.)

3.0.1 Mathematical Preliminaries

The following will provide the necessary tools from algebraic graph theory, for a more detailed exposition of these mathematical objects refer to [31]. For any oriented graph, the column space of ∇ is row$(\nabla^\top)$, and its orthogonal complement is null$(\nabla^\top)$. To cut a subset of the vertices, $A \subset V$, from the graph is to define A to be the positive shore and $\bar{A} = V \backslash A$ the negative shore. Now we define signed characteristic vectors of a cut to be $\chi(A)$ such that

$$\chi(A)_e = \begin{cases} +1 & , e^+ \in A, e^- \in \bar{A} \\ -1 & , e^- \in A, e^+ \in \bar{A} \\ 0 & , \text{otherwise} \end{cases}$$

Notice that $\{\chi(\{v\})\}_{v \in V}$ is precisely the columns of ∇, so by definition it forms a basis for row$(\nabla^\top)$. Moreover $\chi(A) = \sum_{v \in A} \nabla_v$ which we will use often. (We often subscript ∇ with e and v interchangeably where e means rows and v means columns.)

Let us introduce the signed characteristic vectors of cycles, $\chi(\phi)$ where ϕ is an ordered collection of vertices that form a cycle such that $(\phi_i, \phi_{i+1}) \in E$.

$$\chi(\phi)_e = \begin{cases} +1 & , e^+ = \phi_{i+1} \text{ and } e^- = \phi_i \\ -1 & , e^- = \phi_{i+1} \text{ and } e^+ = \phi_i \\ 0 & , \text{otherwise} \end{cases}$$

So, if an edge e is contained in the cycle then $\chi(\phi)_e = 1$ if the orientation of e is in the direction of the cycle and $\chi(\phi)_e = -1$ otherwise. Not only is it the case that null$(\nabla^\top)$ contains $\chi(\phi)$ for all cycles ϕ but it is spanned by all such $\chi(\phi)$. We will denote the projection onto these spaces as $\mathcal{P}_{\text{null}(\nabla^\top)}$ and $\mathcal{P}_{\text{row}(\nabla^\top)}$.

Also, let us denote the largest degree as $d_{\max}$. We also would like to note that the null space and row space (null(∇), row(∇)) is equal to the null and row space of Δ. The null space of ∇ is specifically the vectors that are constant over connected components of G. Furthermore, the projection onto the null space is obtained by averaging a vector within connected components. For an operator Φ define the Moore-Penrose pseudoinverse $\Phi^\dagger$. We will often use the operator norm,

$$|\!|\!|\Phi|\!|\!|_{2,\infty} = \sup_{\|b\|_2 \leq 1} \|\Phi b\|_\infty$$

where the norms ℓ_2 and ℓ_∞ are the Euclidean and max norms. To be clear we define $sign(0) = 0$. For a vector $\mathbf{z} \in \mathbb{R}^E$ and a non-empty set of edges $\mathcal{B} \subset E$, we will denote with $\mathbf{z}_\mathcal{B}$ the vector in $\mathbb{R}^E$ which agrees with $\mathbf{z}$ in the coordinates $\mathcal{B}$ and has zero entries in the coordinates in $\bar{\mathcal{B}}$. Similarly, for a matrix $\nabla \in \mathbb{R}^{V \times E}$, we will write $\nabla_\mathcal{B}$ for the matrix ∇ with the rows in $-\mathcal{B}$ replaced by zero vectors.

3.1 Edge Thresholding

It is natural as a first pass to merely difference observations $y_{e^+} - y_{e^-}$ and hard threshold to obtain an estimator of sign$(\nabla \mathbf{x})$ to achieve changepoint sparsistency. The estimator is given by,

$$\hat{z}_{th,e}(\tau) = (y_{e^+} - y_{e^-}) I\{|y_{e^+} - y_{e^-}| > \tau\} = (\nabla_e \mathbf{y}) I\{|\nabla_e \mathbf{y}| > \tau\}$$

We now characterize necessary and sufficient conditions to obtain changepoint sparsistency of edge thresholding.

Theorem 10. *Suppose that* $\frac{\|\nabla \mathbf{x}\|_0}{|E|} \to 0$ *for simplicity. Let* δ *be the smallest gap between the componentwise signal size between two maximal clusters,* $A, A' \in \mathcal{A}$*.*

1. *If* $\frac{\delta}{\sigma} = \omega(\sqrt{\log |E|})$ *then* $\hat{\mathbf{z}}_{th}$ *is changepoint sparsistent.*
2. *If* $\frac{\delta}{\sigma} = o(\sqrt{\log(|E| - \|\nabla \mathbf{x}\|_0)})$ *then* $\hat{\mathbf{z}}_{th}$ *is not changepoint sparsistent.*

Proof. We will assume that the noise is Gaussian with variance 1, by making the gap $\delta' = \delta/\sigma$. Let us construct the statistics: $z_e = y_{e^+} - y_{e^-}$ for each edge e. Now we are interested in the estimator $\hat{\mathcal{B}} = \{e : |z_e| > \tau\}$. We can show (1) with Markov's inequality and Gaussian concentration,

$$\mathbb{P}\{\hat{\mathcal{B}} \neq \mathcal{B}\} = \mathbb{P}\{\inf_{e \in \mathcal{B}} |z_e| < \tau\} \cup \{\sup_{e \in -\mathcal{B}} |z_e| \geq \tau\} \tag{3.1}$$

$$\leq 2|\mathcal{B}| e^{-(\delta' - \tau)^2/4} + (m - |\mathcal{B}|) e^{-\tau^2/4}$$

The inequality works because within $\mathcal{B}$, z_e is normal with variance 2 and mean of magnitude at least δ'. Also, within $-\mathcal{B}$, z_e is normal with zero mean and variance 2. The RHS of (3.1) is equal to $me^{-\delta'^2/8}$ if we select $\tau = \delta'/2$. Hence, if $\delta' = \Omega(\sqrt{(\log(m))})$ then we obtain dual consistency.

To prove (2) we must employ the generic chaining to control the supremum of a Gaussian process. First let us recall the following fact about zero mean Gaussian processes, X_T, indexed by the set T with $\sigma'^2 \geq \sup_{t\in T} \mathbb{E}X_t^2$.

$$\mathbb{P}\{|\sup_{t\in T} X_t - \mathbb{E}\sup_{t\in T} X_t| \geq u\} \leq e^{-u^2/(2\sigma'^2)}$$

Hence, if $\tau = o(\mathbb{E}\sup_{e\in -\mathcal{B}} z_e)$ then certainly $\hat{\mathcal{B}}$ is not sparsistent. But certainly $\tau < \delta'$ is a necessary condition for sparsistency because there exists an $e \in \mathcal{B}$ such that $\mathbb{P}\{|z_e| \leq \delta'\} \geq C$ for some constant C. Thus, $\delta' = \Omega(\mathbb{E}\sup_{e\in -\mathcal{B}} z_e)$ is necessary for consistency.

Now we use the generic chaining methods developed by [75] to bound $\mathbb{E}\sup_{e\in -\mathcal{B}} z_e$ from below. Let T be an index set with $|T| = n$ and $N_n = 2^{2^n}$ and define an admissible sequence to be a sequence of increasing partitions $\mathcal{H}_n$ such that $|\mathcal{H}_n| \leq N_n$. Let $\text{diam}(H_n(t))$ be the diameter of the cell containing t in the partition. Now define the following functional of metric space (T, d),

$$\gamma_\alpha(T, d) = \inf \sup_{t\in T} \sum_{n\geq 0} 2^{n/\alpha} \text{diam}(H_n(t))$$

where the infimum is over all admissible sequences. Then we have the following majorizing measure theorem,

Theorem 11 ([75])**.** *For Gaussian process X_T, let $d(s,t) = \sqrt{\mathbb{E}(X_s - X_t)^2}$.*

$$\frac{1}{L}\gamma_2(T, d) \leq \mathbb{E}\sup_{t\in T} X_t \leq L\gamma_2(T, d)$$

for some universal constant L.

Now z_e restricted to $-\mathcal{B}$ is a Gaussian process (with mean 0) resulting in the distance,

$$d^2(a, b) = \begin{cases} 0, & a = b \\ 2, & a^+ = b^+ \text{ xor } a^- = b^- \\ 6, & a^+ = b^- \text{ xor } a^- = b^+ \\ 4, & \text{otherwise} \end{cases}$$

Until the allowable partition size $N_n \geq |-\mathcal{B}|$ the largest diameter of a cell is at least $\sqrt{2}$ (because before that point we cannot have a partition of singletons). Hence,

$$\gamma_2(E^\pm, d) \geq \sqrt{2} \sum_{n=0}^{\lfloor \log\log |-\mathcal{B}| \rfloor} 2^{n/2} \geq \sqrt{2\log|-\mathcal{B}|}$$

So,

$$\mathbb{E}\sup_{e\in E} |X_e| \geq \frac{\sqrt{2\log(m - |\mathcal{B}|)}}{L}$$

Thus, $\delta' = \delta/\sigma = \Omega(\log(m - |\mathcal{B}|))$ is necessary for dual consistency. □

We see immediately that the signal to noise ratio must be increasing like the log of the number of edges for edge thresholding to achieve changepoint sparsistency.

3.2 Edge Lasso

In this section we will analyze the edge lasso estimator, which arises as the solution to a generalized fused lasso problem as defined in [78] with the graph constraints specified by the matrix ∇. In particular, the edge lasso is the minimizer of the convex problem

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{x}}\|_2^2 + \lambda \|\nabla \hat{\mathbf{x}}\|_1, \tag{3.2}$$

were $\lambda > 0$ is a tuning parameter. Thus, the edge lasso is the penalized least squares estimator of $\mathbf{x}$ with penalty term given by the ℓ_1-norm of the differences of measurements across edges in G. Using KKT conditions and a primal-dual witness method we are able to extract the following theorem regarding noiseless recovery of $\mathbf{x}$.

As shown in [78], the dual problem to (3.2) is given by

$$\min_{\mathbf{z} \in \mathbb{R}^m} \frac{1}{2}\|\mathbf{y} - \lambda \nabla^\top \mathbf{z}\|_2^2 \text{ such that } \|\mathbf{z}\|_\infty \leq 1, \tag{3.3}$$

and any solution $\hat{\mathbf{z}}$ to the dual problem results in the primal solution (see [78] for details)

$$\hat{\mathbf{x}} = \mathbf{y} - \lambda \nabla^\top \hat{\mathbf{z}} = \mathcal{P}_{\text{null}(\nabla_{-\hat{\mathcal{B}}})}(\mathbf{y} - \lambda \nabla_{\hat{\mathcal{B}}}^\top \hat{\mathbf{z}}_{\hat{\mathcal{B}}}). \tag{3.4}$$

where $\hat{\mathcal{B}} = \{e \in E : |\mathbf{z}_e| = 1\}$. In this way, we can assess if the solution to the dual program is sparsistent. Unlike the primal solution, the solution to the dual problem is not always unique. In fact it may be that there are two solutions with different dual sparsity patterns $\hat{\mathcal{B}}$, but have the same $\hat{\mathbf{x}}$. [78]

3.2.1 Noiseless Changepoint Sparsistency

We now consider the performance of the edge lasso in the noiseless case, i.e. when the vertices are observed without noise. The reasons for investigating this seemingly uninteresting case are two-fold. First, somewhat surprisingly, there are many graphs for which the primal problem will not recover the correct sparsity pattern $\mathcal{A}$, for any value $\lambda > 0$. (See Figure 3.2) For these graphs, consistent noisy recovery with the edge lasso is therefore hopeless. Secondly, and more importantly, as we will show later, in order for noisy recovery to be possible, it is sufficient that strict dual feasibility holds, i.e. $\|\mathbf{z}_{-\mathcal{B}}\|_\infty < a$ for some $a < 1$ for some dual solution $\mathbf{z}$ to the noiseless edge lasso problem.

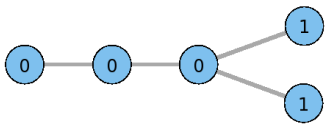


Figure 3.2: An example where for all $\lambda > 0$ the edge lasso does not recover the true $\mathcal{A}$. The 0 to the right is separated into its own element of the partition.

When testing for sparsistency we use the following **primal dual witness** (PDW) construction, pioneered by [82], which results in a pair $(\hat{\mathbf{x}}, \hat{\mathbf{z}})$ of primal and dual solutions. The PDW

construction will be used as sufficient conditions for sparsistency. Note that this is not a practical method for solving the dual problem and is only used as a proof technique. We begin by setting $\mathbf{z}_{\mathcal{B}} = \operatorname{sign}(\nabla \mathbf{x})$, which is equivalent to assuming the knowledge of both $\mathcal{B}$ and the sign differences. Using this knowledge, compute $\hat{\mathbf{x}} = \mathcal{P}_{\text{null}(\nabla_{-\mathcal{B}})}(\mathbf{y} - \lambda \nabla_{\mathcal{B}}^{\top} \hat{\mathbf{z}}_{\mathcal{B}})$. The PDW steps are as follows.

1. Verify the complementary slackness condition $\operatorname{sign}(\nabla_{\mathcal{B}} \hat{\mathbf{x}}) = \hat{\mathbf{z}}_{\mathcal{B}}$.
2. Construct $\tilde{\mathbf{z}}$ by solving the linear program

$$\min_{\tilde{\mathbf{z}} \in Z} \|\tilde{\mathbf{z}}\|_{\infty} \tag{3.5}$$

   where Z is the set of all dual parameters that satisfy the zero-subgradient condition in the noiseless setting, i.e.

$$Z = \{\mathbf{z} \in \mathbb{R}^{-\mathcal{B}} : \nabla_{-\mathcal{B}} \nabla_{-\mathcal{B}}^{\top} \mathbf{z} = -\nabla_{-\mathcal{B}} \nabla_{\mathcal{B}}^{\top} \mathbf{z}_{\mathcal{B}}\}$$

3. Construct the noisy dual by

$$\mathbf{z}_{-\mathcal{B}} = \frac{1}{\lambda} \nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^{\dagger} \boldsymbol{\epsilon} + \tilde{\mathbf{z}}$$

   where $\Delta_{-\mathcal{B}}$ is the Laplacian of the graph $G_{-\mathcal{B}}$.
4. Check the strict dual feasibility condition $\|\mathbf{z}_{-\mathcal{B}}\|_{\infty} < 1$.

Theorem 12. *If the PDW method passes for all large enough p then the solution to the dual program* (3.3) *is sparsistent.*

Before we can prove Theorem 12 we need the following lemma. This lemma will also be used when proving Proposition 22.

Lemma 13. *Suppose we are given the boundary set $\mathcal{B}$ with sign vector $\mathbf{z}_{\mathcal{B}} \in \{-1, 0, 1\}^E$ supported only over $\mathcal{B}$. Let $\hat{\mathbf{z}}_{-\mathcal{B}}^{\dagger} = \nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^{\dagger} (\frac{\mathbf{y}}{\lambda} - \nabla_{\mathcal{B}}^{\top} \mathbf{z}_{\mathcal{B}})$. Set $\mathbf{z}^{\dagger} = \mathbf{z}_{\mathcal{B}} + \hat{\mathbf{z}}_{-\mathcal{B}}^{\dagger}$ and obtain the corresponding primal solution*

$$\hat{\mathbf{x}} = \mathbf{y} - \lambda \nabla^{\top} \hat{\mathbf{z}}^{\dagger}.$$

There exists a solution to the dual problem with $\mathcal{B}$ and $\mathbf{z}_{\mathcal{B}}$ as given if and only if $\exists f \in$ null$(\nabla_{-\mathcal{B}}^{\top})$ such that

1. *Dual feasibility:* $\|\mathbf{z}_{-\mathcal{B}}^{\dagger} + f\|_{\infty} \leq 1$
2. *Complementary slackness:* $\operatorname{sign}(\nabla_{\mathcal{B}} \hat{\mathbf{x}}) \subseteq \mathbf{z}_{\mathcal{B}}$

Where $\subseteq$ in the complementary slackness is taken to mean that it is equal over the support of sign$(\nabla_{\mathcal{B}} \hat{\mathbf{x}})$.

Proof of Lemma 13. We will enumerate the KKT conditions and find that $\hat{\mathbf{z}}^{\dagger}$ arises due to the zero-subgradient conditions leaving only the dual feasibility and complementary slackness to be satisfied. We introduce Lagrangian parameters $\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_-$ and use the following Lagrangian,

$$\frac{1}{2} \|\mathbf{y} - \lambda \nabla^{\top} \mathbf{z}\|_2^2 + \boldsymbol{\gamma}_+^{\top} (\mathbf{z} - 1) + \boldsymbol{\gamma}_-^{\top} (-\mathbf{z} - 1)$$

This was obtained by turning $\|\mathbf{z}\|_{\infty} < 1$ into linear constraints.

The following are the complete KKT conditions:

1. Zero subgradient: $\nabla(\lambda\nabla^\top \mathbf{z} - \mathbf{y}) + \boldsymbol{\gamma}_+ - \boldsymbol{\gamma}_- = 0$
2. Parameter domain: $\boldsymbol{\gamma}_+, \boldsymbol{\gamma}_- \geq 0$
3. Dual feasibility: $\forall e, z_e - 1 \leq 0, -z_e - 1 \leq 0$
4. Complementary slackness: $\gamma_{+,e} = 0$ if $z_e \neq 1$ and $\gamma_{-,e} = 0$ if $z_e \neq -1$

Now define $\boldsymbol{\gamma} = \boldsymbol{\gamma}_+ - \boldsymbol{\gamma}_-$ the KKT conditions may be reduced to,

1. Zero subgradient: $\nabla(\lambda\nabla^\top \mathbf{z} - \mathbf{y}) + \boldsymbol{\gamma} = 0$
2. Dual feasibility: $\|\mathbf{z}\|_\infty \leq 1$
3. Complementary slackness: $\gamma_e \geq 0$ if $z_e = 1$, $\gamma_e \leq 0$ if $z_e = -1$, and $\boldsymbol{\gamma} = 0$ otherwise.

Notice that the existence of such a $\boldsymbol{\gamma}$ is necessary and sufficient for dual optimality due to convexity. Consider the zero subgradient condition only over $-\hat{\mathcal{B}}$,

$$\lambda\nabla_{-\hat{\mathcal{B}}}\nabla_{-\hat{\mathcal{B}}}^\top \mathbf{z}_{-\hat{\mathcal{B}}} + \nabla_{-\hat{\mathcal{B}}}(\lambda\nabla_{\hat{\mathcal{B}}}^\top \mathbf{z}_{\hat{\mathcal{B}}} - \mathbf{y}) = 0$$

because over $-\mathcal{B}$, $\gamma_e = 0$. This is equivalent to

$$\begin{gathered}\exists f \in \text{null}(\nabla_{-\hat{\mathcal{B}}}^\top) \text{ such that} \\ \mathbf{z}_{-\hat{\mathcal{B}}} = (\nabla_{-\hat{\mathcal{B}}}\nabla_{-\hat{\mathcal{B}}}^\top)^\dagger \nabla_{-\hat{\mathcal{B}}}(\tfrac{\mathbf{y}}{\lambda} - \nabla_{\hat{\mathcal{B}}}^\top \mathbf{z}_{\hat{\mathcal{B}}}) + f\end{gathered}$$

Now we will show that $(\nabla_{-\hat{\mathcal{B}}}\nabla_{-\hat{\mathcal{B}}}^\top)^\dagger \nabla_{-\hat{\mathcal{B}}} = \nabla_{-\hat{\mathcal{B}}}\Delta_{-\hat{\mathcal{B}}}^\dagger$. Let $\nabla_{-\hat{\mathcal{B}}} = \mathbf{U}\Lambda\mathbf{V}^\top$ be the singular value decomposition of $\nabla_{-\hat{\mathcal{B}}}$ then

$$\begin{aligned}(\nabla_{-\hat{\mathcal{B}}}\nabla_{-\hat{\mathcal{B}}}^\top)^\dagger \nabla_{-\hat{\mathcal{B}}} &= \mathbf{U}(\Lambda^\dagger)^2\mathbf{U}^\top\mathbf{U}\Lambda\mathbf{V}^\top \\ &= \mathbf{U}\Lambda^\dagger\mathbf{V}^\top = \mathbf{U}\Lambda\mathbf{V}^\top\mathbf{V}\Lambda^\dagger\mathbf{U}^\top \\ &= \nabla_{-\hat{\mathcal{B}}}\Delta_{-\hat{\mathcal{B}}}^\dagger\end{aligned}$$

Furthermore defining $\boldsymbol{\gamma}_{\hat{\mathcal{B}}} = \nabla_{\hat{\mathcal{B}}}^\top \hat{\mathbf{x}}$ is necessary and sufficient for the remainder of the zero subgradient condition. Now the complementary slackness holds if and only if $\text{sign}(\nabla_{\hat{\mathcal{B}}}\hat{\mathbf{x}}) \subseteq \mathbf{z}_{\hat{\mathcal{B}}}$. □

Proof of Theorem 12. We will show that the conditions of Lemma 13 are satisfied if the PDW passes. By construction, if 1 passes then complementary slackness in Lemma 13 will be satisfied. By step 2 of the PDW construction, $\nabla_{-\mathcal{B}}\nabla_{-\mathcal{B}}^\top \tilde{\mathbf{z}} = -\nabla_{-\mathcal{B}}\nabla_{\mathcal{B}}^\top \mathbf{z}_{\mathcal{B}}$ implies that $\mathbf{z} = -\nabla_{-\mathcal{B}}\Delta_{-\mathcal{B}}^\dagger\nabla_{\mathcal{B}}^\top \mathbf{z}_{\mathcal{B}} + f$, for some $f \in \text{null}(\nabla_{-\mathcal{B}}^\top)$, again by the SVD decomposition of $\Delta_{-\mathcal{B}}$. But we know that $\Delta_{-\mathcal{B}}^\dagger \mathbf{x} = 0$ because the Moore-Penrose pseudoinverse has zero action on any vectors that are constant over connected components of $G_{-\mathcal{B}}$. Therefore, $\frac{1}{\lambda}\nabla_{-\mathcal{B}}^\top\Delta_{-\mathcal{B}}^\dagger\boldsymbol{\epsilon} = \frac{1}{\lambda}\nabla_{-\mathcal{B}}^\top\Delta_{-\mathcal{B}}^\dagger\mathbf{y}$. Next, by step 3 we know that

$$\mathbf{z}_{-\mathcal{B}} = \nabla_{-\mathcal{B}}\Delta_{-\mathcal{B}}^\dagger(\frac{\mathbf{y}}{\lambda} - \nabla_{\mathcal{B}}^\top \mathbf{z}_{\mathcal{B}}) + f = \mathbf{z}_{-\mathcal{B}}^\dagger + f$$

If step 4 passes then dual feasibility holds. □

Proposition 14. *For dual solution* $\mathbf{z}$ *the zero-subgradient condition is satisfied if and only if for all* $\hat{A} \in \hat{\mathcal{A}}$*, and for all* $C \subseteq \hat{A}$*,*

$$\chi(C)^\top \mathbf{z}_{-\hat{\mathcal{B}}} = \frac{|C|}{|\hat{A}|}\chi(\hat{A})^\top \mathbf{z}_{\mathcal{B}} - \chi(C)^\top \mathbf{z}_{\mathcal{B}}$$

Proof. The zero-subgradient condition for the noiseless setting can be rewritten as

$$\nabla^\top_{-\hat{\mathcal{B}},v}\mathbf{z}_{-\hat{\mathcal{B}}} = \mathcal{P}_{\text{row}(\nabla^\top_{-\hat{\mathcal{B}}})}(-\nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}}) \tag{3.6}$$

$$= \tfrac{\chi(\hat{A})^\top \mathbf{z}_{\hat{\mathcal{B}}}}{|\hat{A}|} - \nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}} \tag{3.7}$$

Because $\mathbf{z}_{-\hat{\mathcal{B}}}$ is supported only over $-\hat{\mathcal{B}}$ we can rewrite $\nabla^\top_{-\hat{\mathcal{B}},v}\mathbf{z}_{-\hat{\mathcal{B}}} = \nabla^\top_v \mathbf{z}_{-\hat{\mathcal{B}}}$. Similarly, $\nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}} = \nabla^\top_v \mathbf{z}_{\hat{\mathcal{B}}}$. Recall that $\sum_{v\in C}\nabla^\top_v = \chi(C)^\top$ and the result follows by summing each side of eq. (3.8). The other direction follows immediately by setting $C = \{v\}$ and we have that $\chi(\{v\}) = \nabla_v$. □

Below, we will outline sufficient conditions for sparsistency that are based on the topology of the graph G. To this end, recall that, for a given estimated partition $\hat{\mathcal{A}}$, we have an explicit form for the approximation error incurred on $v \in \hat{A}(v) \in \hat{\mathcal{A}}$ using the characteristic vector of the cut $\hat{A}$, namely

$$\hat{\mathbf{x}}_v - \mathbf{x}_v = (\mathcal{P}_{\text{null}(\nabla_{-\hat{\mathcal{B}}})}(\mathbf{x} - \lambda\nabla^\top_{\hat{\mathcal{B}}}\hat{\mathbf{z}}_{\hat{\mathcal{B}}}))_v - \mathbf{x}_v$$

$$= -\lambda\frac{\chi(\hat{A}(v))^\top \hat{\mathbf{z}}_{\hat{\mathcal{B}}}}{|\hat{A}(v)|}$$

Notice that $|\chi(\hat{A}(v))^\top \hat{\mathbf{z}}_{\hat{\mathcal{B}}}| \leq |\partial\hat{A}(v)|$ because of the definition of the characteristic vector. We find that the success and failure of the noiseless edge lasso is dictated by the presence of a bottleneck cut of elements $A \in \mathcal{A}$ in a sense made precise in the following result.

Lemma 15. *Let* $\mathbf{z}$ *be the result of the PDW method and notice that in the noiseless setting* $\tilde{\mathbf{z}} = \mathbf{z}_{-B}$*. Then for some* $A \in \mathcal{A}$ *there exists a cut of* A *with shores* $C, \bar{C}$ *such that*

$$\|\tilde{\mathbf{z}}\|_\infty = \frac{1}{|\partial C \cap \partial\bar{C}|}\left|\frac{|C|}{|A|}\chi(\bar{C})^\top \mathbf{z}_{\mathcal{B}} - \frac{|\bar{C}|}{|A|}\chi(C)^\top \mathbf{z}_{\mathcal{B}}\right|$$

Proof. For the following theorem we will focus on a connected component A that contains an edge e such that $|\tilde{\mathbf{z}}_e| = \|\tilde{\mathbf{z}}\|_\infty$. Let $Q = \{e \in E(A) : |\tilde{\mathbf{z}}_e| = \|\tilde{\mathbf{z}}\|_\infty\}$ and denote $\zeta = \|\tilde{\mathbf{z}}\|_\infty$. Suppose that Q is not a cut of A (the removal of Q does not disconnect A). There exists a spanning tree of A not containing Q as we can take any spanning tree of A with Q removed. Take $e \in Q$ then form a cycle ϕ containing e by including the unique path in the spanning tree from e_h to e_t. Notice that e is the unique element of ϕ such that $|\hat{z}_e| = \zeta$. Construct a new edge vector $z' = \tilde{\mathbf{z}} + \eta\chi(\phi)$ for some small η such that $|z'_e|$ is smaller over the cycle ϕ. Notice that $|z'_e| < \zeta$ and with $|\eta| < \zeta - \max_{e'\neq e}|z_e|$, $|z'_{e'}| < \zeta$ for all $e' \in \phi$. Repeat this procedure for the other elements of Q replacing $\tilde{\mathbf{z}}$ with z'. We obtain a new edge vector that satisfies the zero subgradient condition because we only added elements of the null$(D^\top_{-\mathcal{B}})$. Moreover $\|z'\|_\infty < \|\tilde{\mathbf{z}}\|_\infty$, contradicting the fact that $\tilde{\mathbf{z}}$ is the solution to eqn. (3.5) in PDW step (2). Hence, Q is a cut of A and for all $e \in Q$, $|\tilde{\mathbf{z}}_e| = \|\tilde{\mathbf{z}}\|_\infty$, and let one shore of the cut be C. Notice that $|Q| = |\partial C \cup \partial\bar{C}|$ and $\|\tilde{\mathbf{z}}\|_\infty|\partial C \cup \partial\bar{C}| = |\chi(C)^\top\tilde{\mathbf{z}}| = |\chi(C)^\top\mathbf{z}_{-B}|$. Now, Proposition 16 below states that the zero subgradient condition is equivalent to,

$$\chi(C)^\top \mathbf{z}_{-\mathcal{B}} = \frac{|C|}{|A|}\chi(A)^\top \mathbf{z}_{\mathcal{B}} - \chi(C)^\top \mathbf{z}_{\mathcal{B}}$$

$$= \frac{|C|}{|A|}\chi(A)^\top \mathbf{z}_{\mathcal{B}} - \frac{|C| + |\bar{C}|}{|A|}\chi(C)^\top \mathbf{z}_{\mathcal{B}}$$
$$= \frac{|C|}{|A|}\chi(\bar{C})^\top \mathbf{z}_{\mathcal{B}} + \frac{|\bar{C}|}{|A|}\chi(C)^\top \mathbf{z}_{\mathcal{B}}$$

□

Proposition 16. *For dual solution* $\mathbf{z}$ *the zero-subgradient condition is satisfied if and only if for all* $\hat{A} \in \hat{\mathcal{A}}$*, and for all* $C \subseteq \hat{A}$*,*

$$\chi(C)^\top \mathbf{z}_{-\hat{\mathcal{B}}} = \frac{|C|}{|\hat{A}|}\chi(\hat{A})^\top \mathbf{z}_{\mathcal{B}} - \chi(C)^\top \mathbf{z}_{\mathcal{B}}$$

Proof. The zero-subgradient condition for the noiseless setting can be rewritten as

$$\nabla^\top_{-\hat{\mathcal{B}},v}\mathbf{z}_{-\hat{\mathcal{B}}} = \mathcal{P}_{\text{row}(\nabla^\top_{-\hat{\mathcal{B}}})}(-\nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}}) \tag{3.8}$$
$$= \frac{\chi(\hat{A})^\top \mathbf{z}_{\hat{\mathcal{B}}}}{|\hat{A}|} - \nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}} \tag{3.9}$$

Because $\mathbf{z}_{-\hat{\mathcal{B}}}$ is supported only over $-\hat{\mathcal{B}}$ we can rewrite $\nabla^\top_{-\hat{\mathcal{B}},v}\mathbf{z}_{-\hat{\mathcal{B}}} = \nabla^\top_v \mathbf{z}_{-\hat{\mathcal{B}}}$. Similarly, $\nabla^\top_{\hat{\mathcal{B}},v}\mathbf{z}_{\hat{\mathcal{B}}} = \nabla^\top_v \mathbf{z}_{\hat{\mathcal{B}}}$. Recall that $\sum_{v\in C}\nabla^\top_v = \chi(C)^\top$ and the result follows by summing each side of eq. (3.8). The other direction follows immediately by setting $C = \{v\}$ and we have that $\chi(\{v\}) = \nabla_v$. □

Using the previous Lemma we offer the following sufficient conditions for correct recovery (and strict dual feasibility).

Theorem 17. *Define the following notion of connectivity for each* $A \in \mathcal{A}$*,*

$$\nu(A) = \max_{C\subset A} \frac{|C|}{|\partial C \cap \partial \bar{C}|}\frac{|\partial \bar{C} \cap \partial A|}{|A|} \tag{3.10}$$

Then the noiseless problem recovers the correct $\mathcal{A}$ *if* $\nu(\mathcal{A}) = \max_{A\in\mathcal{A}}\nu(A) < 1/2$*.*

Proof. Consider the C, A pair in the proof of Lemma 15. We need to show that, $\frac{1}{|\partial C\cap\partial\bar{C}|}|\frac{|C|}{|A|}\chi(\bar{C})^\top\mathbf{z} - \frac{|\bar{C}|}{|A|}\chi(C)^\top\mathbf{z}| \le 2\rho(A)$. To this end, note that $\frac{1}{|\partial C\cap\partial\bar{C}|}|\frac{|C|}{|A|}\chi(\bar{C})^\top\mathbf{z} - \frac{|\bar{C}|}{|A|}\chi(C)^\top\mathbf{z}|$ is smaller than

$$\frac{|\chi(\bar{C})^\top\mathbf{z}|}{|\partial C\cap\partial\bar{C}|}\frac{|C|}{|A|} + \frac{|\chi(C)^\top\mathbf{z}|}{|\partial C\cap\partial\bar{C}|}\frac{|\bar{C}|}{|A|}$$

which in turn is smaller than

$$2\max_{C\subset A}\frac{|\chi(\bar{C})^\top\mathbf{z}|}{|\partial C\cap\partial\bar{C}|}\frac{|C|}{|A|} \le 2\max_{C\subset A}\frac{|\partial\bar{C}\cap\partial A|}{|\partial C\cap\partial\bar{C}|}\frac{|C|}{|A|} = 2\rho(A).$$

□

See Figure 3.3 for an illustration of condition (3.10) in the previous theorem. An interpretation of the $\nu(A)$ parameter is that there is no bottleneck within vertices of constant activation, A, relative to the flow coming in and out of A. Later we will describe a class of graphs which we call the nested complete graphs for which $\nu(A) = \frac{1}{|A|}$ for $A \in \mathcal{A}$. We now combine this result with subGaussian concentration to produce a changepoint sparsistency result in the noisy regime.

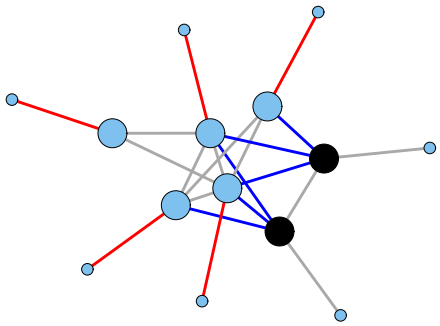


Figure 3.3: An example of the quantities in eq. (3.10) for a cut of set A depicted by the large vertices. The cut C are the black vertices, $\partial C \cap \partial \bar{C}$ are blue edges, and $\partial \bar{C} \cap \partial A$ are red edges. The RHS of eq. (3.10) for this cut is $5/21$.

3.2.2 Noisy Changepoint Sparsistency

We now analyze the performance of the noisy edge lasso estimator. We will rely on the PDW construction and on the results from the previous section to formulate conditions under which the edge lasso achieves sparsistency. All of the proofs in this section are in the supplementary material and are a combination of Gaussian concentration and noiseless recovery. We first provide conditions guaranteeing that, asymptotically, the first step of the PDW construction passes.

Lemma 18. *Let $\hat{\mathbf{x}}$ be the estimated signal resulting from the PDW method and δ is the minimal gap of the symbol. If $\forall A \in \mathcal{A}$,*

$$\frac{\delta}{\sigma} = \omega\left(\frac{1}{\sqrt{|A|}}\right) \text{ and } \lambda = o\left(\delta \frac{|A|}{|\partial A|}\right) \tag{3.11}$$

then step (1) of the PDW method passes with probability tending to 1.

Proof. In the PDW method we use $\hat{\mathcal{B}} = \mathcal{B}$. Recall that the estimated signal is given by,

$$\hat{\mathbf{x}}_v = (\mathcal{P}_{\text{null}(\nabla_{-\mathcal{B}})}(y - \lambda \nabla_{\mathcal{B}}^\top \mathbf{z}_{\mathcal{B}}))_v$$

$$= \mathbf{x}_v + \frac{\sum_{w \in A(v)} \epsilon_w}{|A(v)|} - \lambda \frac{\chi(A(v))^\top \mathbf{z}_{\mathcal{B}}}{|A(v)|}$$

Now by Gaussian concentration, we know that for $\gamma > 0$ with probability at least $1 - 2\gamma$,

$$\frac{|\sum_{w \in A(v)} \epsilon_w|}{|A(v)|} \leq \sigma \sqrt{\frac{2}{|A(v)|} \log(\frac{1}{\gamma})}$$

We intend to show that for v, w such that $A(v) \neq A(w)$,

$$\frac{|\hat{\mathbf{x}}_v - \hat{\mathbf{x}}_w|}{\delta} \geq 1 - o(1)$$

Differencing the equation for $\hat{\mathbf{x}}$ we have,

$$\frac{|\hat{\mathbf{x}}_v - \hat{\mathbf{x}}_w|}{\delta} \geq 1 - \frac{\sigma}{\delta}\sqrt{\frac{2}{|A(v)|}\log(\frac{1}{\gamma})}$$

$$-\frac{\sigma}{\delta}\sqrt{\frac{2}{|A(w)|}\log(\frac{1}{\gamma})} - \frac{\lambda}{\delta}(\frac{|\partial A(v)|}{|A(v)|} + \frac{|\partial A(w)|}{|A(w)|})$$

The conditions of (3.11) imply that

$$\frac{\sigma}{\delta}\sqrt{\frac{2}{|A(v)|}\log(\frac{1}{\gamma})} + \frac{\sigma}{\delta}\sqrt{\frac{2}{|A(w)|}\log(\frac{1}{\gamma})} = o(1)$$

$$\frac{\lambda}{\delta}(\frac{|\partial A(v)|}{|A(v)|} + \frac{|\partial A(w)|}{|A(w)|}) = o(1)$$

□

We continue our study of the noisy reconstruction with the edge lasso by outlining sufficient conditions for sparsistency based on the $2, \infty$ norm of the operator $\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}$. The intuition is that if we have some dual slack in the sense that $\|\tilde{\mathbf{z}}\|_\infty$ is bounded away from 1 and if we bound the maximum of $|(\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\boldsymbol{\epsilon})_e|$ then the PDW method will pass. We show that we can accurately describe $\|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\|_{2,\infty}$ with the spectrum of the Laplacian. Specifically, if the eigenvectors corresponding to low eigenvalues do not differ significantly across an edge then we have a small $\|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\|_{2,\infty}$.

Lemma 19. *Let $\tilde{\mathbf{z}}$ be the result of step (2) of the PDW method. If $\|\tilde{z}_{-\mathcal{B}}\|_\infty < c$ for all large p for some $0 < c < 1$ and*

$$\sigma = o\left(\frac{\lambda}{\|\!|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}|\!\|_{2,\infty}\sqrt{\log(|-\mathcal{B}|)}}\right)$$

Then step (4) in the PDW method passes for large enough p.

Proof. First we know that for each $e \in -\mathcal{B}$,

$$(\nabla_{\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\boldsymbol{\epsilon})_e \sim N(0, \sigma^2\|\nabla_e\Delta^{\dagger}_{-\mathcal{B}}\|_2^2)$$

Notice that $\|\!|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}|\!\|^2_{2,\infty} = \max_e \|\nabla_e\Delta^{\dagger}_{-\mathcal{B}}\|_2^2$ Hence, $\|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\boldsymbol{\epsilon}\|_\infty$ is the maximum of $|-\mathcal{B}|$ Gaussian random variables with maximum variance $\sigma^2\|\!|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}|\!\|^2_{2,\infty}$. By Gaussian concentration and the union bound we know that,

$$\|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\boldsymbol{\epsilon}\|_\infty \leq \sigma\|\!|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}|\!\|_{2,\infty}\sqrt{2\log(\frac{|-\mathcal{B}|}{\gamma})}$$

with probability at least $1-\gamma$. So,

$$\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_\infty + \frac{1}{\lambda}\|\nabla_{-\mathcal{B}}\Delta^{\dagger}_{-\mathcal{B}}\boldsymbol{\epsilon}\|_\infty$$

$$\leq \|\mathbf{z}\|_\infty + \frac{\sigma \|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty}}{\lambda} \sqrt{2 \log(\frac{|-\mathcal{B}|}{\gamma})}$$

So, $\|\mathbf{z}\|_\infty < 1$ with high probability for large p if

$$\frac{\sigma \|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty}}{\lambda} \sqrt{\log(|-\mathcal{B}|)} = o(1)$$

which occurs if

$$\sigma = o(\frac{\lambda}{\|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty} \sqrt{\log(|-\mathcal{B}|)}})$$

□

By putting together the results described so far we arrive at the following conditions for sparsistency.

Theorem 20. *Let $\mathcal{B} = supp(\nabla \mathbf{x})$ the changepoints of $\mathbf{x}$. Suppose that the following conditions hold for all $A \in \mathcal{A}$, and let δ be the gap between signal across clusters of activation in $\mathcal{A}$.*

$$\nu(A) = o(1)$$

$$\frac{\delta}{\sigma} = \omega \left(\frac{|\partial A|}{|A|} \|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty} \sqrt{\log(|-\mathcal{B}|)} \right)$$

$$\frac{\delta}{\sigma} = \omega \left(\frac{1}{\sqrt{|A|}} \right)$$

then the edge lasso is changepoint sparsistent.

Proof. Recall from Theorem 17 that $\|\tilde{\mathbf{z}}\|_\infty \leq 2 \max_{A \in \mathcal{A}} \rho(A)$ and so $\|\tilde{\mathbf{z}}\|_\infty = o(1)$. The second condition implies the conditions of Lemma 19 for some $\lambda = o(\delta \min_{A \in \mathcal{A}} \frac{|A|}{|\partial A|})$. Thus, step (4) of the PDW method passes for large enough p. The third condition completes the conditions of Lemma 18 and step (1) passes for large enough p. □

Thus far our conditions rely on the size of $\|\nabla_{-\mathcal{B}} \Delta_{-B}^\dagger\|_{2,\infty}$ with no obvious validation techniques. We are able to relate this norm to the smoothness of eigenvectors of Laplacians weighted by the reciprocals of eigenvalues.

Proposition 21. *Let the spectral decomposition of the Laplacian for $A \in \mathcal{A}$ be $\Delta_A = \mathbf{U} \Lambda \mathbf{U}^\top$ then $\|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty}$ is equal to*

$$\max_{A \in \mathcal{A}} \max_{e \in A} \sqrt{\sum_{v \in V} (\mathbf{U}_{v,e^+} - \mathbf{U}_{v,e^-})^2 (\lambda_v^2)^\dagger}$$

So, if each eigenvector $\mathbf{U}_v$ is η_v-Lipschitz with respect to the shortest path distance then $(\mathbf{U}_{v,e^+} - \mathbf{U}_{v,e^-})^2 \leq \eta_v^2$

$$\textit{and} \quad \|\nabla_{-\mathcal{B}} \Delta_{-\mathcal{B}}^\dagger\|_{2,\infty} \leq \max_{A \in \mathcal{A}} \sqrt{\sum_{v \in A} \eta_v^2 (\lambda_v^2)^\dagger}$$

Proof. Let $\Delta_{-\mathcal{B}} = U\Lambda U^\top$

$$\begin{aligned}
\|\nabla_{-\mathcal{B}}\Delta^\dagger_{-\mathcal{B}}\|_{2,\infty} &= \sup_{\|\boldsymbol{\alpha}\|_2\leq 1} \|\nabla_{-\mathcal{B}}\mathbf{U}\Lambda^\dagger\mathbf{U}^\top\boldsymbol{\alpha}\|_\infty \\
&= \sup_{\|\boldsymbol{\alpha}\|_2\leq 1} \|\nabla_{-\mathcal{B}}\mathbf{U}\Lambda^\dagger\boldsymbol{\alpha}\|_\infty = \sup_{\|\boldsymbol{\alpha}\|_2\leq 1} \max_{e\in-\mathcal{B}} \|\nabla_e\mathbf{U}\Lambda^\dagger\boldsymbol{\alpha}\|_\infty \\
&= \sup_{\|\boldsymbol{\alpha}\|_2\leq 1} \max_{e\in-\mathcal{B}} |(U_{e^+} - U_{e^-})\Lambda^\dagger\boldsymbol{\alpha}| \\
= \max_{e\in-\mathcal{B}} \sup_{\|\boldsymbol{\alpha}\|_2\leq 1} |(U_{e^+} - U_{e^-})\Lambda^\dagger\boldsymbol{\alpha}| &= \max_{e\in-\mathcal{B}} \|(U_{e^+} - U_{e^-})\Lambda^\dagger\|_2 \\
&= \max_{e\in-\mathcal{B}} \sqrt{\sum_{v\in V}(U_{v,e^+} - U_{v,e^-})^2(\Lambda_v^2)^\dagger}
\end{aligned}$$

The above supremum is achieved for $\boldsymbol{\alpha} = \Lambda^\dagger(U_{e^+} - U_{e^-})^\top/\|(U_{e^+} - U_{e^-})\Lambda^\dagger\|_2$. Because $G_{-\mathcal{B}}$ is disconnected Δ_{-B} is block diagonal. Hence, $\mathbf{U}$ is block diagonal with blocks being the eigenvectors of each component of $\mathcal{A}$. Moreover, $e \in -\mathcal{B}$ is completely internal to some $A \in \mathcal{A}$, so we have that

$$\|\nabla_{-\mathcal{B}}\Delta^\dagger_{-\mathcal{B}}\|_{2,\infty} = \max_{A\in\mathcal{A}} \max_{e\in A} \sqrt{\sum_{v\in V}(U_{v,e^+} - U_{v,e^-})^2(\lambda_v^2)^\dagger}$$

Where the eigenvectors $\mathbf{U}$ are understood to be specific to the $A \in \mathcal{A}$. The only thing remaining to show is that η-Lipschitz in the shortest path distance implies that $(U_{v,e^+} - U_{v,e^-})^2 \leq \eta^2$. But we have that η-Lipschitz in the shortest path distance occurs if and only if $\|\nabla U_v\|_\infty \leq \eta$. □

Proposition 21 provides a more tractable condition that implies that $\|\nabla_{-\mathcal{B}}\Delta^\dagger_{-\mathcal{B}}\|_{2,\infty}$ is small. Our findings suggest that the changepoint sparsistency of the edge lasso is highly dependent on the topology of G and its partition $\mathcal{A}$. In general, it is necessary that there exists no bottleneck cuts (cuts that force $\nu(\mathcal{A})$ to be large). We conclude this section with a final observation and a result that we will use in the next section to analyze the 1D and 2D edge lasso. There are many graphs for which the in degree of a vertex matches the out degree (namely the intersection with $\partial A(V)$), such as the 1D and 2D grids. We show that when the approximation error is small relative to the noise there is no hope of achieving sparsistency.

Proposition 22. *Suppose that $|\mathcal{A}| \geq 2$ and that for some $A \in \mathcal{A}$ there exists $v \in A$ such that $|\nabla^\top_{\mathcal{B},v}\mathbf{z}_\mathcal{B}| = |\{e \notin \mathcal{B} : v \in e\}|$. Let the maximum gap of $\mathbf{x}$ between elements of $\mathcal{A}$ be denoted $\delta_{\max}$. If $\frac{\delta_{\max}}{\sigma} = o(1)$ then edge lasso is not sparsistent.*

Proof. Suppose that the solution to the primal program does recover $\mathcal{A}$ and $\mathbf{z}_\mathcal{B}$ correctly. Then there is a solution to the dual program that recovers $\hat{\mathcal{B}}$ such that $\mathcal{B} \subseteq \hat{\mathcal{B}}$. Let us first find necessary conditions on λ for primal sparsistency. Recall that $\hat{\beta}_v$ is Gaussian with mean $(\mathcal{P}_{\text{null}(\nabla_{-\hat{\mathcal{B}}})}(\beta - \lambda\nabla^\top_{\hat{\mathcal{B}}}\mathbf{z}_{\hat{\mathcal{B}}}))_v$. So this mean is $\beta_v - \lambda\frac{\chi(A(v))^\top\mathbf{z}_{\hat{\mathcal{B}}}}{|A(v)|}$. So, if $\lambda\frac{\chi(A(v))^\top\mathbf{z}_{\hat{\mathcal{B}}}}{|A(v)|} > \delta_{\max}$ then we will not achieve sign consistency with probability at least $1/2$. Thus, $\lambda = o(\frac{\chi(A(v))^\top s}{|A(v)|})$ for all $A \in \mathcal{A}$ is necessary.

Lemma 13 implies

$$\nabla^\top_{-\hat{\mathcal{B}}}\mathbf{z}_{-\hat{\mathcal{B}}} = \nabla^\top_{-\hat{\mathcal{B}}}(\nabla_{-\hat{\mathcal{B}}}\nabla^\top_{-\hat{\mathcal{B}}})^\dagger\nabla_{-\hat{\mathcal{B}}}(\frac{y}{\lambda} - \nabla^\top_{\hat{\mathcal{B}}}\mathbf{z}_{\hat{\mathcal{B}}})$$

$$= \mathcal{P}_{\text{row}(\nabla_{-\hat{\mathcal{B}}})}(\frac{y}{\lambda} - \nabla_{\hat{\mathcal{B}}}^{\top}\mathbf{z}_{\hat{\mathcal{B}}})$$

Using the fact that $\mathcal{P}_{\text{row}(\nabla_{-\hat{\mathcal{B}}})}$ subtracts the average within $A \in \mathcal{A}$, we have,

$$\nabla_{-\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{-\hat{\mathcal{B}}} = -\nabla_{\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{\hat{\mathcal{B}}} + \frac{\chi(A(v))^{\top}\mathbf{z}_{\hat{\mathcal{B}}}}{|A(v)|} +$$

$$\frac{\epsilon_v}{\lambda} - \frac{1}{\lambda|A(v)|}\sum_{w\in A(v)}\epsilon_w$$

Notice that we can decompose the noise terms as, (using the shorthand that $A = A(v)$)

$$\frac{1}{\lambda}(\frac{|A|-1}{|A|}\epsilon_v + \frac{1}{|A|}\sum_{w\neq v}\epsilon_w)$$

which is less than $-\frac{|\partial A|}{|A|}$ with probability bounded from below for all large p if

$$\frac{\sigma|A|}{\lambda|\partial A|} = \omega(1)$$

which happens if

$$\frac{\sigma}{\delta_{\max}} = \omega(1)$$

We know that if the noise term dominates the potential bias $\frac{|\partial A|}{|A|}$ then $|\nabla_{-\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{-\hat{\mathcal{B}}}| > |\nabla_{\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{\hat{\mathcal{B}}}|$. But by the condition there are not enough internal edges to dissipate $|\nabla_{\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{\hat{\mathcal{B}}}|$ without making $\|\nabla_{-\hat{\mathcal{B}},v}^{\top}\mathbf{z}_{-\hat{\mathcal{B}}}\|_{\infty} > 1$, and we arrive at our contradiction. □

3.3 Specific Graph Models

We apply our results to the edge lasso over the 1 and 2 dimensional grids, commonly referred to as the fused lasso. In these cases the SNR must not decrease to achieve changepoint sparsistency, which is in sharp contrast to the performance of the oracle (see Figure 3.1). We provide a topology called the nested complete graph that satisfies the sufficient conditions for changepoint sparsistency. These examples are meant to provide a blueprint for using the previous results to explore topologies for which the edge lasso is changepoint sparsistent.

3.3.1 1D and 2D fused Lasso

Due to the popularity of total variation penalization, it is imperative that we discuss the 1D and 2D fused lasso. In the 1D grid each vertex can be associated with a number in $\{1, ..., p\}$ and we connect the pairs with Euclidean distance less than or equal to 1. Similarly, in the 2D grid each vertex can be associated with a number in $\{1, ..., p_0\} \times \{1, ..., p_1\}$. In the 2D grid we will say that a vertex v is a **corner** if its degree within $A(v)$ (the partition element containing v) is 2. (See Figure 3.4)

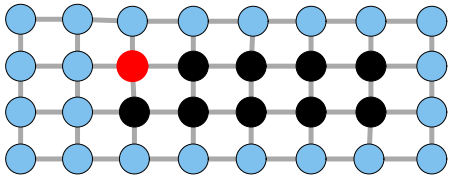


Figure 3.4: A 2D grid with $|\mathcal{A}| = 2$ depicted as union of black and red vertices. The red vertex is an example of a corner.

Corollary 23. *(a) Consider the 1D fused lasso with a non-trivial signal such that $|\mathcal{A}| = 2$. If the signal to noise ratio is decreasing ($\delta_{\max}/\sigma = o(1)$) then the 1D fused lasso is not changepoint sparsistent.*

(b) Consider the 2D fused lasso with a $A \in \mathcal{A}$ such that A contains a corner v and $|\mathcal{A}| = 2$. If the signal to noise ratio is decreasing ($\delta_{\max}/\sigma = o(1)$) then the 2D fused lasso is not changepoint sparsistent.

Proof. If the signal is non-trivial then there is a vertex $v \in A \in \mathcal{A}$ that is adjacent to ∂A. $|\nabla_v^\top \mathbf{z}_\mathcal{B}| = 1$ which is the degree of v within A, so the conditions of Proposition 22 hold. The 2D case follows by considering the corner as v with $|\nabla_v^\top \mathbf{z}_\mathcal{B}| = 2$. □

We see these typical mistakes in the 1D fused lasso in Figure 3.5. Here we observe a small incorrect element of $\hat{\mathcal{A}}$ at the boundary of a true element of $\mathcal{A}$. Here we see that it is due to small deviations from the true clusters that changepoint sparsistency fails in these cases.

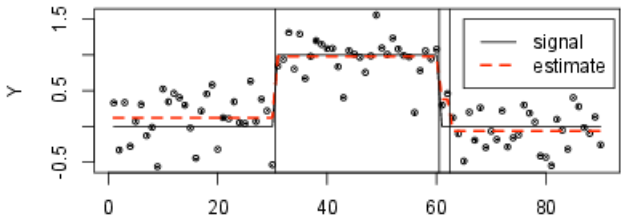


Figure 3.5: A typical mistake in the 1D fused lasso. The vertical lines indicate the beginning and end of estimated $\hat{\mathcal{A}}$.

3.3.2 Nested complete graph

We construct the nested complete graph from $k+1$ copies of the complete graph with k vertices by adjoining each complete graph to each other with one edge. We can form this such that each vertex has only one edge leaving its element in $\mathcal{A}$ which are the original complete graphs. (See Figure 3.6) We find that modulo factors that scale like the $\log p$, the changepoint sparsistency thresholds are the same as that of the oracle.

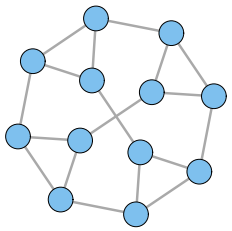


Figure 3.6: Nested complete graph with $k = 3$. $\mathcal{A}$ are the complete subgraphs of size 3.

Corollary 24. *Suppose we construct the nested complete graph with k vertices in A and $k+1$ elements in the partition ($|A| = k$ and $|\mathcal{A}| = k+1$). If the SNR satisfies,*

$$\frac{\delta}{\sigma} = \omega\left(\frac{1}{\sqrt{k}}\sqrt{\log(k(k+1))}\right)$$

Then the edge lasso is changepoint sparsistent.

Proof. Consider a cut C of the complete graph with p vertices. The cut size is $|\partial C \cap \partial \bar{C}| = |C|(p - |C|)$ while the cut boundary is $|\partial \bar{C} \cap \partial A| = p - |C|$. Hence,

$$\frac{|\partial \bar{C} \cap \partial A|}{|\partial C \cap \partial \bar{C}|}\frac{|C|}{|A|} = \frac{(p-|C|)}{|C|(p-|C|)}\frac{|C|}{p} = \frac{1}{p}$$

Thus, $\nu(\mathcal{A}) = o(1)$.

We know that the spectrum of the Laplacian of the p-complete graph has one eigenvalue of 0 and the rest are p. Because the eigenvectors are normalized the Lipschitz constants $\eta_v \leq 1$ as in Proposition 21. Hence, $\sqrt{\sum_{v\in V}\eta_v^2(\xi_v^2)^\dagger} \leq \sqrt{\sum_{v\in A}\frac{1}{p^2}} = \frac{1}{\sqrt{p}}$ Moreover $|\partial A| = |A| = p$ and we have that

$$\max_{A\in\mathcal{A}}\frac{|\partial A|}{|A|}\sqrt{\sum_{v\in V}\eta_v^2(\xi_v^2)^\dagger}\sqrt{\log(|-\mathcal{B}|)} \leq \frac{1}{\sqrt{p}}\sqrt{\log(p(p+1))}$$

By Theorem 20 the result follows. □

This Corollary shows that the edge lasso is changepoint sparsistent when the clusters are easily separable before the observations $\mathbf{y}$ are made.

3.4 Discussion

As we have seen in Corollary 23, it is quite easy to find cases in which changepoint sparsistency is impossible. This illustrates precisely the difficulty in expecting an estimator to precisely determine $\text{supp}(\nabla\mathbf{x})$. While the aforementioned results illustrate the difficulty of obtaining exactly the changepoint sparsity pattern, it is not clear what can be said about approximate recovery of $\mathbf{x}$ through the edge lasso. While some simple adaptations of the primal-dual witness method seem plausible, an exact analysis of approximate recovery is not within the scope of this thesis. We have seen that recovery guarantees for the edge lasso are difficult and nuanced. In the following

Chapter, we will take a more principled approach to a more fundamental problem. We will form a generalized likelihood ratio test for the detection of graph structured signals. This will lead to the development of new statistical methods, the spectral scan statistic and the graph ellipsoid scan statistic.

Chapter 4

Spectral Scan Relaxations for Estimation, Localization and Detection

In this chapter we are concerned with the basic but fundamental task of deciding whether a given graph, over which a noisy signal is observed, contains a cluster of anomalous or activated nodes comprising an induced subgraph. As we have discussed, such a problem is highly relevant in a variety of scientific areas, such as surveillance, disease outbreak detection, biomedical imaging, sensor network detection, gene network analysis, environmental monitoring and malware detection over a computer network. Recent theoretical contributions in the statistical literature (see, e.g., [1, 3, 4, 5]) have detailed the inherent difficulty of such testing problems in relatively simplified settings and under specific conditions on the graph topology. From a practical standpoint, the natural algorithm for detection of anomalous clusters of activity in graphs is the generalized likelihood ratio test (GLRT) or scan statistic, a computationally intensive procedure that entails scanning all clusters with low surface area and testing individually for anomalous activation. Unfortunately, its performance over general graphs is not well understood, and little attention has been paid to determining alternative, computationally tractable, procedures.

We assume that the class of clusters of constant signal consists of sub-graphs of small cut size. We believe this is a natural and realistic assumption which, as we demonstrate below, allows us to explicitly incorporate into the detection problem the properties of the graph topology through its spectrum. In particular, we show that the GLRT is an integer program with a term in the objective that corresponds to the sparsest cut in a graph, a known NP-hard problem [57]. With this in mind, we propose two relaxations of the GLR, called the spectral scan statistic (SSS) and graph ellipsoid scan statistic (GESS), which are based on the combinatorial Laplacian of the graph and, importantly, are computationally efficient programs. We summarize our results as follows:

1. **Balanced Graph Structured Goodness-of-Fit Tests:** We define a few new classes of signals based on the notion of small cut size that reflects in a natural way the topological properties of the graph.
2. **Graph Ellipsoid Scan Statistic:** We analyze the corresponding GLR statistics and show that it is, in fact, related to the problem of finding sparest cuts. We then develop two computationally efficient relaxations of the GLR statistic, called the spectral scan statistic

and the graph ellpsiod scan statistic, and analyze their properties. In our main theoretical result, we show that the performance of the scan statistics depend explicitly on the spectral properties of the graph.

3. **Adaptive Graph Ellipsoid Scan Statistic:** While the GESS requires the specification of the cut sparsity parameter ρ, we develop a competetive procedure that adapts to ρ.
4. **Localization with the Spectral Scan Statistic:** The SSS can be used for localization, for which we develop some theoretical guarantees.
5. **Specific Graph Examples and Comparisons:** Using such results we are able to characterize in a very explicit form the performance of the GESS on a few notable graph topologies and demonstrate its superiority over naive detectors, such as the aggregate and max statistics.

4.1 Balanced Graph Structured Goodness-of-Fit Tests

Detection, also known as goodness-of-fit testing, involves the fundamental statistical question: are we observing merely noise or is there some signal amidst this noise? While the basic detection problem is to determine if the signal $\mathbf{x}$ is constant or not, we analyze four different alternatives: unstructured signal, piece-wise constant graph-structured signal, graph-structured signal with differential activation, and the graph-structured ellipsoid.

Unstructured H_1. Let $\mathcal{X}_U(\mu) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \bar{\mathbf{x}}\| \geq \mu\}$, the complement of the open ball of radius μ in the subspace orthogonal to $\mathbf{1}$. Then the basic 'unstructured' detection problem is testing the null hypothesis against the alternative hypothesis:

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \quad \text{v.s.} \quad H_1^U : \mathbf{x} \in \mathcal{X}_U(\mu)$$

Because there is no a priori structure, i.e. no relationships between the indices, in this setting we are compelled to use tests that are invariant under arbitrary permutation of the indices of $\mathbf{y}$ and $\mathbf{x}$. We will now outline the graph-structured alternative hypotheses.

Piece-wise constant graph-structured H_1. Let $G = (V, E, W)$ be a connected, possibly weighted graph with $p < \infty$ vertices ($V = [p]$). Hence, for each index of $\mathbf{x}$, and consequently the indices of $\mathbf{y}$, we associate a vertex of G. We have already motivated this problem, detailing situations in which it is appropriate. We will assume that there are two groups of constant signal for $\mathbf{x}$, namely that there exists a subset $C \subset V$ ($C \notin \{\emptyset, V\}$) such that $\mathbf{x}$ is constant within both C and its complement $\bar{C} = V \backslash C$. We consider the function class specific to C,

$$\mathcal{X}_{PC}(\mu, C) = \{\mathbf{x} = \alpha\mathbf{1} + \delta\mathbf{1}_C : \mu, \delta \in \mathbb{R}\} \cap \mathcal{X}_U(\mu)$$

The parameter α can be thought of as the magnitude of the background signal and is a nuisance parameter, while δ quantifies the gap in signal between the two clusters. For the signal $\mathbf{x} = \alpha\mathbf{1} + \delta\mathbf{1}_C$ to be contained in $\mathcal{X}_U(\mu)$, it is required that $\sqrt{\frac{|C||\bar{C}|}{n}}\delta \geq \mu$.

We will not assume any knowledge of the true clustering $(C, \bar{C})$, other than that it belongs to a given class $\mathcal{C} \subset 2^{[p]}$ such that for $C \in \mathcal{C}$, C and $\bar{C}$ are both large and can be easily disconnected,

in that they have low cut size. Formally, we define, for some $\rho > 0$,

$$\mathcal{C} = \mathcal{C}(\rho) = \left\{ C \subset V, C \neq \emptyset \colon \frac{|\partial C|}{|C||\bar{C}|} \leq \frac{\rho}{n} \right\}, \tag{4.1}$$

where $\partial C = \{(i,j) \in E : i \in C, j \in \bar{C}\}$ is the boundary of C. Note that $\mathcal{C}$ is a symmetric class in the sense that $C \in \mathcal{C}$ if and only if $\bar{C} \in \mathcal{C}$. The quantity $\frac{n|\partial C|}{|C||\bar{C}|}$ is known as the **cut sparsity** and is equivalent, up to factor of 2, to the **cut expansion** ($\frac{|\partial C|}{\min\{|C|,|\bar{C}|\}}$):

$$\frac{|\partial C|}{\min\{|C|, |\bar{C}|\}} \leq \frac{n|\partial C|}{|C||\bar{C}|} \leq 2\frac{|\partial C|}{\min\{|C|, |\bar{C}|\}}$$

The groundwork for defining our testing hypotheses is set. Define $\mathcal{X}_{PC}(\mu, \rho) = \cup_{C \in \mathcal{C}(\rho)} \mathcal{X}_{PC}(\mu, C)$. Then we will consider testing the null hypothesis against the alternative hypotheses:

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \quad \text{v.s.} \quad H_1^{PC} : \mathbf{x} \in \mathcal{X}_{PC}(\mu, \rho)$$

Graph-structured H_1. We now consider a more general form of alternative, in which the signal is graph-structured, but not necessarily constant over clusters of activation. We will assume that there is again a true cluster $C \in \mathcal{C}(\rho)$ within which the signal differs little and across which the signal differs highly. Specifically, let

$$\mathcal{X}_S(\mu, \rho) = \left\{ \mathbf{x} \in \mathbb{R}^p : \left| \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|} - \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|} \right| \sqrt{\frac{|C||\bar{C}|}{n}} \geq \mu, C \in \mathcal{C} \right\}$$

Notice that if $\mathbf{x} \in \mathcal{X}_S(\mu, \rho)$ then $\|\mathbf{x} - \bar{\mathbf{x}}\| \geq \mu$, so $\mathcal{X}_S(\mu, \rho) \subset \mathcal{X}_U(\mu)$. Furthermore, if $\mathbf{x} = \alpha \mathbf{1} + \delta \mathbf{1}_C \in \mathcal{X}_{PC}(\mu, \rho)$ then

$$\left| \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|} - \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|} \right| = \delta$$

Hence, $\mathcal{X}_{PC}(\mu, \rho) \subset \mathcal{X}_S(\mu, \rho)$. This induces the following hypothesis testing framework:

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \quad \text{v.s.} \quad H_1^{S} : \mathbf{x} \in \mathcal{X}_S(\mu, \rho)$$

Whenever possible we will make statements about this non-constant alternative, for the sake of generality.

Graph-structured ellipsoidal H_1. It will be convenient to introduce a final alternative hypothesis, that is a relaxation of the piece-wise constant graph-structured signals. Define the ellipsoidal alternative space (it is in fact an ellipsoidal cone) to be

$$\mathcal{X}_E(\mu, \rho) = \left\{ \mathbf{x} \in \mathbb{R}^p : \mathbf{z}^\top \Delta \mathbf{z} \leq \rho, \mathbf{z} = \frac{\mathbf{x}}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \right\} \cap \mathcal{X}_U(\mu)$$

This generates the hypothesis test,

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \quad \text{v.s.} \quad H_1^{E} : \mathbf{x} \in \mathcal{X}_E(\mu, \rho)$$

We see that for any $\mathbf{x} = \alpha\mathbf{1} + \delta\mathbf{1}_C \in \mathcal{X}_{PC}(\mu, \rho)$, (here α' is some value)

$$\mathbf{z} = \frac{\mathbf{x}}{\|\mathbf{x} - \bar{\mathbf{x}}\|} = \alpha'\mathbf{1} + \frac{\delta}{\|\mathbf{x} - \bar{\mathbf{x}}\|}\mathbf{1}_C = \alpha'\mathbf{1} + \sqrt{\frac{n}{|C||\bar{C}|}}\mathbf{1}_C$$
$$\Rightarrow \mathbf{z}^\top \Delta \mathbf{z} = \frac{n|\partial C|}{|C||\bar{C}|} \leq \rho$$

Hence, $\mathcal{X}_{PC}(\mu, \rho) \subset \mathcal{X}_E(\mu, \rho)$. The same analysis cannot be repeated for the more general graph-structured signals $\mathcal{X}_S(\mu, \rho)$. We summarize these results.

Proposition 25. *The alternative spaces are nested in the following way:*

(a) $\mathcal{X}_{PC}(\mu, \rho) \subset \mathcal{X}_S(\mu, \rho) \subset \mathcal{X}_U(\mu)$

(b) $\mathcal{X}_{PC}(\mu, \rho) \subset \mathcal{X}_E(\mu, \rho) \subset \mathcal{X}_U(\mu)$

4.2 A Lower Bound and Classical Results

It is our ultimate goal to give a theory of changepoint detection on general graphs. This means that our theorems should apply to all graph structures with only minor, simplifying assumptions, such as connectedness. But as a validation of the theory and methods that we propose, we will pay particular attention to the implications of this on canonical graph structures. We begin by introducing the torus graph structure that will serve as a running example for illustrations. The reader should in no way interpret this to mean that our results necessarily depend on the idiosyncrasies of the torus graph, such as edge transitivity.

Example 1. *(Torus Graph) The $n \times n$ torus graph can be embedded in $V = (\mathbb{Z} \mod n)^2$ where points $(i_1, i_2), (j_1, j_2)$ are connected by an edge if and only if $|i_1 - j_1 \mod n| + |i_2 - j_2 \mod n|$.The class of clusters in the torus under consideration $\mathcal{C}(\rho)$ are those that have sparsity $n|\partial C|/(|C||\bar{C}|) \leq \rho$. For example, rectangles of size $k \times k$ within the torus have sparsity $4kn/(k^2(n-k^2)) \approx 4/k$ for $k^2 \ll n$. Meaning that if we would like to include rectangles of size $k \times k$, it is sufficient that $\rho = 4/k$, and for the largest rectangles we require $\rho = 4/\sqrt{n}$.*

In order to understand the fundamental limitations of the changepoint detection problem, we can provide lower bounds on the performance of any procedure. After this negative result, we will review classical theory about the detection of non-zero means under no constraints and in infinite ellipsoids. While the ellipsoid result can't be exactly ported to our setting, it is the most similar known result to that which we will develop. These will foreshadow the critical SNR scaling that we will discover with the graph ellipsoid scan statistic.

4.2.1 Lower Bound

The first lower bound that we will provide is a simple bound on the performance of the oracle based on the Neyman-Pearson lemma. The second result is more sophisticated and requires that the graph and the ρ parameter allows for unions of disjoint clusters. It requires that the graph has symmetries that we can exploit, but it will not be satisfied by most graphs. We will later show that the conditions are satisfied by the specific graph structures that we will analyze.

Theorem 26. *(a) H_0 and H_1^{PC}, H_1^S, H_1^E are asymptotically indistinguishable if $\mu/\sigma = o(1)$. (b) Suppose that there is a subset of clusters $\mathcal{C}' \subseteq 2^V$ such that all the elements of $\mathcal{C}'$ are disjoint, of the same size ($|C| = c$ for all $C \in \mathcal{C}'$), and*

$$\forall C \in \mathcal{C}', \quad \frac{n|\partial C|}{|C||\bar{C}|} \leq \frac{\rho}{2}$$

i.e., elements of $\mathcal{C}'$ belong to the alternative hypothesis with $\rho/2$ cut sparsity. Furthermore assume that $\frac{c|\mathcal{C}'|}{n} \to 1$. H_0 and H_1^{PC}, H_1^S, H_1^E are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o(|\mathcal{C}'|^{1/4})$$

Proof. Notice that if H_0 is indistinguishable from H_1^{PC} then so is it indistinguishable from H_1^S and H_1^E. For the proof let $\sigma = 1$ as it can be absorbed into the signal size μ. Let the risk be the sum of the probabilities of type 1 and type 2 error. Notice that the risk can be bounded by

$$\sup_{\mathbf{x}=\bar{\mathbf{x}}} \mathbb{E}_{\mathbf{x}} T(\mathbf{y}) + \sup_{\mathbf{x} \in \mathcal{X}_{PC}(\mu,\rho)} \mathbb{E}_{\mathbf{x}}[1 - T(\mathbf{y})] \geq \mathbb{E}_{\mathbf{x}=\mathbf{0}} T(\mathbf{y}) + \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \mathbb{E}_{\mathbf{x}^S}[1 - T(\mathbf{y})] = R^*$$

where $\mathbf{x}^S = \mu\sqrt{\frac{n}{|S||\bar{S}|}}\mathbf{1}_S$ and $\mathcal{S} \subseteq \mathcal{C}$. Then by Proposition 3.2 in [1],

$$R^* \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}\exp\left\{\frac{\mu^2}{\sigma^2}Z\right\} - 1}$$

where

$$Z = \frac{p|S \cap S'|}{\sqrt{|S||\bar{S}||S'||\bar{S}'|}}$$

for S, S' drawn independently uniformly from $\mathcal{S}$. Let $k = \lfloor\sqrt{\lfloor p/(2|C|)\rfloor}\rfloor$, and let S be the union of k elements of $\mathcal{C}$ chosen uniformly at random without replacement (and let S' be an IID copy of S). Notice that $|S| \leq p/2$ and because for all $C \in \mathcal{C}$, $|\partial C|/(|C||\bar{C}|) \leq \rho$ then $|\partial S|/(|S||\bar{S}|) \leq \rho$. Moreover, Z is stochastically bounded by $|K \cap K'|/\sqrt{|K||K'|}$ where K, K' are chosen independently, uniformly from all k-sets of $1, \ldots, m = \lfloor p/|C|\rfloor$. Hence, we can apply Proposition 3.4 from [1] and determine that $R^* > \delta$ if

$$\frac{\mu}{\sigma\sqrt{k}} \leq \sqrt{\log\left(1 + \frac{m\log(1 + 4(1-\delta)^2)}{k^2}\right)}$$

Because $k^2 \asymp m$ we have asymptotic indistinguishability if $\mu/\sigma = o(\sqrt{k}) = o(m^{1/4}) = o((p/|C|)^{1/4})$. □

We illustrate the usefulness of the lower bound with the following example.

Example 2. *(Lower Bound for Torus) We will construct $\mathcal{C}'$ in Theorem 26 (b) from disjoint squares of size a constant multiple of $p^{1-\beta}$ making $|\mathcal{C}'| \asymp p^\beta$. Thus, the critical SNR for H_0 versus any of H_1^{PC}, H_1^E, H_1^S for any estimator is greater than $p^{\beta/4}$.*

This scaling with the $1/4$th power is not a coincidence. We will see that the classical results, both upper and lower bounds, for the unconstrained and ellipsoid constrained alternative hypotheses also provide a critical SNR of this form.

4.2.2 Classical Results

In order to understand the inherent difficulty of distinguishing H_0 from H_1^U, we will recount a result from [36].

Theorem 27. *The critical SNR for any test distinguishing H_0 from H_1^U is given by,*

$$\frac{\mu}{\sigma} \asymp p^{1/4}$$

and it is achieved by the energy test statistic $\|\mathbf{y} - \bar{\mathbf{y}}\|_2^2$.

This result highlights the aforementioned $1/4$th power scaling in critical SNRs. Because of the nested nature of the alternatives, this critical SNR, one would hope that this upper bounds that of any tests that we perform. We will see that this is always achieved by the graph ellipsoid scan statistic, and by it's adaptive version in most cases.

Consider the statistics $\max_{i\in[p]} |y_i - \bar{\mathbf{y}}|$ and $\|\mathbf{y} - \bar{\mathbf{y}}\|_1$, which we will call the *max* statistic and the *aggregate* statistic respectively. Then they have the following critical SNR's for the piecewise alternative structure, H_1^{PC}.

Theorem 28. *(a) Let $c_{\max} = \max_{C\in\mathcal{C}} \min\{|C|, |\bar{C}|\}$. Then if $\log c_{\max} \ll \log p$ then the critical SNR for H_0 versus H_1^{PC} of the max statistic is between the following*

$$\frac{\mu}{\sigma} = \omega(\sqrt{c_{\max}}), \quad \frac{\mu}{\sigma} = o(\sqrt{c_{\max} \log p})$$

while the upper bound is an equality ($\asymp$) if $\log c_{\max} = o(\log p)$.
(b) Let $c_{\min} = \min_{C\in\mathcal{C}} \max\{|C|, |\bar{C}|\}$. The critical SNR for H_0 versus H_1^{PC} of the aggregate test statistic is

$$\frac{\mu}{\sigma} \asymp \sqrt{\frac{p}{c_{\min}}}$$

Remark 29. *Notice that the critical SNR of the max statistic, Theorem 28 (a), can be significantly worse than the energy and aggregate statistics if $c_{\max} \geq \sqrt{p}$. Otherwise the Max statistic is superior. Similarly, (b) provides worse performance than the energy and max statistic if $c_{\min} \leq \sqrt{p}$. For our examples, it will be the case that $c_{\max} \asymp p$ and $c_{\min}$ will be significantly smaller. We can see that for the torus $c_{\max} \asymp p$ and $c_{\min} \asymp p^{1-\beta}$. Moreover, these only hold for the piecewise constant graph structure of H_1^{PC}.*

A closely related testing setting is the *ellipsoid sequence space model*. The main difference between this setting and ours is that the minimax estimator developed in [18] has been shown to be optimal for the setting where the alternative is a fixed infinite-dimensional ellipsoid. The asymptotic statements are made with respect to a decreasing noise variance sequence. To make this precise let the sequence space model be,

$$y_i = \mathcal{N}(x_i, \sigma^2)$$

independent over $i \in \mathbb{Z}_+$. The null and alternative hypotheses are

$$H_0 : x_i = 0, \forall i \in \mathbb{Z}_+ \quad \text{and } H_1^{Er} : \sum_{i=1}^{\infty} a_i x_i^2 \leq 1 \text{ and } \sum_{i=1}^{\infty} x_i^2 \geq \mu^2$$

Ermakov's test statistic is similar to the Pinsker filter in that it applies a linear shrinkage to the observations. Let b_0, b_1 be the solutions to the following equations,

$$\sum_{i=1}^{\infty} a_i(b_0 - b_1 a_i)_+ = 1 \quad \text{and} \quad \sum_{i=1}^{\infty}(b_0 - b_1 a_i)_+ = \mu^2$$

Then the test statistic is

$$\hat{e} = \sigma^{-4} \sum_{i=1}^{\infty} y_i^2 (b_0 - b_1 a_i)_+$$

The following is the relevant portion of the main theorem in [18].

Theorem 30. *Under the ellipsoid sequence space model, the critical SNR achieved by $\hat{e}$ for the hypotheses H_0 and H_1^{Er} is any μ/σ that satisfies,*

$$\sigma^4 \asymp (b_0\mu - b_1\rho)$$

Unfortunately, a convenient form for this SNR is only known for certain forms of $\{a_i\}$, such as the Sobolev-type ellipsoids. Incidentally, the graph ellipsoid scan statistic will take a weighted χ^2 form, but it's derivation will be as a relaxation of the generalized likelihood ratio test. We are able to relate the graph ellipsoid setting to this setting by rotating the signal, $\mathbf{x} \leftarrow \mathbf{U}^\top \mathbf{x}$. We see that in order to even compare our ellipsoid alternative H_1^E to this setting we would need a very special form for the eigenvalues of Δ. At the very least, it would need to happen that there would be some sequence of scalars α_n and a fixed sequence $\{a_i\}$, such that $\{\alpha_n \lambda_i\} \to \{a_i\}$ uniformly. This of course is an overly restrictive assumption when we are considering arbitrary sequences of graph structures. We will compare the performance of Ermakov's statistic to the GESS for specific graph models in a later section.

4.3 Graph Ellipsoid Scan Statistic

In order to derive the Graph Ellipsoid Scan Statistic (GESS), we will consider specifically the piece-wise constant graph structured H_1^{PC}. Before we arrive at the GESS, we must derive the Spectral Scan Statistic (SSS). The SSS was derived originally in [69], but since its publication the authors derived the graph ellipsoid scan statistic. While the GESS is shown to be a relaxation of the SSS, we favor the GESS because it is simple to implement, performs as well as the SSS in practice, and admits a superior theoretical analysis. The superior theoretical analysis has practical implications because p-values and power curves can be derived analytically.

4.3.1 Derivation of GESS

The hypothesis testing problem, with signal in $\mathcal{X}_{PC}(\mu, \rho)$, at hand presents two challenges: (1) the model contains an unbounded nuisance parameter $\alpha \in \mathbb{R}$ and (2) the alternative hypothesis is comprised of a finite disjoint union of composite hypotheses indexed by $\mathcal{C}$. These features set our problem apart from virtually all existing work of structured normal means problems (see, e.g. [1, 3, 4, 5]), which does not consider nuisance parameters and relies on a simplified framework

consisting of a simple null hypothesis and a composite hypothesis consisting of disjoint unions of simple alternatives. Having nuisance parameters and composite hypothesis require a more sophisticated analysis.

We will eliminate the interference caused by the nuisance parameter by considering test procedures that are independent of α (or equivalently $\bar{\mathbf{x}}$). The formal justification for this choice is based on the theory of optimal invariant hypothesis testing (see, e.g., [48]) and of uniformly best constant power tests (see [83]). Due to space limitations we will not provide the details and refer the reader to [7, 22, 23, 25, 26, 67] and references therein for in depth-treatments of these issues related to the model a hand.

For the simpler problem of testing $\mathbf{x} = \bar{\mathbf{x}}$ versus $\mathbf{x} \in \mathcal{X}_{PC}(\mu, C)$ for some $C \subset V$, the optimal test is based on the likelihood ratio (LR) statistic (see the proof of Lemma 31 below for a derivation)

$$2 \log \Lambda_C(\mathbf{y}) = \log \left(\frac{\sup_{\mathbf{x} \in \mathcal{X}_{PC}(\mu, C)} f_{\mathbf{x}}(\mathbf{y})}{\sup_{\mathbf{x} = \bar{\mathbf{x}}} f_{\mathbf{x}}(\mathbf{y})} \right) = \frac{1}{\sigma^2} \frac{p}{|C||\bar{C}|} \left(\sum_{v \in C} \tilde{\mathbf{y}}_v \right)^2, \tag{4.2}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}} = (\tilde{\mathbf{y}}_v, v \in V)$ and f_θ is the Lebesgue density of P_θ. This test rejects H_0 for large values of $\Lambda_C(\mathbf{y})$. Optimality follows from the fact that the statistical model we consider has the monotone likelihood ratio property.

When testing against composite alternatives, like in our case, it is customary to consider instead the generalized likelihood ratio (GLR) statistic, which in our case reduces to

$$\hat{g} = \max_{C \in \mathcal{C}(\rho)} 2\sigma^2 \log \Lambda_C(\mathbf{y}).$$

Through manipulations of the likelihoods, we find that the GLR statistic has a very convenient form which is tied to the spectral properties of the graph G via its Laplacian.

Lemma 31. *Let* $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{1}(\frac{1}{p} \sum_{v \in V} \mathbf{y}_v)$ *and* $\mathbf{K} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$. *Then*

$$\hat{g} = \max_{\mathbf{x} \in \{0,1\}^p} \frac{\mathbf{x}^\top \tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{K} \mathbf{x}} \text{ s.t. } \frac{\mathbf{x}^\top \Delta \mathbf{x}}{\mathbf{x}^\top \mathbf{K} \mathbf{x}} \leq \rho, \tag{4.3}$$

where Δ *is the combinatorial Laplacian of the graph* G.

Proof. First we notice that H_0 and H_1^C are invariant to a group action that will allow us to simplify the LR statistic. Let the group action on $\mathbf{x} \in \mathbb{R}^p$ be

$$G = \{\mathbf{x} \mapsto \mathbf{x} + \Delta : \frac{1}{|C|} \sum_{v \in C} \Delta_v = \frac{1}{|\bar{C}|} \sum_{v \in \bar{C}} \Delta_v = 0, \Delta \in \mathbb{R}^p\}$$

We now are able to obtain the maximal invariant statistics to G, $\mathbf{y}_0 = \frac{1}{|C|} \sum_{i \in C} \mathbf{y}_i \sim N(\beta_0, \sigma_0^2)$ and $\mathbf{y}_1 = \frac{1}{|\bar{C}|} \sum_{i \in \bar{C}} \mathbf{y}_i \sim N(\beta_1, \sigma_1^2)$ for $\sigma_0 = \sigma / \sqrt{|C|}$ and $\sigma_1 = \sigma / \sqrt{|\bar{C}|}$. By the theory of most powerful invariant tests we can reduce the LR statistic for $\mathbf{y}$ to it's maximal invariants $\mathbf{y}_0, \mathbf{y}_1$. Then, we obtain

$$2 \log \Lambda_C(\mathbf{y}) = \frac{1}{\sigma_0^2} (\mathbf{y}_0 - \hat{\beta})^2 + \frac{1}{\sigma_1^2} (\mathbf{y}_1 - \hat{\beta})^2$$

where $\hat{\beta} = \frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}\mathbf{y}_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2}\mathbf{y}_1$ is the MLE under H_0. (The likelihood under the alternative balances with the normalizing constant of the null likelihood.) Thus,

$$\begin{aligned}
2\log\Lambda_C(\mathbf{y}) &= \frac{1}{\sigma_0^2}\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2}(\mathbf{y}_0-\mathbf{y}_1)\right)^2 + \frac{1}{\sigma_1^2}\left(\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}(\mathbf{y}_0-\mathbf{y}_1)\right)^2 \\
&= \frac{(\mathbf{y}_0-\mathbf{y}_1)^2}{\sigma_0^2+\sigma_1^2} = \frac{1}{\sigma^2}\frac{|C||\bar{C}|}{|V|}(\mathbf{y}_0-\mathbf{y}_1)^2 \\
&= \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\frac{|\bar{C}|}{|V|}\sum_{v\in C}\mathbf{y}_v - \frac{|C|}{|V|}\sum_{v\in\bar{C}}\mathbf{y}_v\right)^2 \\
&= \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\sum_{v\in C}\mathbf{y}_v - \frac{|C|}{|V|}\sum_{v\in V}\mathbf{y}_v\right)^2 = \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\sum_{v\in C}\tilde{\mathbf{y}}_v\right)^2. \qquad (4.4)
\end{aligned}$$

Now we let $\mathbf{x} = \mathbf{1}_C$, making the statistic above

$$2\sigma^2\log\Lambda_C(\mathbf{y}) = \frac{\mathbf{x}^\top\tilde{\mathbf{y}}\tilde{\mathbf{y}}\mathbf{x}}{\mathbf{x}^\top\mathbf{K}\mathbf{x}} \text{ and } \frac{|\partial C||V|}{|C||\bar{C}|} = \frac{\mathbf{x}^\top\mathbf{L}\mathbf{x}}{\mathbf{x}^\top\mathbf{K}\mathbf{x}}.$$

The result now follows by considering all the indicator functions corresponding to the sets in $\mathcal{C}$. □

The savvy reader will notice the connection between (4.3) and the graph sparsest cut program. By Lagrangian duality, we see that the program (4.3) is equivalent to (for some Lagrangian parameter ν)

$$\min_{C\subseteq V}\frac{|\partial C|}{|C||\bar{C}|} - \nu\frac{(\sum_{i\in C}\tilde{y}_i)^2}{|C||\bar{C}|}$$

the first term of which is precisely the *sparsest cut* objective, and the second term drives the solution C to have positive within cluster empirical correlations. The sparsest cut program is known to be NP-hard, with poly-time algorithms known for trees and planar graphs([57]). Because of this fact, approximate algorithms have been proposed over the past two decades, most notably the uniform multicommodity flow approach of ([49, 72]) and the semi-definite relaxation of the cut metric ([6]). While it is tempting to apply these same relaxation techniques to (4.3), we will see in the Section 4 that the addition of the data driven objective $\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top$ renders these approaches invalid. [32] observed that the minimum cut sparsity is bounded by the algebraic connectivity (λ_2), suggesting the Fiedler vector (i.e. the second eigenvector of Δ) to be an appropriate relaxation of the characteristic vector of the cut. Moreover, the well known Cheeger inequality shows that the minimum cut sparsity (in a regular graph) is bounded by the algebraic connectivity (see [13]). We will follow the tradition of bounding sparsity with the algebraic connectivity, and provide a surrogate estimator to the scan statistic based on this simple spectral relaxation.

Proposition 32. *Define the Spectral Scan Statistic (SSS) as*

$$\hat{s} = \sup_{\mathbf{x}\in\mathbb{R}^n}(\mathbf{x}^\top\tilde{\mathbf{y}})^2 \text{ s.t. } \mathbf{x}^\top\Delta\mathbf{x} \le \rho, \|\mathbf{x}\| \le 1, \mathbf{x}^\top\mathbf{1} = 0.$$

Then the GLR statistic is bounded by the SSS: $\hat{g} \le \hat{s}$.

Proof. First let us notice that $\mathbf{K} = \mathbf{I} - \frac{1}{p}\mathbf{1}\mathbf{1}^\top$ is the projection onto the subspace orthogonal to $\mathbf{1}$. Because $\mathbf{K}$ is thus idempotent, $\tilde{\mathbf{y}}\mathbf{1} = 0$, and $\Delta\mathbf{1} = 0$ we can rewrite

$$\hat{g} = \max_{\mathbf{x}\in\{0,1\}^p\setminus\{\mathbf{0},\mathbf{1}\}} \frac{(\mathbf{K}\mathbf{x})^\top\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top(\mathbf{K}\mathbf{x})}{(\mathbf{K}\mathbf{x})^\top(\mathbf{K}\mathbf{x})} \text{ s.t. } \frac{(\mathbf{K}\mathbf{x})^\top\Delta(\mathbf{K}\mathbf{x})}{(\mathbf{K}\mathbf{x})^\top(\mathbf{K}\mathbf{x})} \leq \rho$$

So, we have the following relaxation,

$$\hat{g} \leq \max_{\mathbf{x}\neq 0,\mathbf{x}^\top\mathbf{1}=0} \frac{\mathbf{x}^\top\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \text{ s.t. } \frac{\mathbf{x}^\top\Delta\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \rho = \hat{s}$$

□

Notice that because the domain $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^\top\Delta\mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top\mathbf{1} = 0\}$ is symmetric around the origin, this is precisely the square of the solution to

$$\sqrt{\hat{s}} = \sup_{\mathbf{x}\in\mathbb{R}^p} \mathbf{x}^\top\mathbf{y} \text{ s.t. } \mathbf{x}^\top\Delta\mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top\mathbf{1} = 0, \tag{4.5}$$

where we have used the fact that $\mathbf{x}^\top\tilde{\mathbf{y}} = ((\mathbf{I} - \frac{1}{p}\mathbf{1}\mathbf{1}^\top)\mathbf{x})^\top\mathbf{y} = \mathbf{x}^\top\mathbf{y}$ because $\mathbf{x}^\top\mathbf{1} = 0$ within $\mathcal{X}$.

Proposition 33. *Define the Graph Ellipsoid Scan Statistic (GESS) as*

$$\hat{t} = \sum_{i=2}^{p} \min\{1, \frac{\rho}{\lambda_i}\}[(\mathbf{u}_i^\top\mathbf{y})^2 - 1]$$

The SSS as a function of ρ can be bounded above and below in the following:

$$\hat{t} + \sum_{i=2}^{p} \min\{1, \frac{\rho}{\lambda_i}\} \leq \hat{s} \leq 2(\hat{t} + \sum_{i=2}^{p} \min\{1, \frac{\rho}{\lambda_i}\})$$

Proof. Let $a_i = \lambda_i/\rho$. We have shown that

$$\|\mathbf{z}\| \leq 1, \mathbf{z}^\top\mathbf{A}\mathbf{z} \leq 1 \Rightarrow \sum_{i>1} \max\{1, a_i\}z_i^2 \leq 2$$

It is also the case that

$$\sum_{i>1} \max\{1, a_i\}z_i^2 \leq 1 \Rightarrow \|\mathbf{z}\| \leq 1, \mathbf{z}^\top\mathbf{A}\mathbf{z} \leq 1$$

Hence,

$$\{\mathbf{U}\mathbf{z} \in \mathbb{R}^p \colon \mathbf{z} \in \mathbb{R}^{p-1}, \sum_{i=2}^{p} \max\{1, a_i\}z_i^2 \leq 1\} \subseteq \{\mathbf{U}\mathbf{z} \in \mathbb{R}^p \colon \mathbf{z} \in \mathbb{R}^{p-1}, \frac{1}{\rho}\mathbf{z}^\top\Lambda\mathbf{z} \leq 1, \mathbf{z}^\top\mathbf{z} \leq 1\}$$

$$\subseteq \{\mathbf{U}\mathbf{z} \in \mathbb{R}^n \colon \mathbf{z} \in \mathbb{R}^{p-1}, \sum_{i=2}^{p} \frac{\max\{1, a_i\}}{2}z_i^2 \leq 1\}$$

If we denote $\mathbf{A}' = \text{diag}\{\max\{1, a_i\}\}_{i=2}^p$, then it is the case that

$$\sup_{\mathbf{x}^\top\mathbf{U}\mathbf{A}'\mathbf{U}^\top\mathbf{x}\leq 1} \tilde{\mathbf{y}}^\top\mathbf{x} \leq \sqrt{\hat{s}} \leq \sup_{\mathbf{x}^\top\mathbf{U}\mathbf{A}'\mathbf{U}^\top\mathbf{x}\leq 2} \tilde{\mathbf{y}}^\top\mathbf{x}$$

Evaluating these programs (one can use the Fenchel dual of indicators over ellipsoids) gives us the lemma. □

While it is not obvious that it is important that the GESS is a relaxation of the GLRT, it is a natural starting point for the development of a suite of changepoint detectors on graphs. Understanding its performance is an important step in the possibly long process of characterizing the difficulty of distinguishing H_0 from H_1^{PC}. It is not clear, even if it is possible to obtain a poly-time algorithm that can distinguish H_0 from H_1^{PC} over *any* graph under the critical SNR regime. Before we get lost in such speculation, let us return to the GESS and show what it does on our torus example.

Example 3. *(GESS for the Torus) By a simple Fourier analysis (see [71]), we know that the Laplacian eigenvalues are* $2(2 - \cos(2\pi i_1/\ell) - \cos(2\pi i_2/\ell))$ *for all* $i_1, i_2 \in [\ell]$*. The eigenvectors correspond to that of the discrete Fourier transform. So the GESS for the Torus graph corresponds to linear shrinkage in the frequency domain.*

In order to quantify the uncertainty inherent in this testing problem, we develop a theory of the GESS.

4.3.2 Theoretical Analysis of GESS

A thorough theoretical analysis of the GESS has several uses. Information theoretic results, such as Corollary 37, in which we characterize the critical signal-to-noise ratio, enable us to determine the strength of the GESS as a detector on theoretical grounds. More statistical results, such as Corollary 36, give the practicing statistician ways to quantify the uncertainty inherent in the GESS. We give the practitioner a simple way to calculate p-values and power curves without need for computationally expensive simulations from the null hypothesis.

The following main result bounds the test statistic under H_0 and under the piece-wise constant (H_1^{PC}), general graph structured (H_1^S), and the graph ellipsoid (H_1^E) alternatives. It is based on the concentration of weighted sums of independent χ^2 random variables found in [46].

Theorem 34. *Under the null hypothesis,* H_0*, with probability at least* $1 - \delta_0$*,*

$$\hat{t} \leq 2\left(\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\} \log(1/\delta_0)} + \log(1/\delta)\right) \tag{4.6}$$

Under the alternative hypotheses, H_1^{PC}, H_1^S, H_1^E*, with probability at least* $1 - \delta_1$*,*

$$\hat{t} \geq \frac{\mu^2}{2\sigma^2} - \frac{2\mu}{\sigma}\sqrt{\log(2/\delta_1)} - 2\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\} \log(2/\delta_1)} \tag{4.7}$$

Proof. We will use the following lemma regarding the concentration of χ^2 random variables.

Lemma 35 ([46])**.** *Let for* $i \in \{2, \ldots, p\}$*,* $a_i \geq 0$ *and* $\{X_i\}_{i=1}^p$ *be independent* χ_1^2 *random variables. Define* $Z = \sum_{i=1}^p a_i(X_i - 1)$

$$\mathbb{P}\{Z \geq 2\|\mathbf{a}\|_2\sqrt{x} + 2\|\mathbf{a}\|_\infty x\} \leq e^{-x}$$
$$\mathbb{P}\{Z \leq -2\|\mathbf{a}\|_2\sqrt{x}\} \leq e^{-x}$$

The probability of error under the null, (4.6), follows from Lemma 35. Consider any of the alternatives, then $\hat{t}$ can be written,

$$\hat{t} = \mathbf{y}^\top \mathbf{A}\mathbf{y} - \operatorname{tr} \mathbf{A} = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{x}^\top \mathbf{A}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} - \operatorname{tr} \mathbf{A}$$

where $\mathbf{A} = \mathbf{U}\operatorname{diag}(\{\min\{1, \rho/\lambda_i\}\}_{i=2}^p)\mathbf{U}^\top$. By Gaussian concentration, with probability at least $1 - \delta_1$,

$$\mathbf{x}^\top \mathbf{A}\boldsymbol{\epsilon} \geq -\sqrt{2\mathbf{x}^\top \mathbf{A}^2 \mathbf{x} \log(1/\delta_1)}$$

Because $\mathbf{A}$ is PSD and symmetric, we have that $\mathbf{x}^\top \mathbf{A}^2 \mathbf{x} \leq \mathbf{x}^\top \mathbf{A}\mathbf{x}$. We will now show that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq \mu^2/2$ under H_1^{PC}, H_1^S, H_1^E. Recall that by Fenchel conjugacy,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sup_{\mathbf{z}^\top \mathbf{A}^\dagger \mathbf{z} \leq 1} (\mathbf{z}^\top \mathbf{x})^2$$

Case 1: H_1^{PC}. In this case,

$$\frac{(\mathbf{x} - \bar{\mathbf{x}})^\top}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \Lambda \frac{(\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \leq \rho$$

while $\|(\mathbf{x} - \bar{\mathbf{x}})/\|\mathbf{x} - \bar{\mathbf{x}}\|\| = 1$. Thus,

$$\frac{(\mathbf{x} - \bar{\mathbf{x}})^\top}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \mathbf{A}^\dagger \frac{(\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \leq 2$$

because $\mathbf{A}^\dagger = \mathbf{U}\operatorname{diag}(\{\max\{1, \lambda_i/\rho\}\}_{i=2}^p)\mathbf{U}^\top$. So,

$$\frac{(\mathbf{x} - \bar{\mathbf{x}})^\top}{\sqrt{2}\|\mathbf{x} - \bar{\mathbf{x}}\|} \mathbf{A} \frac{(\mathbf{x} - \bar{\mathbf{x}})}{\sqrt{2}\|\mathbf{x} - \bar{\mathbf{x}}\|} \leq 1$$

$$\Rightarrow \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq \left(\frac{(\mathbf{x} - \bar{\mathbf{x}})^\top}{\sqrt{2}\|\mathbf{x} - \bar{\mathbf{x}}\|} \mathbf{x} \right)^2 = \|\mathbf{x} - \bar{\mathbf{x}}\|^2/2 \geq \mu^2/2$$

Case 2: H_1^S. Let $\mathbf{x} \in \mathcal{X}_S(\mu, C)$. In this case we will let $\mathbf{K}_C$ be the projection onto the span of $\mathbf{1}_C, \mathbf{1}_{\bar{C}}$ and orthogonal to $\mathbf{1}$. So, $\mathbf{K}_C\mathbf{x} = \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|}\mathbf{1}_C + \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|}\mathbf{1}_{\bar{C}} - \bar{\mathbf{x}}$. Let $\mathbf{z} = \mathbf{K}_C\mathbf{x}/\|\mathbf{K}_C\mathbf{x}\|$ that $\mathbf{z}^\top \mathbf{x} = \|\mathbf{K}_C\mathbf{x}\|$. Let $\bar{\mathbf{x}}_C = \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|}\mathbf{1}_C$ and $\bar{\mathbf{x}}_{\bar{C}} = \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{\bar{C}|}\mathbf{1}_{\bar{C}}$.

$$\begin{aligned}
\bar{\mathbf{x}}_C - \bar{\mathbf{x}} &= (\frac{1}{|C|} - \frac{1}{n})\mathbf{1}_C^\top \mathbf{x} - \frac{1}{n}\mathbf{1}_{\bar{C}}^\top \mathbf{x} \\
&= \frac{|\bar{C}|}{n}(\bar{\mathbf{x}}_C - \bar{\mathbf{x}}_{\bar{C}})
\end{aligned}$$

Similarly, $\bar{\mathbf{x}}_{\bar{C}} - \bar{\mathbf{x}} = \frac{|C|}{n}(\bar{\mathbf{x}}_{\bar{C}} - \bar{\mathbf{x}}_C)$. And so,

$$\mathbf{z}^\top \mathbf{x} = \|\mathbf{K}_C\mathbf{x}\| = |C|\frac{|\bar{C}|^2}{n^2}(\bar{\mathbf{x}}_C - \bar{\mathbf{x}}_{\bar{C}})^2 + |\bar{C}|\frac{|C|^2}{n^2}(\bar{\mathbf{x}}_{\bar{C}} - \bar{\mathbf{x}}_C)^2 = \frac{|C||\bar{C}|}{n}(\bar{\mathbf{x}}_{\bar{C}} - \bar{\mathbf{x}}_C)^2 \geq \mu^2$$

Now we can go through the same proof as the previous case substituting $\mathbf{z}$ for $\mathbf{x} - \bar{\mathbf{x}}/\|\mathbf{x} - \bar{\mathbf{x}}\|$.

Case 3: H_1^E. This follows directly from the definition of $\mathcal{X}_E(\mu, \rho)$ and the considerations in Case 1.

The error bound, (4.7) follows from these facts and the Lemma 35 applied to $\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} - \operatorname{tr} A$. □

Theorem 34 shows that by setting a threshold to be the right hand side of (4.6), we have a size δ_0 test. If we then set the right hand side of (4.7) to be this threshold, and solve for μ/σ, then we get the lowest SNR such that the test has power $1-\delta_1$ under the alternative. Equivalently, we can obtain p-values and power curves for the GESS test statistic.

Corollary 36. *By Theorem 34, we have that a valid p-value, $\alpha(\hat{t})$, is (i.e. $\mathbb{P}_0\{\alpha(\hat{t}) \geq \delta_0\} \leq \delta_0$)*

$$\alpha(\hat{t}) = \exp\left\{\frac{1}{2}\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\}} - \frac{1}{2}\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\} + 2\hat{t}}\right\}$$

A valid power curve, $\gamma(\hat{t})$, for H_1^{PC}, H_1^S, H_1^E is (i.e. $\mathbb{P}_1\{\gamma(\hat{t}) \geq \delta_1\} \leq \delta_1$)

$$\gamma(\hat{t}) = \frac{1}{2}\exp\left\{-\frac{1}{2}\left(\frac{\mu^2/2\sigma^2 - \hat{t}}{\mu/\sigma + \sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\}}}\right)^2\right\}$$

While the above corollary is useful for the practicing statistician, the corollary below allows us to compare the GESS to other tests on asymptotic theoretical grounds.

Corollary 37. *The GESS, $\hat{t}$, can asymptotically distinguish H_0 from H_1^{PC}, H_1^S, H_1^E if the SNR is stronger than*

$$\frac{\mu}{\sigma} = \omega\left(\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\}\right)^{1/4}$$

Most notably the critical SNR is lower than $p^{1/4}$ which is the critical SNR enjoyed by the energy test statistic. The most unreasonable assumption that we have made thus far is that the cut sparsity, ρ, is known. This unreasonable advantage is especially apparent when we compare the GESS to the max and aggregate statistics that do not require the knowledge of ρ. The following section develops a test that adapts to ρ.

4.4 Adaptive Graph Ellipsoid Scan Statistic

Notice that our estimators, the SSS and GESS, require that the statistician prespecify the cut sparsity parameter, ρ. While the user may have certain shapes in mind, such as large rectangles in a lattice, it is not reasonable to assume that this can be done for arbitrary graph structure. In order to adapt to ρ we will consider the test statistic, $\hat{t}(\rho)$, as a function of ρ, as it is allowed to vary.

Definition 38. *Let $\delta_0 > 0$ and*

$$\tau(\rho) = 2\left(\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^2}{\lambda_i^2}\}\log((p-1)/\delta_0)} + \log((p-1)/\delta_0)\right)$$

The adaptive GESS test is the test that rejects H_0 if $\exists \rho > 0$ such that

$$\hat{t}(\rho) > \tau(\rho) \tag{4.8}$$

Of course it is not reasonable to assume that the statistician can calculate with finite resources the entire curve $\hat{t}(\rho)$ without some a priori knowledge of its stability as ρ varies. It is our good fortune that it is sufficient to evaluate $\hat{t}(\rho)$ only at $p-1$ points, because $\hat{t}$ is piecewise linear with knots at the eigenvalues and $\tau(\rho)$ is similarly well behaved. Let $k = \max\{k : \lambda_k \leq \rho\}$ then

$$\hat{t}(\rho) = \rho \sum_{i=k+1}^{p} \frac{(\mathbf{u}_i^\top \mathbf{y})^2 - 1}{\lambda_i} + \sum_{i=2}^{k} ((\mathbf{u}_i^\top \mathbf{y})^2 - 1)$$

Hence, $\hat{t}(\rho)$ is piecewise linear with knots at $\{\lambda_i\}_{i=2}^{p}$. Similarly,

$$\tau(\rho) = \sqrt{4\log((p-1)/\delta_0)\left(\rho^2 \sum_{i=k+1}^{p} \lambda_i^{-2} + k\right)} + 2\log((p-1)/\delta_0)$$

Define the following shorthand:

$$A = 4\log((p-1)/\delta_0) \sum_{i=k+1}^{p} \lambda_i^{-2}, \quad B = 4k\log((p-1)/\delta_0)$$

$$D = 2\log((p-1)/\delta_0), \quad E = \sum_{i=k+1}^{p} \frac{(\mathbf{u}_i^\top \mathbf{y})^2 - 1}{\lambda_i}, \quad F = \sum_{i=2}^{k} ((\mathbf{u}_i^\top \mathbf{y})^2 - 1)$$

Then we reject iff

$$\tau(\rho) = \sqrt{\rho^2 A + B} + D < \rho E + F = \hat{t}(\rho)$$

Notice that $A, B, D > 0$, so $\tau(\rho)$ has strictly positive curvature and is convex. Thus, $\tau(\rho) - \hat{t}(\rho)$ is convex within $\lambda_k \leq \rho \leq \lambda_{k+1}$ and has a unique minimum. We can minimize the unrestricted function,

$$\rho^* = \arg\min_{\rho} \sqrt{\rho^2 A + B} + D - \rho E - F$$

and we find that this is attained at

$$\rho^* = \begin{cases} 0, & E^2 \geq A \\ \sqrt{\frac{E^2 B}{A^2 - E^2 A}}, & \text{otherwise} \end{cases}$$

We know by convexity that if $\rho^* < \lambda_k$ then the constrained maximum is attained at λ_k, and if $\rho^* > \lambda_{k+1}$ then it is attained at λ_{k+1}. For each k, we can construct A, B, D, E, F and define

$$\rho_k = \begin{cases} \lambda_k, & E^2 \geq A \text{ or } \sqrt{\frac{E^2 B}{A^2 - E^2 A}} \leq \lambda_k \\ \lambda_{k+1}, & \sqrt{\frac{E^2 B}{A^2 - E^2 A}} \geq \lambda_{k+1} \\ \sqrt{\frac{E^2 B}{A^2 - E^2 A}}, & \text{otherwise} \end{cases}$$

Then the following proposition holds,

Proposition 39. *The adaptive GESS test rejects H_0 if and only if*

$$\exists k \in \{2, \dots, p\}, \quad \tau(\rho_k) < \hat{t}(\rho_k)$$

This proposition has theoretical implications as well as practical. It shows us that we only need to provide a theoretical control of p separate GESS values. We see that Proposition 39 was foreshadowed by the specific form of $\tau(\rho)$ in (4.8) The clever choice of threshold function $\tau(\rho)$ naturally gives us a control on the false alarm (type 1 error).

Theorem 40. *Under H_0, the probability of false alarm (type 1 error) is bounded by*

$$\mathbb{P}_0\{\exists \rho, \hat{t}(\rho) > \tau(\rho)\} \leq \delta_0$$

Consider models from the alternative hypotheses, H_1^{PC}, H_1^S, H_1^E as functions of ρ. Let ρ^* be the smallest such ρ^* such that $\mathbf{x} + \boldsymbol{\epsilon}$ is contained in the alternative hypotheses. Then the probability of type 2 error is bounded by $\delta_1 > 0$ if,

$$\frac{\mu^2}{2\sigma^2} - 2\frac{\mu}{\sigma}\sqrt{2\log(2/\delta_1)} - 2\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^{*2}}{\lambda_i^2}\} \log(2/\delta_1)} > \tau(\rho^*)$$

Proof. The type 1 error control follows directly from a union bound over k,

$$\mathbb{P}_0\{\exists k \in \{2 \dots p\}, \hat{t}(\rho) > \tau(\rho)\} \leq \sum_{k=2}^{p} \mathbb{P}_0\{\hat{t}(\rho_k) > \tau(\rho_k)\} \leq \delta_0$$

because $\tau(\rho)$ was chosen that at any one ρ the probability that $\hat{t}(\rho) > \tau(\rho)$ under H_0 is less than $\delta_0/(p-1)$. The type 2 error control follows by applying Theorem 34 to the GESS test at ρ^*. □

Corollary 41. *The adaptive GESS asymptotically distinguishes H_0 from H_1^{PC}, H_1^S, H_1^E if*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho^{*2}}{\lambda_i^2}\} \log p} + \log p\right)^{1/2}$$

So we are able to make all the same theoretical guarantees with the adaptive GESS as the GESS with only a loss of $(\log p)^{1/4}$. We will now show how this theory is applicable by developing corollaries for different specific graph topologies.

4.5 Localization

It turns out that while the GESS lends itself to superior detection results the SSS can be used to estimate the mean $\mathbf{x}$. It is apparent from its definition in Proposition 32, that the spectral scan statistic attempts to estimate $\mathbf{x}$ modulo $\bar{\mathbf{x}}$ and $\|\mathbf{x} - \bar{\mathbf{x}}\|$. Thus, we will attempt to estimate the mean, modulo it's average, $\mathbf{x} - \bar{\mathbf{x}}$ and its norm. In this section, we will assume that $\mu = \|\mathbf{x} - \bar{\mathbf{x}}\|$ precisely. This should not create much confusion because we are no longer in the hypothesis

testing framework. We will first define the following notion of localization and show that this leads directly to an estimation bound. Suppose that we have estimated $(\mathbf{x}-\bar{\mathbf{x}})/\|\mathbf{x}-\bar{\mathbf{x}}\|$ with $\hat{\mathbf{x}}_S$ then the localization distance is

$$\tilde{d}(\mathbf{x},\hat{\mathbf{x}}_S) = 2\left(1-\frac{(\mathbf{x}-\bar{\mathbf{x}})^\top(\hat{\mathbf{x}}_S - \mathbf{1}^\top\hat{\mathbf{x}}_S\mathbf{1}/p)}{\|\mathbf{x}-\bar{\mathbf{x}}\|\|\hat{\mathbf{x}}_S\|}\right) \tag{4.9}$$

The idea is that we are attempting only to localize the direction of the vector $\hat{\mathbf{x}}_S$. We will consider the following spectral scan localization algorithm

$$\hat{\mathbf{x}}_S = \arg\max \mathbf{x}^\top(\mathbf{y}-\bar{\mathbf{y}}) \text{ s.t. } \mathbf{x}^\top\Delta\mathbf{x} \le \rho, \|\mathbf{x}\| = 1, \mathbf{x}^\top\mathbf{1} = 0 \tag{4.10}$$

Because $\|\hat{\mathbf{x}}_S\| = 1$ we have the following bound on $\tilde{d}(\hat{\mathbf{x}}_S,\mathbf{x})$,

$$\begin{aligned}
\hat{\mathbf{x}}_S^\top\frac{(\mathbf{x}-\bar{\mathbf{x}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} &= \hat{\mathbf{x}}_S^\top\frac{(\mathbf{y}-\bar{\mathbf{y}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} - \hat{\mathbf{x}}_S^\top\frac{(\boldsymbol{\epsilon}-\bar{\boldsymbol{\epsilon}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} \\
&\ge \frac{(\mathbf{x}-\bar{\mathbf{x}})^\top}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\frac{(\mathbf{y}-\bar{\mathbf{y}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} - \hat{\mathbf{x}}_S^\top\frac{(\boldsymbol{\epsilon}-\bar{\boldsymbol{\epsilon}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} \\
&\ge 1 - \left(\frac{(\mathbf{x}-\bar{\mathbf{x}})^\top}{\|\mathbf{x}-\bar{\mathbf{x}}\|} + \hat{\mathbf{x}}_S^\top\right)\frac{(\boldsymbol{\epsilon}-\bar{\boldsymbol{\epsilon}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|} \\
&\ge 1 - \frac{2}{\mu}\sqrt{\hat{s}(\boldsymbol{\epsilon})}
\end{aligned}$$

Here $\hat{s}(\boldsymbol{\epsilon})$ is the SSS evaluated on $\boldsymbol{\epsilon}$. Hence, we obtained

$$\tilde{d}(\mathbf{x},\hat{\mathbf{x}}_S) \le \frac{4}{\mu}\sqrt{\hat{s}(\boldsymbol{\epsilon})}$$

By Proposition 33 we know that

$$\hat{s}(\boldsymbol{\epsilon}) \le 2(\hat{t}(\boldsymbol{\epsilon}) + \sum_{i=2}^{p}\min\{1,\frac{\rho}{\lambda_i}\})$$

Hence, by Theorem 34,

$$\hat{s}(\boldsymbol{\epsilon}) \le 2\sigma^2\left(\sum_{i=2}^{p}\min\{1,\frac{\rho}{\lambda_i}\} + 2\left(\sqrt{\sum_{i=2}^{p}\min\{1,\frac{\rho^2}{\lambda_i^2}\}\log(1/\delta)} + \log(1/\delta)\right)\right)$$

with probability $1-\delta$. For convenience, define $\mathbf{z} = (\mathbf{x}-\bar{\mathbf{x}})/\|\mathbf{x}-\bar{\mathbf{x}}\|$. With an estimate of the unit vector $\mathbf{z}$ at hand, we will proceed to estimate $\mathbf{x}$,

$$\hat{\mathbf{x}} = (\mathbf{y}^\top\hat{\mathbf{x}}_S)\hat{\mathbf{x}}_S + \bar{\mathbf{y}} \tag{4.11}$$

Throughout the following assume that $\sqrt{\hat{s}(\boldsymbol{\epsilon})} = \omega_{\mathbb{P}}(\mu)$.

$$\begin{aligned}
\|\hat{\mathbf{x}}-\mathbf{x}\|^2 &= \|\mathbf{y}^\top\hat{\mathbf{x}}_S\hat{\mathbf{x}}_S - (\mathbf{x}-\bar{\mathbf{x}})\|^2 + \|\bar{\mathbf{y}}-\bar{\mathbf{x}}\|^2 \\
&\le (\|(\mathbf{x}-\bar{\mathbf{x}})^\top\hat{\mathbf{x}}_S\hat{\mathbf{x}}_S - (\mathbf{x}-\bar{\mathbf{x}})\| + |\boldsymbol{\epsilon}^\top\hat{\mathbf{x}}_S|)^2 + \|\bar{\mathbf{y}}-\bar{\mathbf{x}}\|^2 \\
\|(\mathbf{x}-\bar{\mathbf{x}})^\top\hat{\mathbf{x}}_S\hat{\mathbf{x}}_S - (\mathbf{x}-\bar{\mathbf{x}})\| &\le \|\mathbf{x}-\bar{\mathbf{x}}\|\sqrt{2(1-(\mathbf{z}^\top\mathbf{x})^2)} \\
\le \mu\sqrt{2(1-(\min\{0,1-\frac{2}{\mu}\sqrt{\hat{s}(\boldsymbol{\epsilon})}\})^2)} &= O_{\mathbb{P}}\left(\sqrt{\mu\sqrt{\hat{s}(\boldsymbol{\epsilon})}}\right)
\end{aligned}$$

Hence, we can bound the square error by

$$\|\hat{\mathbf{x}} - \mathbf{x}\|^2 = O_{\mathbb{P}}\left((\sqrt{\mu\sqrt{\hat{s}(\boldsymbol{\epsilon})}} + \sqrt{\hat{s}(\boldsymbol{\epsilon})})^2 + \sigma^2\right) = O_{\mathbb{P}}\left(\mu\sqrt{\hat{s}(\boldsymbol{\epsilon})} + \hat{s}(\boldsymbol{\epsilon}) + \sigma^2\right)$$

due to the fact that $\|\bar{\mathbf{y}} - \bar{\mathbf{x}}\|^2 = O_{\mathbb{P}}(\sigma^2)$. Therefore we have the following corollary.

Corollary 42. *Consider the spectral scan estimator* $\hat{\mathbf{x}}$ *defined by* (4.10)*. Then* $\hat{\mathbf{x}}$ *is localization consistent (i.e.* $\tilde{d}(\mathbf{x}, \hat{\mathbf{x}}) \to 0$*) if*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\sum_{i=2}^{p} \min\{1, \frac{\rho}{\lambda_i}\}}\right)$$

and $\hat{\mathbf{x}}$ *defined in* (4.11) *is* ℓ_2 *consistent if in addition*

$$\sigma = o\left(\left(\sum_{i=2}^{p} \min\{1, \frac{\rho}{\lambda_i}\}\right)^{-\frac{1}{2}}\right)$$

Hence, we have that we can obtain an ℓ_2 consistent estimator by individually estimating $\mathbf{x}$ in the direction of $\hat{\mathbf{x}}_S$, the average, $\bar{\mathbf{x}}$, and the direction $\mathbf{z}$. While the GESS requires the entire spectrum the SSS and it's estimator may be solved by first order methods with linear solvers.

Remark 43. *Through a reparametrization we can show that the program* (4.5) *has a linear objective and only quadratic constraints. After forming the Lagrangian we can show that this is equivalent to*

$$\inf_{\nu_0,\nu_1 \geq 0} \nu_0\rho + \nu_1 + \frac{1}{4}\tilde{\mathbf{y}}^\top[\nu_0\Delta + \nu_1\mathbf{I}]^{-1}\tilde{\mathbf{y}}$$

which can be solved by first order interior point methods over the parameters ν_0, ν_1 *where the gradient calculation requires the solution to a linear system. Furthermore, the linear systems are semidefinite, diagonally dominant, hence by the recent work of [45], has a running time of* $O(|E| \log n)$ *modulo precision factors for the interior point algorithm.*

4.6 Specific Graph Models and Experiments

In this section we demonstrate the power and flexibility of Theorem 34 by analyzing in detail the performance of the GESS over three important graph topologies: balanced binary trees, the torus graph and Kronecker graphs (see [50, 51]). The explicit goals of this section are as follows:

1. Determine the implications of Theorem 34 in these specific graph examples;
2. Demonstrate the competitiveness of the GESS and the adaptive GESS against the agglomerative and max statistics;
3. Assess the tightness of our critical SNR bounds for the specific graph examples;
4. Provide an example of the general graph structure;
5. Provide a comparison to Ermakov's test.

4.6.1 Balanced Binary Trees

Balanced trees are graph structures of particular interest because they provide a hierarchical structure that is common in many social structures. Furthermore, the behavior of the graph spectra for the balanced binary tree provides a natural multiscale basis [71, 73]. We begin the analysis of the spectral scan statistic by applying it to the balanced binary tree (BBT) of depth ℓ. The class of signals that we will consider have clusters of constant signal which are subtrees of size at least cp^α for $0 < c \leq 1/2, 0 < \alpha \leq 1$. Hence, the cut size of the signals are 1 and $\rho = [cp^\alpha(1 - cp^{\alpha-1})]^{-1}$.

Corollary 44. *For the balanced binary tree with p vertices, the GESS can asymptotically distinguish H_0 from signals within H_1^{PC}, H_1^S, H_1^E where $\rho = p[cp^\alpha(p - cp^\alpha)]^{-1}$ if the SNR is stronger than*

$$\frac{\mu}{\sigma} = \omega(p^{\frac{1-\alpha}{4}}(\log p)^{1/4}).$$

(b) The adaptive GESS distinguishes the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega(p^{\frac{1-\alpha}{4}}(\log p)^{1/2}).$$

(c) H_0 and H_1^{PC} are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o(p^{\frac{1-\alpha}{4}}).$$

(d) The spectral scan estimator is localization consistent if

$$\frac{\mu}{\sigma} = o(p^{\frac{1-\alpha}{2}} \log p).$$

Proof. The study of the spectra of trees really began in earnest with the work of [21]. Notably, it became apparent that trees have eigenvalues with high multiplicities, particularly the eigenvalue 1. [60] gave a tight bound on the algebraic connectivity of balanced binary trees (BBT). They found that for a BBT of depth ℓ, the reciprocal of the smallest eigenvalue ($\lambda_2^{(\ell)}$) is

$$\begin{aligned} \frac{1}{\lambda_2^{(\ell)}} \leq 2^\ell - 2\ell + 2 - \frac{2^\ell - \sqrt{2}(2\ell - 1 - 2^{\ell-1})}{2^\ell - 1 - \sqrt{2}(2^{\ell-1} - 1)} + (3 - 2\sqrt{2}\cos(\frac{\pi}{2\ell - 1}))^{-1} \\ \leq 2^\ell + 105I\{\ell < 4\} \end{aligned} \tag{4.12}$$

[65] gave a more exact characterization of the spectrum of a balanced binary tree, providing a decomposition of the Laplacian's characteristic polynomial. Specifically, the characteristic polynomial of $\mathbf{L}$ is given by

$$\det(\lambda\mathbf{I} - \mathbf{L}) = p_1^{2^{\ell-2}}(\lambda)p_2^{2^{\ell-3}}(\lambda)...p_{\ell-3}^{2^2}(\lambda)p_{\ell-2}^2(\lambda)p_{\ell-1}(\lambda)s_\ell(\lambda) \tag{4.13}$$

where $s_\ell(\lambda)$ is a polynomial of degree ℓ and $p_i(\lambda)$ are polynomials of degree i with the smallest root satisfying the bound in (4.12) with ℓ replaced with i. In [66], they extended this work to more general balanced trees.

By (4.13) we know that at most $\ell + (\ell-1) + (\ell-2)2 + ... + (\ell-j)2^{j-1} \leq \ell 2^j$ eigenvalues have reciprocals larger than $2^{\ell-j} + 105I\{j < 4\}$. Let $k = \max\{\lceil \frac{\ell}{c} 2^{\ell(1-\alpha)} \rceil, 2^3\}$, then we have ensured that at most k eigenvalues are smaller than ρ. For n large enough

$$\sum_{i>1} \min\{1, \rho^2 \lambda_i^{-2}\} \leq k + \rho^2 \sum_{j > \log k}^{\ell} \ell 2^j 2^{2(\ell-j)} \leq k + \frac{\ell}{k} p^2 \rho^2 = O(p^{1-\alpha} \log p)$$

□

The conclusion is that for the BBT the GESS and the adaptive GESS is near optimal with respect to critical SNR. The proof is based on the special form of the spectrum of the BBT. It is reasonable to suspect that this strong result is due to the specific structure of the BBT, and will not hold for other graph structures. For the Torus and the Kronecker graph, there appears to be some gap between the performance of the GESS and the lower bound. But with this said, the GESS consistently dominates the naive estimators and the closeness of the theoretical results is surprisingly close to the lower bounds in all cases.

We simulate the probability of correct discovery of change-points (rejecting H_0 when the truth is H_1) versus the probability of false alarm (falsely rejecting H_0). These are given for the four estimators in Figure 4.1 as $p = 2^{\ell+1} - 1$ increases. Different estimators dominate under different choices of cluster size parameter, α. When $\alpha = 1$, corresponding to large clusters, where the size is on the same order as p, the aggregate statistic is competitive with the adaptive statistic. When $\alpha = 0.5$, corresponding to clusters of size $\approx p^{1/2}$, the aggregate becomes less competitive and the max more competitive, where the GESS remains the dominating test.

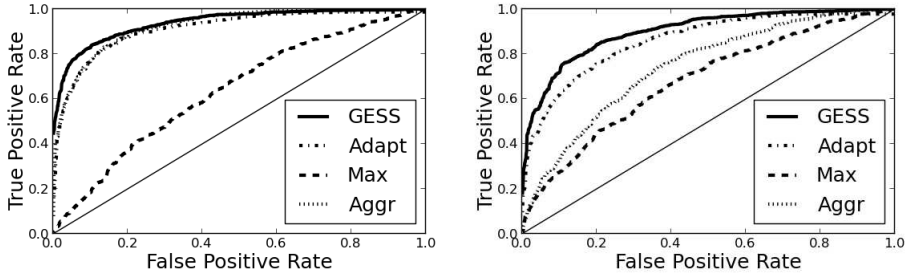


Figure 4.1: **(BBT Comparisons)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the balanced binary tree of the GESS, adaptive GESS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for the tree of depth $\ell = 6$, $p = 2^{\ell+1} - 1 = 127$, with choice of $\alpha = 1$ (left) and $\alpha = 0.5$ (right).

One may rightly ask if the gap between the bound in Cor. 44 is just due to a lack of theoretical know-how and the estimator is truly optimal. We attempt to assuage these fears by plotting the performance of the GESS with constant SNR as the graph size increases alongside that with the SNR increasing according to the scaling dictated by Cor. 44 (Figure 4.2). As we can see for both $\alpha = 1, 0.5$ when the SNR is rescaled the ROC curves line up, and the performance does not noticeably increase.

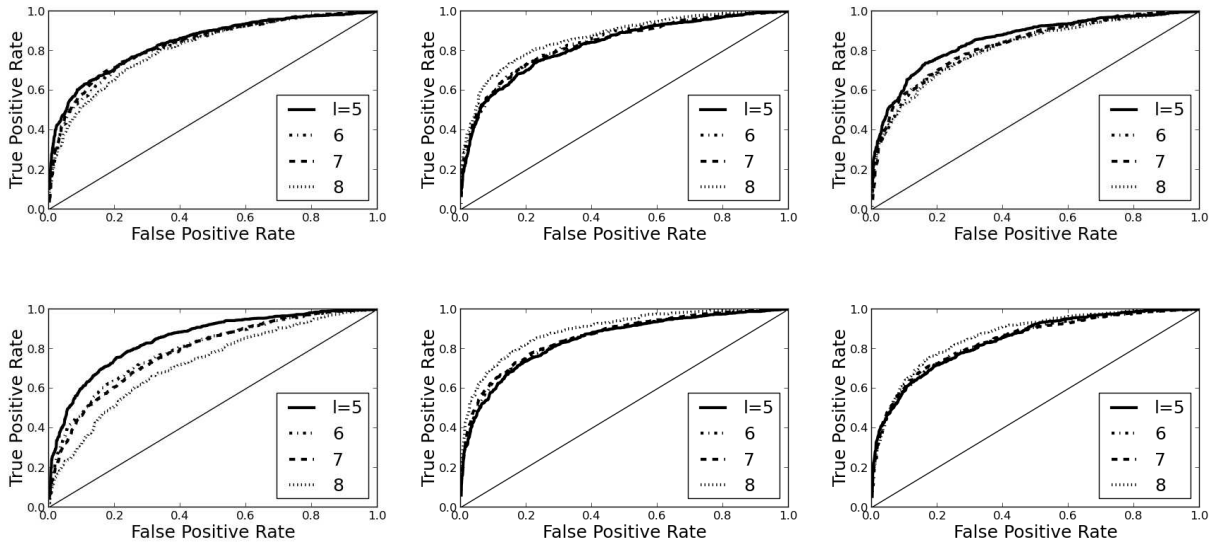


Figure 4.2: **(BBT Rescaling)** The size (false positive rate) and power (true positive rate) of the GESS as p increases from $\ell = 5, p = 2^6 - 1 = 63$ to $\ell = 8, p = 2^9 - 1 = 511$, with $\alpha = 1$ (top) and $\alpha = .5$ (bottom). The SNR was held fixed (left), then was allowed to scale according to Cor. 44 (b) (middle), and scaled according to the lower bound Cor. 44 (c) (right).

4.6.2 Torus Graph

The torus has been a pedagogical example, but it is also an important example as it models a mesh of sensors in two dimensions. We will analyze the performance guarantees of the GESS over our running example, the 2-dimensional torus graph with ℓ vertices along each dimension ($p = \ell^2$). To include squares of size $p^{1-\beta}$, as in the examples, then we would obtain $\rho \asymp p^{-(1-\beta)/2}$. The following result is due to a detailed analysis of the spectrum of the torus.

Corollary 45. *Let G be the $\ell \times \ell$ square lattice ($p = \ell^2$), and let $\rho = Cp^{-(1-\beta)/2}$ for $\beta \in [0, 1)$.*
(a) The GESS can asymptotically distinguish H_0 from H_1^{PC}, H_1^S, H_1^E if the SNR satisfies

$$\frac{\mu}{\sigma} = \omega(p^{\frac{1+\beta}{8}})$$

(b) The adaptive GESS can asymptotically distinguish the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega(p^{\frac{1+\beta}{8}}(\log p)^{1/4})$$

(c) H_0 and H_1^{PC} are asymptotically indistinguishable if the SNR is weaker than

$$\frac{\mu}{\sigma} = o(p^{\frac{\beta}{4}})$$

(d) The spectral scan estimator is localization consistent (i.e. $\tilde{d}(\hat{\mathbf{x}}, \mathbf{x}) \to 0$) if

$$\frac{\mu}{\sigma} = \omega(p^{\frac{1+\beta}{4}}\sqrt{\log p})$$

Proof. By a simple Fourier analysis (see [71]), we know that the Laplacian eigenvalues are $2(2-\cos(2\pi i_1/\ell)-\cos(2\pi i_2/\ell))$ for all $i_1, i_2 \in [\ell]$. Let us denote the ℓ^2 eigenvalues as $\lambda_{(i_1,i_2)}$ for $i_1, i_2 \in [\ell]$. Notice that for $i \in [\ell]$, $|\{(i_1,i_2) : i_1 \vee i_2 = i\}| \leq 2i$. For simplicity let ℓ be even. We know that if $i_1 \vee i_2 \leq \ell/2$ then $\lambda_{(i_1,i_2)} = 2-\cos(2\pi i_1/\ell)-\cos(2\pi i_2/\ell) \geq 1-\cos(2\pi(i_1 \vee i_2)/\ell)$. Let $k \ll \ell$ which we will specify later. Thus,

$$\begin{aligned}
&\sum_{(i_1,i_2)\neq(1,1)\in[\ell]^2} 1 \wedge \frac{\rho^2}{\lambda^2_{(i_1,i_2)}} \\
&\leq 2\sum_{i\in[\ell/2]} 2i\left(1 \wedge \frac{\rho^2}{(1-\cos(2\pi i/\ell))^2}\right) \\
&\leq 4\sum_{i=1}^{k} i + \rho^2\frac{\ell^2}{2}\frac{2}{\ell}\sum_{k<i\leq\ell/2} 2\frac{i/\ell}{(1-\cos(2\pi i/\ell))^2} \\
&\leq 4k^2 + \rho^2\frac{\ell^2}{2}\int_{k/\ell}^{1/2}\frac{x\,dx}{1-\cos(2\pi x)} \\
&= 4k^2 + \rho^2\frac{\ell^2}{2}\left(\frac{\ell^2}{k^2} + o\left(\frac{\ell}{k}\right)\right)
\end{aligned}$$

The above followed by the Taylor expansion about 0 of the integral. Let us choose k to be the point at which $1-\cos(2\pi k) \approx \rho$. By the first order approximation of $\arccos$ this occurs around $k = \lfloor\sqrt{\rho}\ell\rfloor$. The inequalities above hold regardless of the choice of k, as long $k \ll \ell$, so we have the freedom to tune it to our liking. Plugging this in we obtain,

$$\sum_{(i_1,i_2)\neq(1,1)\in[\ell]^2} 1 \wedge \frac{\rho^2}{\lambda^2_{(i_1,i_2)}} = O(\rho\ell^2) = O(n^{(1+\beta)/2})$$

□

The implication of Cor. 45 is that when β is close to 1 (small clusters), the GESS has nearly optimal critical SNR. When β is small (large clusters), we suffer an additional factor of $p^{(1-\beta)/8}$ in the upper bound. We conclude from the simulations (Figure 4.3) and our theory that the GESS dominates when $\beta \approx 1/2$, so clusters are of intermediate size.

We again see by rescaling the SNR (see Figure 4.4) according to the bound in Cor. 45 that the theory appears to be correct. (After the rescaling the SNR the ROC curves uniformly do not decrease as n increases.) While the curves do not exactly line up after rescaling for $\beta = .5$, this seems to be a low n effect as the ROC curves for $\ell = 20, 30, 40$ match.

4.6.3 Kronecker Graphs

Much of the research in complex networks has focused on observing statistical phenomena that is common across many data sources. The most notable of these are that the degree distribution obeys a power law ([19]) and networks are often found to have small diameter ([59]). A class of graphs that satisfy these, while providing a simple modelling platform are the Kronecker graphs

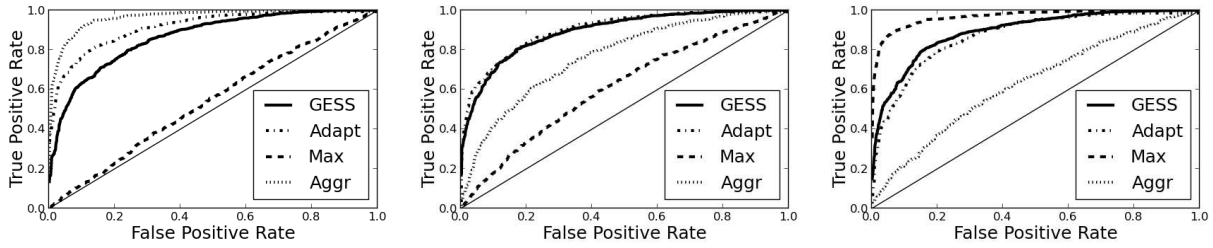


Figure 4.3: **(Torus Comparisons)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the Torus of the GESS, adaptive GESS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for the size length of $\ell = 30$, $p = \ell^2 = 900$, with choice of $\beta = 0$ (left), $\beta = .5$ (middle) and $\beta = .75$ (right).

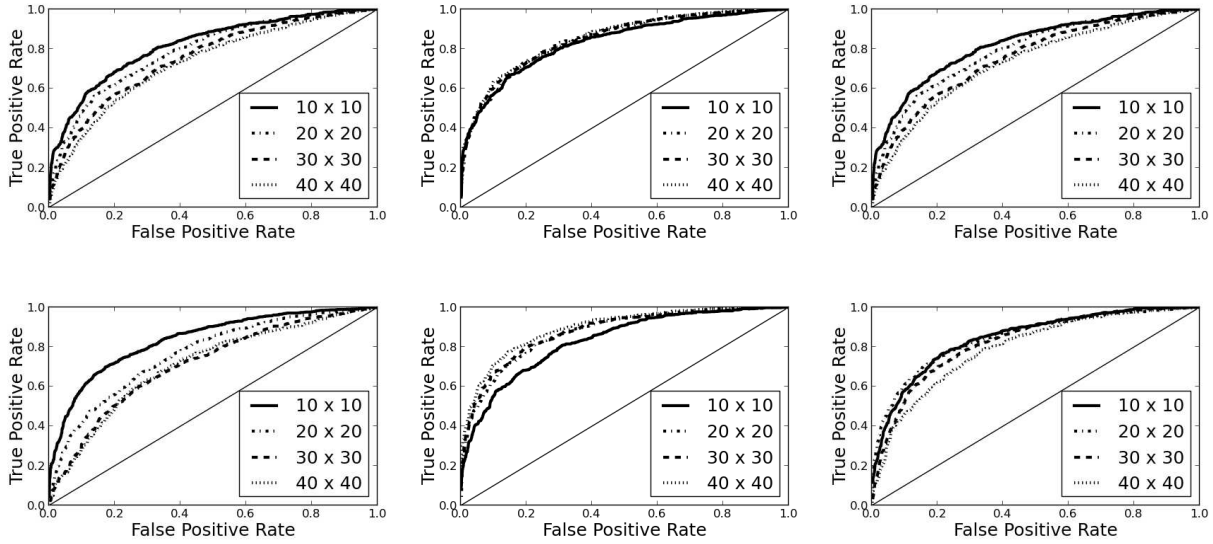


Figure 4.4: **(Torus Rescaling)** The size (false positive rate) and power (true positive rate) of the GESS as n increases from $\ell = 10, n = 100$ to $\ell = 40, n = 1200$, with $\beta = 0$ (top) and $\beta = .5$ (bottom). The SNR was held fixed (left), then was allowed to scale according to Cor. 45 (b) (middle), and scaled according to the lower bound Cor. 45 (c) (right).

(see [50, 51]). Let H_1 and H_2 be graphs on p_0 vertices with Laplacians Δ_1, Δ_2 and edge sets E_1, E_2 respectively. The Kronecker product, $H_1 \otimes H_2$, is the graph over vertices $[p_0] \times [p_0]$ such that there is an edge $((i_1, i_2), (j_1, j_2))$ if $i_1 = j_1$ and $(i_2, j_2) \in E_2$ or $i_2 = j_2$ and $(i_1, j_1) \in E_1$. We will construct graphs that have a multi-scale topology using the Kronecker product. Let the multiplication of a graph by a scalar indicate that we multiply each edge weight by that scalar. First let H be a connected graph with p_0 vertices. Then the graph G for $\ell > 0$ levels is defined as

$$\frac{1}{p_0^{\ell-1}} H \otimes \frac{1}{p_0^{\ell-2}} H \otimes ... \otimes \frac{1}{p_0} H \otimes H$$

The choice of multipliers ensures that it is easier to make cuts at the more coarse scale. Notice that all of the previous results have held for weighted graphs.

Corollary 46. *For G be the Kronecker product of the base graph H described above with $p = p_0^\ell$ vertices, the GESS can asymptotically distinguish H_0 from signals from H_1^{PC}, H_1^S, H_1^E with $\rho \propto p_0^{2k-\ell-1}$ (which includes cuts within the k coarsest scale), if the SNR is stronger than,*

$$\frac{\mu}{\sigma} = \omega(p^{k/2\ell}(\textit{diam}(H))^{1/4})$$

where diam(H) is the diameter of the base graph H.
(b) The adaptive GESS can distinguish the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega(p^{k/2\ell}(\textit{diam}(H) \log p)^{1/4})$$

(c) H_0 and H_1^{PC} are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o(p^{k/4\ell})$$

(d) The spectral scan estimator is localization consistent (i.e. $\tilde{d}(\hat{\mathbf{x}}, \mathbf{x}) \to 0$)

$$\frac{\mu}{\sigma} = \omega(p_0^2(\ell+1)p^{(2k+1)/\ell})$$

Proof. The Kronecker product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$ is defined as $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{(p\times p)\times(p\times p)}$ such that $(\mathbf{A} \otimes \mathbf{B})_{(i_1,i_2),(j_1,j_2)} = A_{i_1,j_1} B_{i_2,j_2}$. Some matrix algebra shows that if H_1 and H_2 are graphs on p_0 vertices with Laplacians Δ_1, Δ_2 then the Laplacian of their Kronecker product, $H_1 \otimes H_2$, is given by $\Delta = \Delta_1 \otimes \mathbf{I}_{p_0} + \mathbf{I}_{p_0} \otimes \Delta_2$ ([58]). Hence, if $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{p_0}$ are eigenvectors, viz. $\Delta_1 \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$ and $\Delta_2 \mathbf{v}_2 = \lambda_2 \mathbf{v}_2$, then $\Delta(\mathbf{v}_1 \otimes \mathbf{v}_2) = (\lambda_1 + \lambda_2)\mathbf{v}_1 \otimes \mathbf{v}_2$, where $\mathbf{v}_1 \otimes \mathbf{v}_2$ is the usual tensor product. This completely characterizes the spectrum of Kronecker products of graphs.

We should argue the choice of $\rho \propto p_0^{2k-\ell-1}$, by showing that it is the results of cuts at level k. We say that an edge $e = ((i_1, ..., i_\ell), (j_1, ..., j_\ell))$ has scale k if $i_k \neq j_k$. Furthermore, a cut has scale k if each of its constituent edges has scale at least k. Each edge at scale k has weight $p_0^{k-\ell}$ and there are $p_0^{\ell-1}$ such edges, so cuts at scale k have total edge weight bounded by

$$p_0^{\ell-1} \sum_{i=1}^{k} p_0^{i-\ell} = p_0^{k-1} \frac{p_0 - \frac{1}{p_0^{k-1}}}{p_0 - 1} \leq \frac{p_0^k}{p_0 - 1}$$

Cuts at scale k leave components of size $p_0^{\ell-k}$ intact, meaning that $\rho \propto p_0^{2k-\ell-1}$ for large enough p_0.

We now control the spectrum of the Kronecker graph. Let the eigenvalues of the base graph H be $\{\nu_j\}_{j=1}^{p_0}$ in increasing order. The eigenvalues of G are precisely the sums

$$\lambda_i = \frac{1}{p_0^{\ell-1}}\nu_{i_1} + \frac{1}{p_0^{\ell-2}}\nu_{i_2} + ... + \frac{1}{p_0}\nu_{i_{\ell-1}} + \nu_{i_\ell}$$

for $i = (i_j)_{j=1}^{\ell} \subseteq [p_0]$. The eigenvalue distribution $\{\lambda_i\}$ stochastically bounds

$$\lambda_i \geq \sum_{j=1}^{\ell} \frac{1}{p_0^{\ell-j}}\nu_2 I\{\nu_{i_j} \neq 0\} \geq \frac{\nu_2}{p_0^{Z(i)}}$$

where $Z(i) = \min\{j : \nu_{i_{\ell-j}} \neq 0\}$. Notice that if i is chosen uniformly at random then $Z(i)$ has a geometric distribution with probability of success $(p_0 - 1)/p_0$. Hence,

$$\begin{aligned}
&\frac{1}{p_0^{\ell}} \sum_{i \in [p_0]^\ell} \min\{1, \frac{\rho^2}{\lambda_i^2}\} \leq \mathbb{E}_Z \min\{1, \frac{\rho^2 p_0^{2Z}}{\nu_2^2}\} \\
&\leq \mathbb{P}_Z\{Z \geq 2k - \ell - 1 + \log_{p_0} \nu_2\} + \frac{1}{\nu_2^2} \sum_{z=1}^{\lfloor \ell+1-2k+\log_{p_0} \nu_2 \rfloor} {p_0}^{2(2k-\ell-1+z)} \mathbb{P}_Z\{Z = z\} \\
&\leq {p_0}^{2k-\ell-1+\log_{p_0} \nu_2} + \frac{1}{\nu_2^2} \sum_{z=1}^{\lfloor \ell+1-2k+\log_{p_0} \nu_2 \rfloor} {p_0}^{2(z+2k-\ell-1)} \frac{1}{{p_0}^z} \frac{p_0 - 1}{p_0} \\
&= O((\nu_2 + \nu_2^{-1}){p_0}^{2k-\ell-1}) = O({p_0}^{2k-l}\text{diam}(H))
\end{aligned}$$

where $\text{diam}(H)$ is the diameter of the base graph H. Hence,

$$\sum_{i \in [p_0]^\ell} \min\{1, \frac{\rho^2}{\lambda_i^2}\} = O(n^{2k/\ell}\text{diam}(H))$$

□

The proof and an explanation of ρ is in the appendix. The implication of Cor. 46 is that only for k small is the GESS nearly optimal. Generally, one will suffer an additional term of $p^{k/4\ell}$. As we can see from the simulations the $k = 1$ case is exactly when the aggregate statistic dominates the ROC curve (see Figure 4.5). It is in the k larger than 1 and less than ℓ case that the GESS improves on the aggregate and the max statistics.

We see by rescaling the SNR (see Figure 4.4) according to the bound in Cor. 45 that the theory appears to be correct if loose. (After the rescaling the SNR the ROC curves uniformly do not decrease as p increases.) For $k = 2$ the rescaling seems to be overkill, if lining up the ROC curves is desired. This indicates that the gap between the upper and lower bounds may be artificial.

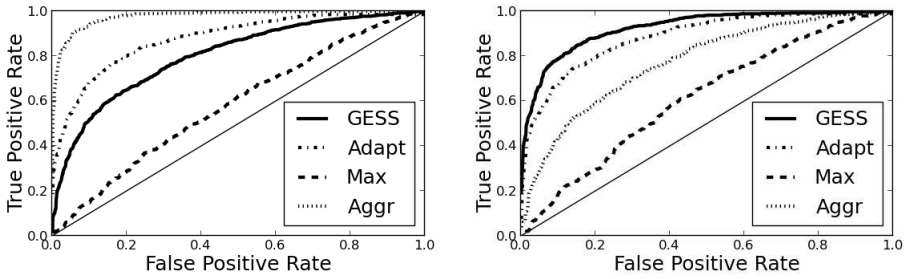


Figure 4.5: **(Kronecker Comparison)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the Kronecker graph of the GESS, adaptive GESS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for a base graph of size $p_0 = 6$ and Kronecker power of $\ell = 3$, so $p = p_0^\ell = 216$. The cuts were chosen at the coarsest scale, $k = 1$, (left) and at the second coarsest, $k = 2$ (right).

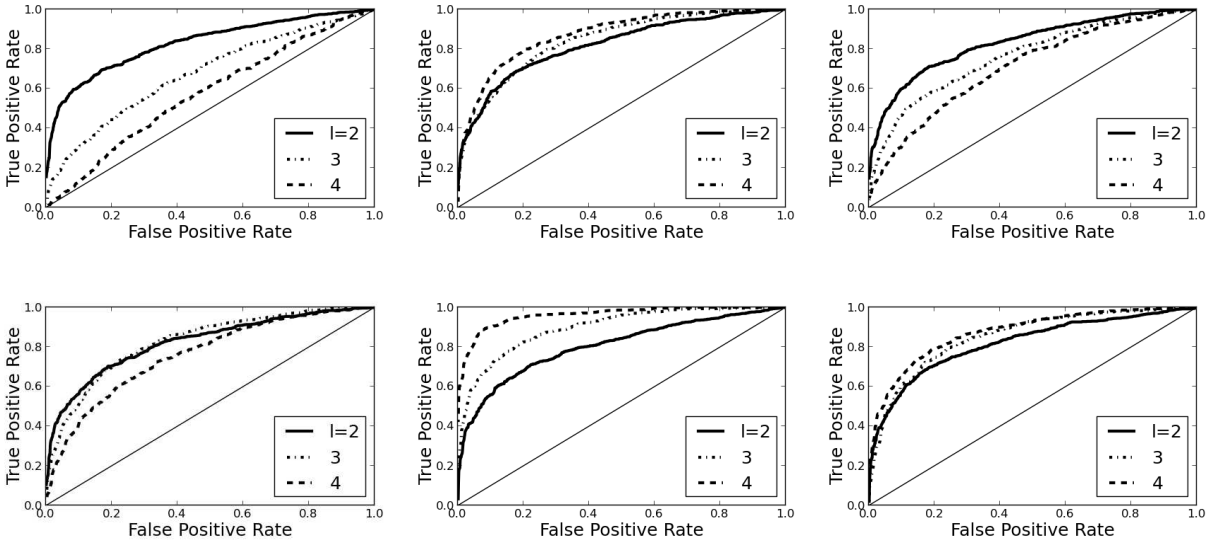


Figure 4.6: **(Kronecker Rescaling)** The size (false positive rate) and power (true positive rate) of the GESS on the Kronecker graph with base graph size $p_0 = 6$; p increases from $\ell = 2, p = 6^2 = 64$ to $\ell = 4, p = 216$, with $k = 1$ (top) and $k = 2$ (bottom). The SNR was held fixed (left), then was allowed to scale according to Cor. 46 (b) (middle), and scaled according to the lower bound Cor. 46 (c) (right).

4.6.4 General Graph Structure, H_1^S

The piecewise constant alternative hypothesis H_1^{PC} is amenable to a sophisticated theoretical analysis and it motivates the GESS. Unfortunately, it is very easy to modify signals in $\mathcal{X}_{PC}(\mu, \rho)$ by slight perturbations and find a signal that is outside our supposed class. This lack of robustness is rightly alarming, and it is through the general graph structured class, $\mathcal{X}_S(\mu, \rho)$, that we intended to include these perturbations. We know will show that the alternative H_1^S achieves this claim, through an example, and will demonstrate the performance of the GESS on the example.

Suppose that we begin with a signal $\mathbf{x} \in \mathcal{X}_{PC}(\mu, \rho)$ such that $\mathbf{x} = \delta \mathbf{1}_C$ and modify it in the following way: let $C' \subset C$ and make $\mathbf{x}' \propto \mathbf{1}'_C$ such that $\mathbf{x}' \in \mathcal{X}_S(\mu, \rho)$. We now determine the normalization that would make this so. Notice that

$$\left| \frac{\mathbf{x}'^\top \mathbf{1}_C}{|C|} - \frac{\mathbf{x}'^\top \mathbf{1}_{\bar{C}}}{|\bar{C}|} \right| \sqrt{\frac{|C||\bar{C}|}{p}} \geq \mu$$

Hence, $\mathbf{x}' = \delta' \mathbf{1}_{C'}$ implies that

$$\delta' = \frac{|C|}{|C'|} \delta$$

So for the subsampled signal $\mathbf{x}'$ to remain in $\mathcal{X}_S(\mu, \rho)$ we will need to boost the signal by a factor of $|C|/|C'|$. Figures 4.7, 4.8, 4.9 show the ROC curves for the GESS when the signal cluster C' is formed by including each vertex in C according to independent Bernoulli(q) random variables. To make the comparisons fair we boosted the signals according to the above formulation. As one can see in all the cases the subsampling does not make any ROC curves worse. In the case of the torus (Figure 4.9), subsampling the cluster seems to help the performance. It is guessed that because in this case when the scan is performed not only does the cluster match, but individual node clusters begin to match the data.

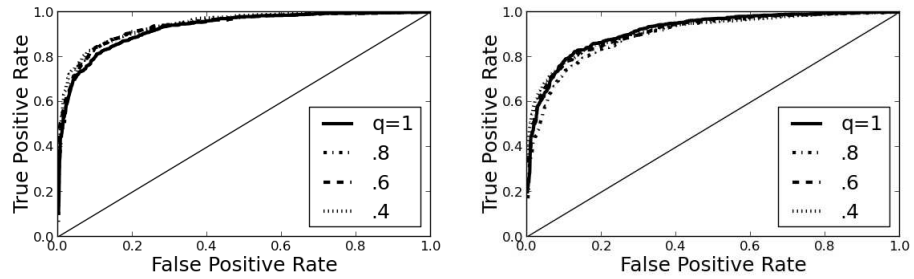


Figure 4.7: **(BBT Perturbations)** Simulations of the size (false positive rate) and the power under H_1^S for the balanced binary tree of the GESS with changing cluster sampling probability, q. The figures are for the tree of depth $\ell = 6$, $p = 2^{\ell+1} - 1 = 127$, with choice of $\alpha = 1$ (left) and $\alpha = 0.5$ (right).

4.6.5 Ermakov Comparison

This study would certainly be incomplete without a comparison to Ermakov's test for the ellipsoidal alternative hypothesis. We have shown that if $\mathbf{x} \in \mathcal{X}_{PC}(\mu, \rho)$ then if we define $a_i =$

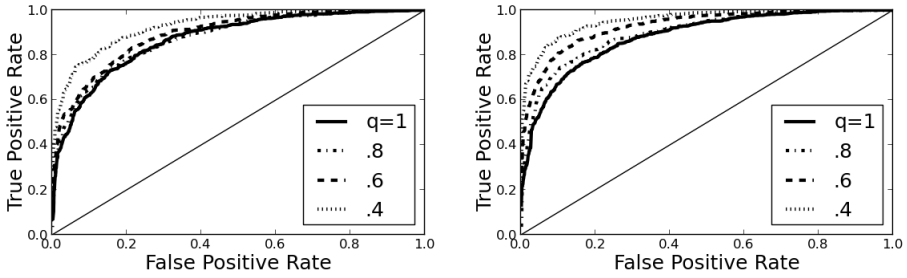


Figure 4.8: **(Torus Perturbations)** Simulations of the size (false positive rate) and the power under H_1^S for the Torus of the GESS with changing cluster sampling probability, q. The figures are for the size length of $\ell = 30$, $p = \ell^2 = 900$, with choice of $\beta = 0$ (left) and $\beta = .75$ (right).

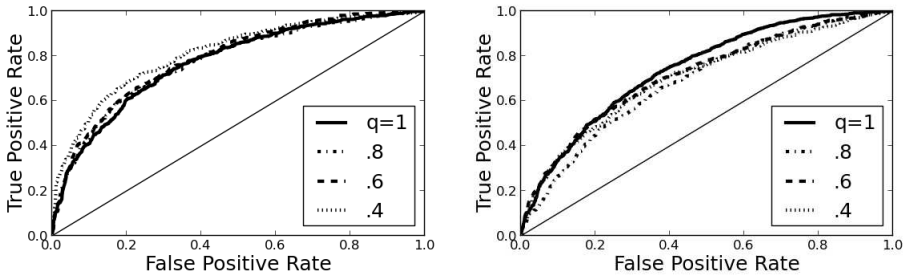


Figure 4.9: **(Kronecker Perturbations)** Simulations of the size (false positive rate) and the power under H_1^S for the Kronecker graph of the GESS with changing cluster sampling probability, q. The figures are for a base graph of size $p_0 = 6$ and Kronecker power of $\ell = 3$, so $p = p_0^\ell = 216$. The cuts were chosen at the coarsest scale, $k = 1$, (left) and at the second coarsest, $k = 2$ (right).

$\min\{1, \rho/\lambda_i\}$ we know that $\sum_i a_i(\mathbf{U}\mathbf{x})_i^2 \leq 2$. Hence, we are partially justified in using Ermakov's statistic $\hat{e}$ where the axis length parameters are $\{a_i/2\}$. The results of simulations are in Figure 4.10. We see that only for the Torus does Ermakov's test have a slightly better ROC curve. For the Kronecker graph, the GESS dominates Ermakov's test, and for the BBT they are all competitive.

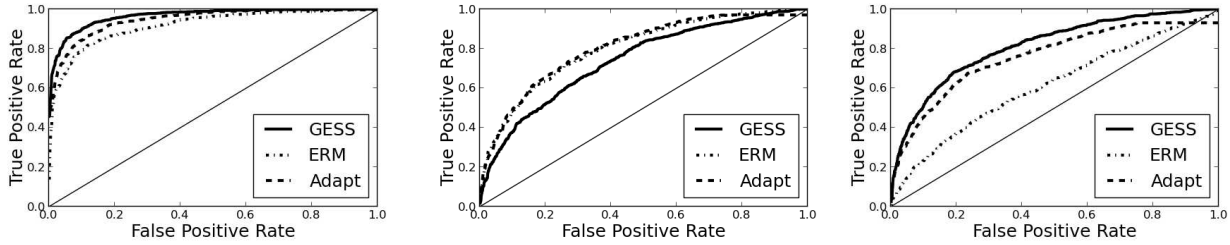


Figure 4.10: **(Ermakov's test)** Simulations of the size (false positive rate) and the power (false negative rate) under H_1^{PC} for the Torus of the GESS, Ermakov's test (ERM), and the adaptive GESS (Adapt). The figures are for the tree of depth $\ell = 6$, $p = 2^{\ell+1} - 1 = 127$, with choice of $\alpha = 1$ (left); the torus of side length, $\ell = 30$, $p = \ell^2 = 900$, $\beta = 0$ (middle); Kronecker graph with base graph of size $p_0 = 6$ and Kronecker power of $\ell = 3$, so $p = p_0^\ell = 216$.

4.7 Discussion

In this Chapter, we studied balanced graph structured activation detection, in which we require that activations not only have low graph cut, but are large relative to the graph cut. We observed that the GLR statistic was a computationally intractable combinatorial optimization. This led naturally to the proposal of relaxations of the GLR, specifically the SSS and the GESS. The GESS provided us with superior theoretical risk bounds for detection, which can be stated in terms of the graph spectrum. The SSS provided a natural estimator of the true mean, $\mathbf{x}$, in this setting. We showed that these performed nearly optimally on the balanced binary tree.

While perhaps our upper bound in the Kronecker graph is not as small as it could be, we have shown ample evidence that this is not the case for the Torus. It is natural to ask, why does the GESS perform suboptimally in the Torus graph? It is believed that the relaxation is not constrictive enough, and that perhaps there are relaxations with a smaller integrality gap that we can exploit. Answering this question fully was not within the scope of this thesis, and it remains an open problem.

Chapter 5

Spanning Tree Multiscale Basis for Estimation, Localization and Detection

In this chapter, we will be testing if there is a non-zero piece-wise constant activation pattern on the graph given observations that are corrupted by Gaussian white noise. We show that correctly distinguishing the null and alternative hypotheses is impossible if the signal-to-noise ratio does not grow quickly with respect to the allowable number of discontinuities in the activation pattern. With that said, we are able to out perform naive detectors that do not exploit the graph structure. We summarize our results in this chapter:

1. **Univeral Lower Bound and Unstructured Tests:** We prove a lower bound for the ℓ_0-structured detection problem, which will serve as a benchmark for all tests. We also describe the SNR required for the max and aggregate tests to asymptotically distinguish H_0 from H_1.
2. **Detection with Spanning Tree Decompositions:** We develop a decomposition of the graph based on a spanning tree of the graph. This naturally leads to a test which admits theoretical risk bound.
3. **Localizing Activity with Spanning Tree Wavelets:** We give a computationally tractable algorithm for the construction of a wavelet basis over the graph from a spanning tree. We show that thresholding in the wavelet basis admits a localization guarantee.
4. **Uniform Spanning Tree Basis:** We prove that if one constructs the spanning tree randomly, then we can express the aforementioned results in terms of electric network properties of the graph.
5. **Specific Graph Models:** We find that the UST wavelet estimator and the UST decomposition detector provides near optimal performance on the torus graph, the kNN graph, and the $\epsilon-$graph.

First, we will remind the reader of the details for the testing problem. The structure of activation pattern $\mathbf{x}$ is determined by the graph G. Specifically, we assume that there are parameters ρ, μ (possibly dependent on p[1]) such that the class of graph-structured activation patterns $\mathbf{x}$ is

[1]We suppress dependence on the number of edges m as we focus on graph models where m depends on n.

given as follows.

$$\mathcal{X}_0 = \left\{ \frac{\mu}{\sqrt{|C|}} \mathbf{1}_C : C \subset V, |\partial C| \leq \rho \right\}$$

where the boundary is defined as $\partial C = \{(v,w) \in E : v \in C, w \notin C\}$ and the indicator $\mathbf{1}_{C,v} = \mathbf{1}\{v \in C\}$. Hence, the possible patterns have few edges across which the values of $\mathbf{x}$ differ. In other words, the set of activated vertices C have a small *cut size* in the graph G, where cut size is defined to be $\sum_{v \in C} \sum_{w \notin C} \mathbf{1}[(v,w) \in E]$. If the number of activated vertices is $|C|$, then values of $\rho < |C| \cdot \text{degree}(G)$ [2] imply that the activation is localized on the graph. In ℓ_0-structured activation detection we are concerned with statistically testing the null and alternative hypotheses,

$$\begin{aligned} H_0 &: \mathbf{y} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \\ H_1 &: \mathbf{y} \sim N(\mathbf{x}, \sigma^2 \mathbf{I}), \mathbf{x} \in \mathcal{X}_0, \|\mathbf{x}\| \geq \mu \end{aligned} \tag{5.1}$$

H_0 represents business as usual while H_1 encompasses all of the foreseeable anomalous activity.

5.1 Universal Lower Bound and Unstructured Tests

In order to more completely understand the problem of detecting weak activations on graphs, we prove that there is a universal minimum signal strength under which H_0 and H_1 are asymptotically indistinguishable. The proof is based on a result developed in [4], with a new construction of prior distribution over worst case patterns.

Theorem 47. *Hypotheses H_0 and H_1 defined in Eq. (5.1) are asymptotically indistinguishable if*

$$\frac{\mu}{\sigma} = o\left(\sqrt{\min\left\{ \frac{\rho}{d_{\max}} \log\left(\frac{p d_{\max}^2}{\rho^2} \right), \sqrt{p} \right\}} \right)$$

where $d_{\max}$ is the maximum degree of graph G.

Proof. We begin by constructing a prior distribution over $\mathcal{X}$. Consider $k = \min\{\rho/d_{\max}, \sqrt{p}\}$ and form a subset $\mathcal{S} \subseteq 2^{[p]}$ that consists of all subsets of k vertices, i.e. $|S| = k$ for all $S \in \mathcal{S}$. For every $S \in \mathcal{S}$ construct a pattern $\mathbf{x} = \frac{\mu}{\sqrt{|S|}} \mathbf{1}_S$ and denote the collection of such patterns as $\mathcal{X}'$. Note that for all $\mathbf{x} \in \mathcal{X}'$, $\|\mathbf{x}\|_2 = \mu$ and by construction $|\{(v,w) \in E : x_v \neq x_w\}| \leq k d_{\max} \leq \rho$. Hence, $\mathcal{X}' \subseteq \mathcal{X}$. The prior we construct assigns uniform probability to all patterns in $\mathcal{X}'$ and zero probability to patterns in $\mathcal{X} \setminus \mathcal{X}'$. This gives us a prior distribution π over $\mathcal{X}$.

The Bayes risk associated with a prior π is defined as $R^* = \inf_T R^*(T) = \inf_T \{\mathbb{E}_{\mathbf{0}}[T] + \mathbb{E}_\pi \mathbb{E}_{\mathbf{y}}[1 - T]\}$. Notice that $R(T) \geq R^*(T) \geq R^*$ for all prior distributions π. Our prior satisfies the conditions of Proposition 3.4 in [1], and hence we get that $R^* \geq \delta$ for all μ such that

$$\frac{\mu}{\sigma\sqrt{k}} \leq \sqrt{\log\left(1 + \frac{p \log(1 + 4(1-\delta)^2)}{k^2} \right)}$$

[2] degree(G) is the maximum degree of any $v \in G$.

from which the theorem follows by considering the two cases $\rho/d_{\max} \leq \sqrt{p}$ and $\rho/d_{\max} > \sqrt{p}$. □

We illustrate that there is a gap between this lower bound and tests that do not incorporate the graph structure. We analyze two test statistics: vertex-wise thresholding and the vertex averaging detector. Vertex-wise thresholding is typically used when it's believed that the parameter $\mathbf{x}$ is sparse, in the sense that it has few non-zero coordinates, or the signal strength at each vertex is high. The following proposition characterizes its performance.

Proposition 48. *Consider the vertex-wise thresholding test statistic* $\max_{v\in V}|y_v|$*, then a necessary and sufficient condition for asymptotic distinguishability of* H_0 *and* H_1 *is*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{|C|}(\sqrt{\log n} - \sqrt{\log |C|})\right)$$

where $|C| = o(p)$ *denotes the number of activated vertices.*

Proof. By Cirelson's theorem, we know that

$$\mathbb{P}\{\max_{v\in C}\epsilon_v \geq \mathbb{E}\max_{v\in C}\epsilon_v + u\} \leq e^{-u^2/2\sigma^2}$$

and we have that $\mathbb{E}\min_{v\in C}\epsilon_v$ is within a constant factor of $-\sigma\sqrt{2\log|C|}$ by the Majorizing-Measure theorem. Hence, under H_1^C we have that $\max_{v\in C} y_v \geq \mu/\sqrt{|C|} + L\sigma\sqrt{2\log|C|} - \sigma\sqrt{2\log(1/\delta)}$ with probability at least $1-\delta$ for some universal constant L. By similar reasoning, under H_0 we have that $\max_{v\in V}|y_v| \leq L\sigma\sqrt{2\log p} + \sigma\sqrt{2\log(2/\delta)}$. We have analogous opposite bounds that prove the necessary condition. Hence, a threshold, τ, distinguishes H_0 and H_1 if and only if

$$\begin{aligned}\mu/\sqrt{|C|} + L\sigma\sqrt{2\log|C|} - \sigma\sqrt{2\log 1/\delta} &> \tau \\ &> L\sigma\sqrt{2\log p} + \sigma\sqrt{2\log 1/\delta}\end{aligned}$$

□

When $\rho/d_{\max} = o(|C|)$ (i.e. C is structured or localized on the graph) and $|C|$ is increasing with p, then there can be a significant gap between the lower bound of Theorem 47 and the upper bound of Proposition 48. In other words, vertex thresholding does not take advantage of the pattern structure when it is localized on the graph. On the other hand, if the signal is unstructured or very sparse (number of activations do not increase with size of the graph), then the max statistic is nearly optimal (up to log factors).

Another natural test statistic is vertex averaging, $|\frac{1}{p}\sum_{v\in V} y_v|$, whose performance is characterized below.

Proposition 49. *Consider the test statistic* $|\frac{1}{n}\sum_{v\in V} y_v|$*, then a necessary and sufficient condition for asymptotic distinguishability of* H_0 *and* H_1 *is*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{p}{|C|}}\right)$$

where $|C|$ *denotes the number of activated vertices.*

Proof. Under H_0, $\frac{1}{p}\sum_{v\in V} y_v$ is normally distributed with mean 0 and variance σ^2/p. Meanwhile under H_1^C, $\frac{1}{p}\sum_{v\in V} y_v$ is normally distributed with mean $\mu\sqrt{|C|}/p$ and variance σ^2/p. Hence, the test statistic asymptotically distinguished H_0 from H_1 if and only if

$$\Phi\left(-\frac{\mu}{\sigma}\sqrt{\frac{|C|}{p}}\right) = o(1)$$

where Φ is the CDF of the standard normal. □

Since $\rho \leq |C| d_{\max}$, if $|C| = o(\sqrt{p/\log p})$ then there can be a significant gap between the lower bound of Theorem 47 and the upper bound of Proposition 49. Intuitively, if the number of activated vertices is small, then globally averaging the observations at all vertices is suboptimal. We will show that testing with spanning tree wavelets is near optimal (up to log factors) and hence bridges the gap in performance exhibited by unstructured tests (both when the structure is localized and when the number of activated vertices is small).

5.2 Detection with Spanning Tree Decompositions

In this section, we show that given a spanning tree $\mathcal{T}$ one can decompose the graph by splitting $\mathcal{T}$ in a strategic manner. Then we will see how using this decomposition, one can form a test statistic for testing H_0 versus H_1. In the following section we will use this decomposition to show that one can localize the cluster C by forming a wavelet basis from the decomposition.

5.2.1 The Decomposition Construction

The graph decomposition, also known as a laminar decomposition, is a set of clusters $\mathcal{D} = \{D_i\}_{i=1}^{|\mathcal{D}|} \subset 2^{[p]}$ such that the following hold: for any two $D \neq D' \in \mathcal{D}$ either $D \subset D'$, $D' \subset D$, or $D \cap D' = \emptyset$; $\{v\} \in \mathcal{D}, \forall v \in [p]$; and $[p] \in \mathcal{D}$. Thus we can associate a dendrogram, given by the partial ordering $D \subset D'$, to the decomposition $\mathcal{D}$. For each $D \in \mathcal{D}$, let level$(D) = |\{D' \in \mathcal{D} : D \subset D'\}|$. We will furthermore denote height$(\mathcal{D}) = \max_{v\in[p]}$ level$(\{v\})$.

We construct our decomposition $\mathcal{D}$ recursively from $\mathcal{T}$, by first finding a seed vertex in the spanning tree such that the subtrees adjacent to the seed have at most $\lceil p/2 \rceil$ vertices and then by including these subtrees to the decomposition. We recurse on each subtree, each time finding a balancing vertex and splitting around said vertex. There are two properties of this decomposition that we would like: that it has logarithmic height and it preserves the connectivity of the graph. This will be made precise in Proposition 50.

The balancing step ensures (1), while the splitting method ensures (2). Finding a balancing vertex in the tree parallels the technique in [61], which finds a balancing edge. The algorithm starts from any vertex $v \in \mathcal{T}$ and moves along $\mathcal{T}$ to a neighboring vertex w that lies in the largest connected component of $\mathcal{T} \setminus v$. The algorithm repeats this process (moving from v to w) until the largest connected component of $\mathcal{T} \setminus w$ is larger than the largest connected component of $\mathcal{T} \setminus v$ at which point it returns v. We call this the *FindBalance* algorithm (Algorithm 1). The decomposition method that calls *FindBalance* as a subroutine is explained in Algorithm 2.

Algorithm 1 FindBalance

Require: $\mathcal{T}$ is a spanning tree of G and initialize $v \in V$ arbitrarily

loop

Let T' be the component of $\mathcal{T}\backslash\{v\}$ of largest size

Let w be the unique neighbor of v in T'.

Let T'' be the component of $\mathcal{T}\backslash\{w\}$ of largest size.

Stop and return v if $|T''| \geq |T'|$.

$v \leftarrow w$.

end loop

Algorithm 2 Spanning Tree Decomposition

(0) Initialize $\mathcal{D} = \{[p]\}$.

(1) Let v be the output of *FindBalance* applied to $\mathcal{T}$.

(2) Let $\mathcal{T}_1, \ldots, \mathcal{T}_{d_v}$ be the connected components of $\mathcal{T}\backslash v$ and add v to the smallest component where d_v is the degree of v in $\mathcal{T}$.

(3) Add $V(\mathcal{T}_1), \ldots, V(\mathcal{T}_{d_v})$ to $\mathcal{D}$.

(4) For each $i \in \{1, \ldots, d_v\}$, recursively apply (1) - (4) on $\mathcal{T}_i$ as long as $|\mathcal{T}_i| \geq 2$.

Proposition 50. *Let $\mathcal{T}$ be any spanning tree of G, and $\mathcal{D}$ be the decomposition formed by Algorithm 2.*

1. *height$(\mathcal{D})$ is logarithmic in p and*
2. *For any $D \subset D'$ in the decomposition there is an edge in G that is adjacent to both D and $D'\backslash D$.*

Proof. (1) follows from the fact that at each split the size of the largest subtree is at least halved. (2) follows from the fact that each subtree in the recursion $\mathcal{T}_1, \ldots, \mathcal{T}_{d_v}$ is connected its complement within $\mathcal{T}$ by a unique edge. □

With the decomposition we are able to immediately construct a detection scheme with theoretical guarantees that is easily implemented.

5.2.2 Spanning Tree Decompositions Scan Test

First, let us describe the following notation for a graph decomposition. Let $\ell \in 1, \ldots, L = \text{height}(\mathcal{T})$ denote a level of the dendrogram (decomposition), and $\mathcal{D}^\ell = \{D \in \mathcal{D} : \text{level}(D) = \ell\}$ (enumerate the elements of $\mathcal{D}^\ell = \{D_i^\ell\}_{i=1}^{m_\ell}$). The following lemma is the primary mechanism that allows us to obtain our detection results.

Lemma 51. *Let $\{Z_i\}_{i=1}^m$ be independent standard normal variables.*

$$\max_{1 \leq i \leq m} Z_i \leq \sqrt{2 \log m} + \sqrt{2 \log 1/\delta_0}$$

with probability at least $1 - \delta_0$.

Proof. By Cirelson's theorem, we know that

$$\max_{1\le i\le m} Z_i \le \mathbb{E}\max_{1\le i\le m} Z_i + \sqrt{2\log 1/\delta_0}$$

We then apply Dudley's bound, in which $N(\eta)$ is the η covering number of the discrete metric over $[p]$ given by $d(i,j) = I\{i=j\}$.

$$\mathbb{E}\max_{1\le i\le m} Z_i \le \int_0^\infty \sqrt{2\log N(\eta)}d\eta$$

Here $N(\eta) = 0$ if $\eta \ge 1$ and $N(\eta) = p$ otherwise. Evaluating the integral gives us our result. □

Let us introduce the scan test that rejects when

$$\exists \ell \in [L] \text{ s.t. } \max_{i\in m_\ell} \frac{\sum_{v\in D_i^\ell} y_v}{\sigma\sqrt{|D_i^\ell|}} \ge \sqrt{2\log m_\ell} + \sqrt{2\log(L/\delta_0)}$$

Call this test $T(\mathbf{y})$, then we have the following result.

Theorem 52. *Under the null hypothesis, H_0, the probability of type 1 error is bounded by*

$$\mathbb{E}_0 T(\mathbf{y}) \le \delta_0$$

Moreover, under the alternative hypothesis with active cluster C, let ℓ, D_j^ℓ, be any level and cluster in the dendrogram contained in C, $D_j^\ell \subseteq C$. Then the probability of type 2 error is bounded by δ_1, $\mathbb{E}_1(1-T(\mathbf{y})) \le \delta_1$, if

$$\frac{\mu}{\sigma} \ge \sqrt{\frac{|C|}{|D_j^\ell|}}\left(\sqrt{2\log m_\ell} + \sqrt{2\log(L/\delta_0)} + \sqrt{2\log 1/\delta_1}\right)$$

Proof. Consider the null hypothesis, H_0, and the Z-scores,

$$Z_i^\ell = \frac{\sum_{v\in D_i^\ell} y_v}{\sigma\sqrt{|D_i^\ell|}}$$

By Lemma 51, we know that with probability $1-\delta_0$

$$\max_i Z_i^\ell \le \sqrt{2\log m_\ell} + \sqrt{2\log 1/\delta_0}$$

Then we apply the union bound to show that

$$\max_\ell \max_i Z_i^\ell \le \sqrt{2\log m_\ell} + \sqrt{2\log L/\delta_0}$$

Let ℓ, D_j^ℓ be the aforementioned cluster, then under H_1 the Z-score has mean $\mu\sqrt{|D_j^\ell|/|C|}/\sigma$ and variance 1. Hence, we know that with probability $1-\delta_1$,

$$Z_j^\ell \ge \frac{\mu}{\sigma}\sqrt{\frac{|D_j^\ell|}{|C|}} - \sqrt{2\log 1/\delta_1}$$

from which the theorem follows. □

Corollary 53. *Let $r = \max_{C \in \mathcal{C}} |\partial_{\mathcal{T}} C|$, then $T(\mathbf{y})$ asymptotically distinguishes H_0 from H_1 if*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{r \log d_{\max}} \log p\right)$$

Proof. Let $\{D_{j_i}^{\ell}\}_i^{I_\ell}$ be the unique set of clusters in the dendrogram of maximal size within level ℓ such that $D_{j_i}^{\ell} \subseteq C$ for all $i \in I^{\ell}$. For every parent of a maximal cluster there exists a unique edge that disconnects We know that there is a unique edge for each I^{ℓ} in ∂C, because it would require an edge between $D_{j_i}^{\ell}$ and its parent cluster since these clusters are maximal. Hence, $|I^{\ell}| \leq r \log d_{\max}$ and so $\sum_{\ell} |I^{\ell}| \leq r \log d_{\max} \log p$, and so there is a cluster $i' \in I$ such that $|D_{j_{i'}}^{\ell_{i'}}| \geq |C|/(r \log d_{\max} \log p)$. Because $m_{\ell_{i'}} \leq p$ and the height of the dendrogram is less than p, we obtain the result. □

Thus, the performance of the detector is dependent on how many edges in the unknown boundary ∂C are contained in the tree $\mathcal{T}$. We will see in Section 5.4 that this can be guaranteed to be small by merely choosing a spanning tree randomly.

5.3 Localizing Activity with Spanning Tree Wavelets

In this section, we present an algorithm for constructing a wavelet basis over the graph given a decomposition $\mathcal{D}$. We then show that using sparse normal means estimation with this wavelet basis one can recover the true cluster of activation.

5.3.1 Wavelet Construction

We present an algorithm for constructing a wavelet basis given a spanning tree $\mathcal{T}$ and its decomposition $\mathcal{D}$. Informally, we would like to construct a basis $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ such that each edge $e \in \mathcal{T}$ is in the support of few basis elements. By e is in the support of $\mathbf{b}$, we mean that $b_{e^+} \neq b_{e^-}$. Now, to form Haar wavelets we need that clusters are split in a binary fashion, which at present is not the case (a cluster can have as many children clusters as the degree of the balancing vertex). A naive procedure to make this dendrogram binary may result in an intolerable increase in dendrogram height. The primary mechanism that circumvents this is that for a cluster in the decomposition and its children we have to form an intermediate balanced binary dendrogram with the children at its leaves. These smaller intermediate trees have heights that are logarithmic in the number of children, and hence we are able to effectively form a binary decomposition that is has a height logarithmic in the max degree of a vertex and in p. Algorithm 3 explains in detail this procedure.

Proposition 54. *The basis, $\mathbf{B}$, constructed in Algorithm 3 is an orthonormal basis for $\mathbb{R}^p$.*

Proof. We can form a tree with the basis elements as vertices by connecting parents to their children in the construction. This tree has leaves that are the singletons of $[p]$ and is a binary tree. The number of internal vertices in this tree are p, hence this is precisely the number of basis

Algorithm 3 FormWavelets

Require: $\mathcal{D}$ is the spanning tree decomposition.

Initialize $\mathbf{B} = [\frac{1}{\sqrt{p}}\mathbf{1}]$ and $\mathcal{D}_b = \{[p]\}$.

for $D \in \mathcal{D}$ such that D is not singleton **do**

(1) Let $\{D_i\}_{i=1}^k$ be the children of D in $\mathcal{D}$.

(2) Let $D_L = \cup_{i \leq k/2} D_i$ and $D_R = \cup_{i > k/2} D_i$. Add D_L, D_R to $\mathcal{D}_b$ and call these the children of D in this construction.

(3) Form the following basis element and add it to $\mathbf{B}$:

$$\mathbf{b} = \frac{\sqrt{|D_L||D_R|}}{\sqrt{|D_L| + |D_R|}} \left[\frac{1}{|D_L|} \mathbf{1}_{D_L} - \frac{1}{|D_R|} \mathbf{1}_{D_R} \right]$$

(3) Recurse at (1) with D_L and D_R with partitions $\{D_i\}_{i \leq k/2}$ and $\{D_i\}_{i > k/2}$ respectively if the partition is not a single set.

end for

elements. Now, we know that if $\mathbf{b}_0, \mathbf{b}_1$ are constructed from $D_{L,0}, D_{R,0}$ and $D_{L,1}, D_{R,1}$ then

$$\mathbf{b}_0^\top \mathbf{b}_1 = \frac{\sqrt{|D_{L,0}||D_{R,0}|}}{\sqrt{|D_{L,0}| + |D_{R,0}|}} \frac{\sqrt{|D_{L,1}||D_{R,1}|}}{\sqrt{|D_{L,1}| + |D_{R,1}|}} \left(\frac{1}{|D_{L,0}||D_{L,1}|} |D_{L,0} \cap D_{L,1}| \right.$$
$$\left. + \frac{1}{|D_{R,0}||D_{R,1}|} |D_{R,0} \cap D_{R,1}| - \frac{1}{|D_{L,0}||D_{R,1}|} |D_{L,0} \cap D_{R,1}| - \frac{1}{|D_{R,0}||D_{L,1}|} |D_{R,0} \cap D_{L,1}| \right)$$
$$= \begin{cases} 1, & D_{L,0} = D_{L,1}, D_{R,0} = D_{R,} \\ 0, & otherwise \end{cases}$$

This follows from the fact that if the pairs are not identical or all disjoint then either D_{L_0} and D_{R_0} are both contained in D_{L_1} or D_{R_1} or D_{L_1} and D_{R_1} are both contained in D_{L_0} or D_{R_0}. In these latter cases the basis elements are constructed such that their inner product is 0. Finally, we see that

$$\mathbf{b}^\top \mathbf{b} = \frac{|D_L||D_R|}{|D_L| + |D_R|} (1/|D_L| + 1/|D_R|) = 1$$

where $\mathbf{b}$ corresponds to D_L, D_R. □

One main idea behind the specifics of the basis construction is that if $D \in \mathcal{D}$ has children $\{D_i\}_{i=1}^k$, then if an edge e is adjacent to both D_i and $D \backslash D_i$ then the number of basis elements that support e is logarithmic in k. This gives us the following approximation theoretic guarantee.

Proposition 55. *For any $C \subseteq [p]$, $\|\mathbf{B}\mathbf{1}_C\|_0 \leq (1 + |\partial C \cap \mathcal{T}|)(\textit{height}(\mathcal{D}) \lceil \log(d_{\max}) \rceil)$.*

Proof. The idea is that the height of $\mathcal{D}_b$ is logarithmic.

$$\begin{aligned} \|\mathbf{B}\mathbf{1}_C\|_0 &\leq 1 + |\{D_L, D_R \text{ in Alg. 3} : \neg\{D_L \subseteq C, D_R \subseteq C \text{ or } D_L \cap C = \emptyset, D_R \cap C = \emptyset\}\}| \\ &\leq 1 + |\{D_L, D_R \text{ in Alg. 3} : \exists e \in \partial C \cap \mathcal{T}, e^+ \in D_L, e^- \in D_R\}| (\text{height}(\mathcal{D}) \lceil \log(d_{\max}) \rceil) \\ &\leq (1 + |\partial C \cap \mathcal{T}|)(\text{height}(\mathcal{D}) \lceil \log(d_{\max}) \rceil) \end{aligned}$$

The first inequality follows from the definition of the basis elements. The second inequality is due to the fact that for a cluster $D \in \mathcal{D}$ either $D \cap C = \emptyset$, $D \subseteq C$, or $\exists D' \in \mathcal{D} : D' \subseteq C$, for each maximal $D \subseteq C$ we may at most suffer height$(\mathcal{D})\lceil d_{\max} \rceil$ non-zero basis elements. The third inequality follows from the fact that D_L, D_R has a unique $e \in \mathcal{T}$ that is adjacent to both. □

This result gives us the ability to apply any sparse normal means estimation or testing procedure over $\mathbf{By}$ because we know that $\mathbf{Bx}$ is sparse.

5.3.2 Estimation and Localization with Wavelet Thresholding

We will focus on reconstructing C from $\mathbf{y}$. Because we now know that $\|\mathbf{Bx}\|_0$ is small, using hard and soft thresholding of the wavelet coefficients can give us an estimator, $\hat{\mathbf{x}}$, such that the mean square error (MSE), $\|\hat{\mathbf{x}} - \mathbf{x}\|$, is small. After rotation by the eigenvectors the problem falls squarely in the wheelhouse of the sparse normal means estimation work in [15]. Namely, they show that if the signal under a wavelet basis is sparse then hard and soft thresholding can recover the signal with small MSE.

We now recall how one can easily construct an estimate $\hat{C}$ of C from $\hat{\mathbf{x}}$ according to (2.11) in Section 2.4 which we repeat below.

$$\hat{C} = \arg\max_{C \subseteq [n]} \hat{\mathbf{x}}^\top \frac{\mathbf{1}_C}{\sqrt{|C|}}$$

Notice that this is easily computed by greedily including the ordered components of $\hat{\mathbf{x}}$ until the objective is maximized. With the estimators $(\hat{\mathbf{x}}, \hat{C})$ in hand, we can make the following statistical guarantee.

Theorem 56. *Consider the hard thresholding estimator*

$$\hat{u}_i = (\mathbf{b}_i^\top \mathbf{y}) I\{|\mathbf{b}_i^\top \mathbf{y}| > \sigma\sqrt{2\log p}\}$$

Then form the estimate $\hat{C}$ from the estimate $\hat{\mathbf{x}} = \mathbf{B}\hat{\mathbf{u}}$ according to (2.11)*. We have the following risk bound for estimation,*

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \leq (2\log p + 1.2)(\sigma^2 + \min\{\mu^2, k\sigma^2\})$$

and the following risk bound for localization,

$$d(\hat{C}, C) \leq 2\left(4\frac{\sigma^2}{\mu^2}(k+1)(2\log p + 1.2) + 2\frac{\sigma}{\mu}\sqrt{(k+1)(2\log p + 1.2)}\right)$$

assuming that $\mu/\sigma = \omega(1)$ and $k = |\partial C \cap \mathcal{T}|L$ where L is the height of the spanning tree basis dendrogram. The same bound holds for soft thresholding with $(2\log p + 1)$.

Proof. The theorem follows by applying the following risk bound of the hard thresholding (it can be found as Proposition 8.6 in [41]).

$$\|\hat{\mathbf{u}} - \mathbf{B}^\top \mathbf{u}\|^2 \leq (2\log p + 1.2)(\sigma^2 + \sum_{i=1}^{p} \min\{(\mathbf{b}_i^\top \mathbf{x})^2, \sigma^2\}) \leq (2\log p + 1.2)(\sigma^2 + \min\{\mu^2, k\sigma^2\})$$

where $k = rL$ where L is the height of the spanning tree basis dendrogram. We can then combine this with Proposition 9 to obtain the result. □

Corollary 57. *The hard (and soft) thresholding are localization and ℓ_2 consistent if*

$$\frac{\mu}{\sigma} = \omega(\sqrt{r \log d_{\max}} \log p)$$

where $r = \max\{|\partial C \cap \mathcal{T}| : |\partial C| \leq \rho\}$.

Remark 58. *Of course we know that $r \leq \rho$ regardless of the spanning tree $\mathcal{T}$. Hence, regardless of the spanning tree construction we know that*

$$\frac{\mu}{\sigma} = \omega(\sqrt{\rho \log d_{\max}} \log p)$$

is sufficient for localization consistency of the hard thresholding wavelet estimator.

As in the detection performance, Corollary 53, we relate the performance of the estimator to how many edges in the unknown boundary ∂C are contained in the tree $\mathcal{T}$. The following section discusses the choice of spanning tree, and how it effects these bounds.

5.4 Uniform Spanning Tree Basis

The uniform spanning tree (UST) is a random spanning tree generation technique that we will use to construct wavelet bases. We will first examine the deep connection between electrical networks, USTs and random walks. Because the UST is randomly generated, the test statistic $T(\mathbf{y})$ when conditioned on $\mathbf{y}$ will also be random. Moreover the estimator $\hat{C}$ is random even when conditioned on $\mathbf{y}$. Due to results from cut sparsification, we can relate the performance of the UST decomposition detector to effective resistances.

5.4.1 Cuts and Effective Resistance

Effective resistances have been extensively studied in electrical network theory. We define the combinatorial Laplacian of G to be $\Delta = \nabla^\top \nabla$. A *potential difference* is any $\mathbf{z} \in \mathbb{R}^{|E|}$ such that it satisfies *Kirchoff's potential law*: the total potential difference around any cycle is 0. Algebraically, this means that $\exists \mathbf{x} \in \mathbb{R}^{|V|}$ such that $\nabla \mathbf{x} = \mathbf{z}$. The *Dirichlet Principle* states that any solution to the following program gives an absolute potential $\mathbf{x}$ that satisfies Kirchoff's potential law:

$$\min_{\mathbf{x}} \mathbf{x}^\top \Delta \mathbf{x} \text{ s.t. } \mathbf{x}_S = \mathbf{v}_S$$

for source/sinks $S \subset V$ and some voltage constraints $\mathbf{v}_S \in \mathbb{R}^{|S|}$. By Lagrangian calculus, the solution to the above program is given by $\mathbf{x} = \Delta^\dagger \mathbf{v}$ where $\mathbf{v}$ is 0 over S^C and $\mathbf{v}_S$ over S, and $\dagger$ indicates the Moore-Penrose pseudoinverse. The effective resistance between a source $v \in V$ and a sink $w \in V$ is the potential difference required to create a unit flow between them. Hence, the effective resistance between v and w is $r_{v,w} = (\delta_v - \delta_w)^\top \Delta^\dagger (\delta_v - \delta_w)$, where δ_v is the Dirac delta function.

A massively useful characterization of effective resistance is the random walk interpretation. Let X_t be the location of a random walker on G at time t. The hitting time $H(v, w)$ is then

$$H(v, w) = \mathbb{E}[\min\{t > 0 : X_t = w\} | X_0 = v]$$

We find that the effective resistance is related to the hitting time by,

$$r_{v,w} = \frac{H(v, w) + H(w, v)}{2m}$$

The numerator is also known as the commute time. As we will see, this characterization of effective resistance is useful when bounding it for specific graph models.

5.4.2 UST Detector and Estimator

We will now examine the performance of the spanning tree detector and wavelet estimator, when the spanning tree is drawn according to a UST. First, we will explore the construction of the UST and examine key properties. The UST is a random spanning tree, chosen uniformly at random from the set of all distinct spanning trees. The foundational Matrix-Tree theorem [43] describes the probability of an edge being included in the UST. The following lemma can be found in [53] and [54].

Lemma 59. *Let G be a graph, $\mathcal{T}$ a draw from $\mathrm{UST}(G)$, and let e be any edge in E. Then,*

$$\mathbb{P}\{e \in \mathcal{T}\} = r_e$$

Hence, we can expect that for a given cut in the graph, the cut size in the tree will look like the sum of effective resistances of edges in the cut. If these effective resistances are small then $|\partial C \cap \mathcal{T}|$ will be small and we can improve on the bound in Remark 58. While it is infeasible to enumerate all spanning trees of a graph, the Aldous-Broder algorithm is an efficient method for generating a draw from $\mathrm{UST}(G)$ [2]. The algorithm simulates a random walk $\{X_t\}$ on G, stops when all of the vertices have been visited, and defines the spanning tree $\mathcal{T}$ by the edges $\{(X_{H(X_0,v)-1}, v) : v \in V\}$. The computational complexity of the Aldous-Broder algorithm is the expected cover time of G, which is $O(p \log p)$ for expander graphs, d-regular graphs, and many other graph models, but is $O(p^3)$ in the worst case [10].

In order to control $\|\partial C \cap \mathcal{T}\|_0$, we need to control the overlap between a cut in the graph and the UST. Clearly the UST does not independently sample edges, but it does have the well documented property of negative association, that the inclusion of an edge decreases the probability that another edge is included. The following lemma states a concentration result for the UST, based on negative association, and can be found in [28]. The proof is a simple extension of the concentration results in [29].

Lemma 60. *Let $B \subset E$ be a fixed subset of edges, and $|\mathcal{T} \cap B|$ denote the number of edges in $\mathcal{T}$ also in B.*

$$\mathbb{P}\{|\mathcal{T} \cap B| \geq (1 + \delta) \sum_{e \in B} r_e\} \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{\sum_{e \in B} r_e}$$

We use this result to give conditions under which the UST decomposition detector asymptotically distinguishes H_0 from H_1.

Theorem 61. *Let $r_{\max} = \max_{C \in \mathcal{C}} \sum_{e \in C} r_e$ (the maximum effective resistance of the cut of a pattern in $\mathcal{C}$).*
(1) H_0 and H_1 are asymptotically distinguished by $T(\mathbf{y})$ if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{r_{\max} \log p}\right)$$

(2) The estimator $\hat{C}$ defined in (2.11) *for hard or soft thresholding is localization consistent ($d(\hat{C}, C) \to 0$) if*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{r_{\max} \log d_{\max}} \log p\right)$$

Proof. Let $r_B = \sum_{e \in B} r_e$ for $B \subset E$. By some basic calculus, and the fact that $\log(1+x) \geq x/(1+x/2)$, we see that

$$\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{r_B} \leq \exp\left(-\frac{\delta^2 r_B}{2+\delta}\right)$$

Rewriting the Lemma 60, we obtain with probability $> 1 - \gamma$

$$\begin{aligned} |\mathcal{T} \cap B| &\leq r_B + \sqrt{2 r_B \log\frac{1}{\gamma} + \frac{1}{4}(\log\frac{1}{\gamma})^2} + \frac{1}{2}\log\frac{1}{\gamma} \\ &\leq \left(r_B + \sqrt{2 r_B \log\frac{1}{\gamma}} + \log\frac{1}{\gamma}\right) \end{aligned}$$

Now, because $\|\nabla_{\mathcal{T}} \mathbf{x}\|_0 = |\mathcal{T} \cap B|$ for $B = \text{supp}(\nabla_{\mathcal{T}} \mathbf{x})$, we know by Theorem 52 if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\left(r_B + \sqrt{2 r_B \log\frac{1}{\gamma}} + \log\frac{1}{\gamma}\right) \log d} \log p\right)$$

then H_0 and H_1 are asymptotically distinguished and the result follows because we guarantee this for all such B. The same analysis holds for the hard thresholding estimator. □

While these guarantees may at first glance be merely of theoretical importance, we know that one can easily compute the effective resistances of a graph through a pseudoinverse calculation. Thus, the practicing statistician is able to precompute the distribution of effective resistances and determine how well or how poorly the UST might subsample the cut boundary. In the following section we will see that in fact we can bound theoretically the effective resistance of cuts for a variety of real world graph models.

5.5 Specific Graph Models

In this section we study our detection problem for several different families of graphs. Specifically, we control the effective resistance r_e for each graph family, which when combined with

Theorem 61 gives a lower bound on the SNR for which $T(\mathbf{y})$ asymptotically distinguishes H_0 and H_1. Moreover, this will give us the localization consistency of $\hat{C}$ defined (2.11).

In Theorem 61, we showed that the distinguishability regime depends on the effective resistances of the cuts induced by the class of activation patterns $\mathcal{X}$. On its own, it is not immediately clear that this result is an improvement over Remark 58 that we would obtain from any spanning tree. However, Foster's theorem highlights why we expect the effective resistance to be less than the cut size.

Theorem 62 (Foster's Theorem [24, 76])**.**

$$\sum_{e \in E} r_e = p - 1$$

Hence, if we select an edge uniformly at random from the graph, we expect its effective resistance to be $(p-1)/m \approx \bar{d}^{-1}$ (the reciprocal of the average degree $\bar{d}$) where $m = |E|$.

5.5.1 Edge Transitive Graphs

An edge transitive graph, G, is one such that for any edges e_0, e_1, there is a graph automorphism that maps e_0 to e_1. Examples of edge transitive graphs include the l-dimensional torus and the complete graph K_n. For such a graph, every edge has the same effective resistance, and Foster's Theorem then shows that $r_e = (n-1)/m$ where m is the number of edges. Moreover since edge transitive graphs must be d-regular for some degree d, we see that $m = \Theta(nd)$ so the $r_e = \Theta(1/d)$. This leads us to the following corollary, which we note matches the lower bound in Theorem 47 modulo logarithmic terms if $\rho/d \leq \sqrt{n}$:

Corollary 63. *Let G be edge transitive with common degree d. Then for each edge $e \in E(G)$, $r_e = (p-1)/m$.*
(1) H_0 and H_1 are asymptotically distinguished by $T(\mathbf{y})$ if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{d}\log p}\right)$$

(2) The hard (and soft) thresholding wavelet estimator is localization consistent ($d(\hat{C}, C) \to 0$) if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{d}\log d \log p}\right)$$

We simulate the detection performance for the 2 dimensional torus graph. We report the ROC curve, the true positive rate versus the false positive rate, for the tree wavelet detector with 1000 simulations from H_0 and H_1 (Figure 5.1). The cluster, C, is chosen to be a square of size $\asymp p^{1-\beta}$, which means that the cut size is $\rho \asymp p^{(1-\beta)/2}$. We compare the performance of the UST detector to the adaptive GESS, max and aggregate statistic. We find that the UST detector performs better than the adaptive GESS in the sparse regime ($\beta = 0.75$), and is competitive otherwise.

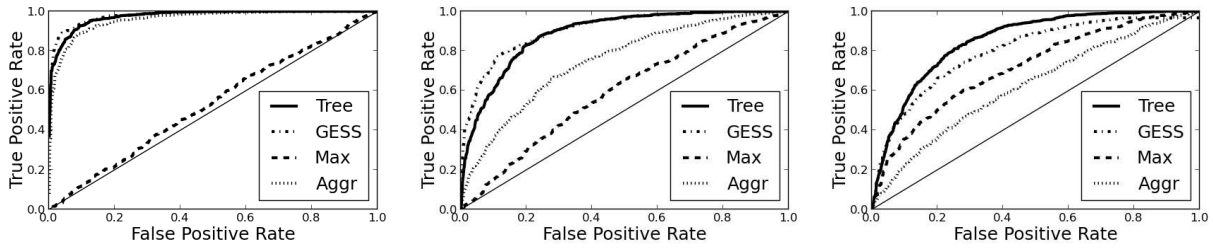


Figure 5.1: **(Torus Comparison)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the Torus of the UST detector (Tree), adaptive GESS (GESS), Max statistic (Max), and Aggregate statistic (Aggr). The figures are $p = 40^2 = 1200$, with choice of $\beta = .25$ (left), $\beta = .5$ (middle) and $\beta = .75$ (right).

5.5.2 kNN Graphs

Oftentimes in applications, the graph topology is derived from data. In this case, the randomness of the data means that the graph itself is inherently random. Commonly, these graphs are modeled as random geometric graphs, and in this section we will devote our attention to the *symmetric k-nearest neighbor graphs.* Specifically, suppose that $\mathbf{z}_1, ..., \mathbf{z}_p$ are drawn IID from a density p supported over $\mathbb{R}^D$. Then we form the graph G over $[p]$ by connecting vertices i, j if $\mathbf{z}_i$ is amongst the k-nearest neighbors of $\mathbf{z}_j$ or vice versa. Some regularity conditions of p are needed for our results to hold; they can be found in [81].

To bound the effective resistance r_e, Corollary 9 in [81] shows that $H_{ij}/2m \to 1/d_j$ and by the definition of r_e we see that $r_{ij} \to \frac{1}{d_i} + \frac{1}{d_j} \leq \frac{2}{k}$, since $d_i \geq k$ for each i. A formal analysis leads to the following corollary, which we prove in [68] with more precise concentration arguments:

Corollary 64. *Let G be a k-NN graph with $k/p \to 0$ and $k(k/p)^{2/D} \to \infty$ and where the density p satisfies the regularity conditions in [81].*
(1) H_0 and H_1 are asymptotically distinguished by $T(\mathbf{y})$ if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{k}\log p}\right)$$

(2) The hard (and soft) thresholding wavelet estimator is localization consistent ($d(\hat{C}, C) \to 0$) if

$$\frac{\mu}{\sigma} = \omega(\sqrt{\frac{\rho}{k}\log k \log p})$$

We simulate the detection performance for the kNN graph. We report the ROC curve, the true positive rate versus the false positive rate, for the tree wavelet detector with 1000 simulations from H_0 and H_1 (Figures 5.2, 5.3). The cluster, C, is chosen by selecting a random centroid vertex and expanding a ball in the extrinsic space until it contains the desired number of vertices. k was set to be $\lfloor\sqrt{p}\rfloor$ while $|C|$ varied between 10, 15, and 20. Somewhat surprisingly the UST detector does not dominate the adaptive GESS, but in all cases it is competitive.

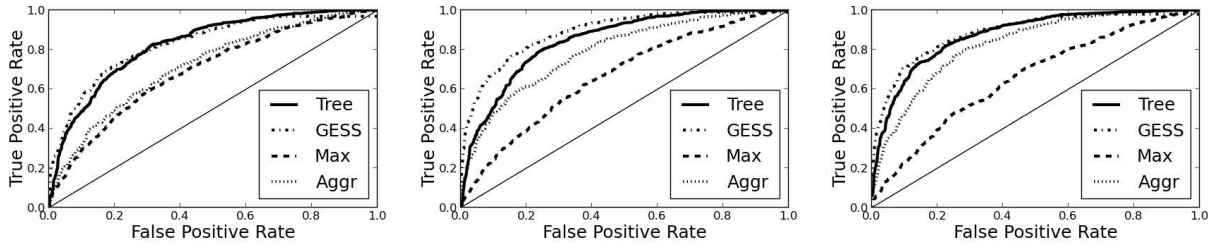


Figure 5.2: **(kNN Comparison)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the kNN of the tree detector (Tree), adaptive GESS (GESS), Max statistic (Max), and Aggregate statistic (Aggr). The figures are $p = 200$, $k \approx p^{.5}$, with choice of $|C| = 10$ (left), $|C| = 15$ (middle) and $|C| = 20$ (right).

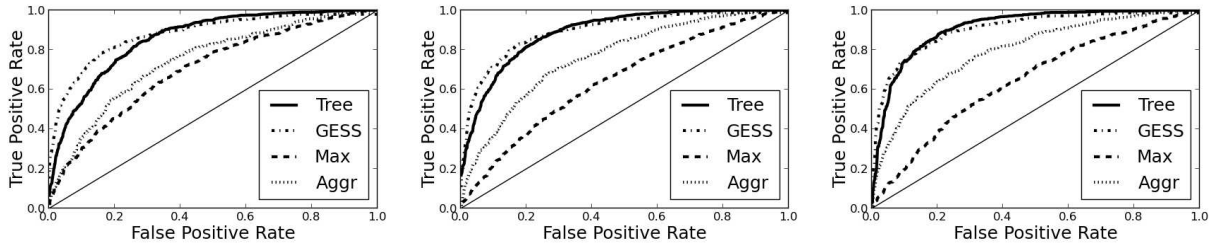


Figure 5.3: **(kNN Comparison)** Simulations of the size (false positive rate) and the power under H_1^{PC} for the kNN of the tree detector (Tree), adaptive GESS (GESS), Max statistic (Max), and Aggregate statistic (Aggr). The figures are $p = 200$, $k \approx p^{.33}$, with choice of $|C| = 10$ (left), $|C| = 15$ (middle) and $|C| = 20$ (right).

5.5.3 ϵ-Graphs

The ϵ-graph is another widely used random geometric graph in machine learning and statistics. As with the k-NN graph, the vertices are embedded into $\mathbb{R}^D$ and edges are added between pairs of vertices that are within distance ϵ of each other. As with the k-NN graph, Corollary 8 from [81] shows that $H_{ij} \to m/d_j$ for each pair of vertices. This leads us to believe that $r_{ij} \to 1/(d_i) + 1/(d_j)$. If the density p from which we draw data points is bounded from below by some constant, then we can uniformly lower bound all of the degrees d_i using fairly elementary concentration results, which results in an upper bound on r_e. Formalizing this intuition, we have the following corollary, which we prove in [68]:

Corollary 65. *Let G be a ϵ-graph with points $X_1, \dots X_n$ drawn from a density, which satisfies the regularity conditions in [81] and is lower bounded by some constant (independent of n) and $\epsilon \to 0, n\epsilon^{D+2} \to \infty$.*
(1) H_0 and H_1 are asymptotically distinguished by $T(\mathbf{y})$ if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{p\epsilon^D}\log p}\right)$$

(2) The hard (and soft) thresholding wavelet estimator is localization consistent ($d(\hat{C}, C) \to 0$) if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{p\epsilon^D}\log d_{\max}}\log p\right)$$

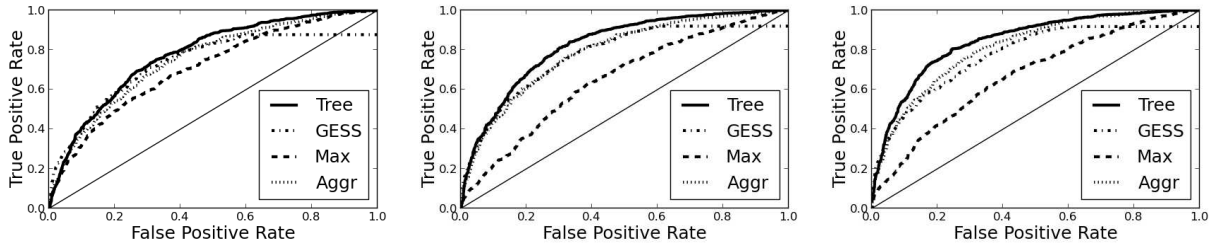


Figure 5.4: (**ϵ-graph Comparison**) Simulations of the size (false positive rate) and the power under H_1^{PC} for the Torus of the tree detector (Tree), adaptive GESS (GESS), Max statistic (Max), and Aggregate statistic (Aggr). The figures are $p = 300$, with choice of $|C| = 10$ (left), $|C| = 15$ (middle) and $|C| = 20$ (right).

We simulate the detection performance for the $\epsilon-$graph. We report the ROC curve, the true positive rate versus the false positive rate, for the tree wavelet detector with 1000 simulations from H_0 and H_1 (Figure 5.4). The cluster, C, is chosen by selecting a random centroid vertex and expanding a ball in the extrinsic space until it contains the desired number of vertices. ϵ was set to be $p^{-1/5}$ while $|C|$ varied between 10, 15, and 20. For these simulations, the UST detector dominates the adaptive GESS and the naive estimators uniformly.

5.6 Discussion

In this Chapter, we studied the detection of piece-wise constant activation patterns over graphs, and provided a necessary condition for the asymptotic distinguishability of signals that are assumed to have few discontinuities. We gave a novel spanning tree wavelet construction, that is the extension of the Haar wavelet basis, for arbitrary graphs and proposed a detector relying on the largest wavelet coefficient obtained by projecting the observations onto the basis. The decomposition detector constructed using a uniform spanning tree was shown to have strong theoretical guarantees that in many cases gives us near optimal performance. This means that under adversarial choice of patterns, our randomized algorithm asymptotically distinguishes H_0 from H_1 at near optimal signal-to-noise ratios. Alternatively, this means that for any given activation pattern (non-adversarial setting) that the vast majority of spanning trees induce detectors that asymptotically distinguish H_0 from H_1 near optimally.

The UST wavelet construction is done is such a way to ensure that the resulting basis is localized and approximates signals with low cut size. If one were to ask for a localized basis that approximated other signal classes, then this might radically alter the wavelet construction. For example, if one desired a localized basis that approximated the ℓ_2 graph structured signals then it would be more appropriate to turn to the Laplacian eigenvectors, using some method such as binary cuts to localize the basis. Furthermore, one promising direction is to use the work of [62], which has a decomposition technique with guarantees that may be extended to an approximation theoretic result.

Notice that in this Chapter, we observe a very different behavior in the critical SNRs than those in Chapter 4. This highlights some of the fundamental difference between ℓ_0 graph structure, $\mathcal{X}_0$, and balanced graph structure, $\mathcal{X}_{PC}$, as alternative spaces. Moreover, we have argued the estimators from very different perspectives. The GESS is a relaxation of the GLR statistic, which is computationally infeasible. The decomposition is motivated as a reasonable alternative to likelihood based procedures for the ℓ_0 graph structure. This leads us to the obvious question, is a more likelihood based approach, such as the edge lasso superior to the decomposition approach? This is an object of current study, and remains an open problem. Much of the impetus for the formation of a decomposition on general graphs was explicitly the construction of the wavelet basis over general graphs. The fact that this wavelet basis resulted in an estimation procedure is a byproduct of the approximation guarantee Proposition 55.

5.6.1 Discussion of the Thesis

The main message of this thesis is that by exploiting graph structure as informed prior information for statistical estimation, one is able to detect, localize, and estimate under lower SNR regimes than would be possible without such prior knowledge. We saw this in Chapter 2, where Theorem 1 shows that one can bound the MSE of Laplacian eigenmaps through quantiles of the graph spectrum. In Chapter 3, Theorem 20 states that under conditions that could be related to the graph Laplacian and the underlying signal lead to near optimal changepoint sparsistency. In Chapter 4, the performance of the GESS is quantified in Theorem 34, where the critical SNR is equivalent to a functional of the graph spectrum. In Chapter 5, we developed a graph wavelet basis, and in Theorem 56, we proved that thresholding in this basis had a natural risk bound.

Throughout this thesis, we made a variety of simplifying assumptions. Generally, this work is the first comprehensive study of graph structured normal means inference, and simplifying assumptions were made because it would not be in the scope of the thesis to generalize the setting any more. Specifically, the noise variance σ is assumed to be known. If it were not known, one recommendation would be to estimate it from the eigenvectors of the Laplacian with large eigenvalue. This would introduce an additional tuning parameter, but perhaps an adaptive method would allow us to find it effectively.

Another assumption is that the graph is assumed to be known. In practice one may obtain a graph with edges missing, falsely added, or misspecified weights. There are two approaches that one may take when it is known that the graph is misspecified. First, one may proceed with the inference procedures given in the proceeding chapters, and verifying that the effect of deviations from the assumed graph does not have a significant effect on the statistical power. Second, one may model the graph misspecification, which would require new estimators with a new theoretical analysis.

This work is only the opening salvo in a possibly long line of research. Indeed, much of the work in this thesis was merely finding appropriate formulations of the detection and estimation problems. It is no coincidence that each chapter left its author with more questions than answers. It seems apparent that this line of research exemplifies the statistical and computational tradeoffs inherent in combinatorial testing. We have seen this in the need for a spectral relaxation of the GLR statistic, and the computational feasibility of the UST wavelet basis. While there is a multitude of results waiting to be discovered regarding graph structured normal means estimation, we feel confident that this thesis provides insights into the key theoretical and computational issues.

Bibliography

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010. 1, 1.1.3, 1.2, 4, 4.2.1, 4.3.1, 5.1

[2] D. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3(4):450–465, 1990. 5.4.2

[3] E. Arias-Castro, E. Candes, and A. Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011. 1, 1.1.3, 1.2, 4, 4.3.1

[4] E. Arias-Castro, E. Candes, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36(4):1726–1757, 2008. 1.2, 4, 4.3.1, 5.1

[5] E. Arias-Castro, D. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, 2005. 1.2, 4, 4.3.1

[6] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009. 4.3.1

[7] B. Baygün and A. O. Hero. Optimal simultaneous detection and estimation under a false alarm constraint. *Signal Processing, IEEE Transactions on*, 41(3):688–703, 1995. 1.2, 4.3.1

[8] E. Belitser and B. Levit. Asymptotically minimax nonparametric regression in l2. *Statistics: A journal of theoretical and applied statistics*, 28(2):105–122, 1996. 1.2

[9] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001. 2.2.3

[10] A. Broder. Generating random spanning trees. *Foundations of Computer Science*, 1989. 5.4.2

[11] L. Carvalho and C. Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences*, 105(9):3209, 2008. 1.2

[12] V. Cevher, C. Hegde, M. Duarte, and R. Baraniuk. Sparse signal recovery using markov random fields. Technical report, DTIC Document, 2009. 1, 1.2

[13] F. Chung. Discrete isoperimetric inequalities. *Surveys in Differential Geometry IX, International Press*, pages 53–82, 2004. 4.3.1

[14] R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006. 1.2

[15] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage:

asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995. 5.3.2

[16] S. Efroimovich and M. Pinsker. Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii*, 18(3):19–38, 1982. 1.2

[17] P. Erdös and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960. 2.2.3

[18] M. S. Ermakov. Minimax detection of a signal in a gaussian white noise. *Theory of Probability & Its Applications*, 35(4):667–679, 1990. 4.2.2

[19] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999. 4.6.3

[20] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek. Spectra of real-world graphs: Beyond the semi-circle law. *Physical Review E*, 64:1–12, 2001. 2.2.3, 2.3

[21] M. Fiedler. Eigenvectors of acyclic matrices. *Czechoslovak Mathematical Journal*, 25(4):607–618, 1975. 4.6.1

[22] L. Fillatre. Asymptotically uniformly minimax detection and isolation in network monitoring. to appear in Signal Processing, IEEE Transactions on, 2012. 4.3.1

[23] L. Fillatre and I. Nikiforov. Non-bayesian detection and detectability of anomalies from a few noisy tomographic projections. *Signal Processing, IEEE Transactions on*, 55(2):401–413, 2007. 4.3.1

[24] R. Foster. The average impedance of an electrical network. *Contributions to Applied Mechanics (Reissner Anniversary Volume)*, pages 333–340, 1949. 62

[25] M. Fouladirad, L. Freitag, and I. Nikiforov. Optimal fault detection with nuisance parameters and a general covariance matrix. *International Journal of Adaptive Control and Signal Processing*, 22(5):431–439, 2008. 1.2, 4.3.1

[26] M. Fouladirad and I. Nikiforov. Optimal statistical fault detection with nuisance parameters. *Automatica*, 41(7):1157–1171, 2005. 1.2, 4.3.1

[27] B. Friedman. Eigenvalues of composite matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57:37–49, 1961. 2.2.2

[28] W. Fung and N. Harvey. Graph sparsification by edge-connectivity and random spanning trees. *Arxiv preprint arXiv:1005.0265*, 2010. 5.4.2

[29] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006. 5.4.2

[30] M. Gavish, B. Nadler, and R. Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proc. International Conference on Machine Learning, Haifa, Israel*, 2010. 1.2, 2.2.1

[31] C. Godsil and G. Royle. *Algebraic graph theory*, volume 8. Springer New York, 2001. 3.0.1

[32] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(9):1074–1085, 1992. 4.3.1

[33] Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010. 1.2

[34] H. Hoefling. A path algorithm for the fused lasso signal approximator. Technical report, October 2009. 1.2

[35] Y. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability and its Applications*, 31:333, 1987. 1.2

[36] Y. Ingster and I. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Verlag, 2003. 1.2, 4.2.2

[37] L. Jacob, P. Neuvial, and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. *Arxiv preprint arXiv:1009.5173*, 2010. 1.2

[38] S. Jalan and J. N. Bandyopadhyay. Random matrix analysis of network laplacians. Technical Report cond-mat/0611735, Nov 2006. 8

[39] J. Jin and D. L. Donoho. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004. 2.5

[40] I. Johnstone. Minimax bayes, asymptotic minimax and sparse wavelet priors. *Statistical Decision Theory and Related Topics, Springer*, pages 303–326, 1994. 1.2

[41] I. Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2002. 2.5, 5.3.2

[42] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 321, 2003. 1.2

[43] G. Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508, 1847. 5.4.2

[44] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 315–322. Citeseer, 2002. 1.2

[45] I. Koutis, A. Levin, and R. Peng. Faster spectral sparsification and numerical algorithms for sdd matrices. *arXiv preprint arXiv:1209.5821*, 2012. 43

[46] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The annals of Statistics*, 28(5):1302–1338, 2000. 4.3.2, 35

[47] A. Lee, B. Nadler, and L. Wasserman. Treeletsan adaptive multi-scale basis for sparse unordered data. *The Annals of Applied Statistics*, 2(2):435–471, 2008. 1.2

[48] E. Lehmann and J. Romano. *Testing statistical hypotheses*. Springer Verlag, 2005. 4.3.1

[49] T. Leighton and S. Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. In *Foundations of*

Computer Science, 1988., 29th Annual Symposium on, pages 422–431. IEEE, 1988. 4.3.1

[50] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010. 4.6, 4.6.3

[51] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *Proceedings of the 24th international conference on Machine learning*, pages 497–504. ACM, 2007. 4.6, 4.6.3

[52] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *In ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010. 1.2

[53] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993. 5.4.2

[54] R. Lyons and Y. Peres. *Probability on trees and networks*. 2000. 5.4.2

[55] A. Madry, G. Miller, and R. Peng. Electrical flow algorithms for total variation minimization. *Arxiv preprint arXiv:1110.1358*, 2011. 1

[56] P. Mahé, N. Ueda, T. Akutsu, J. Perret, and J. Vert. Extensions of marginalized graph kernels. In *Proceedings of the twenty-first international conference on Machine learning*, page 70. ACM, 2004. 1.2

[57] D. Matula and F. Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1):113–123, 1990. 4, 4.3.1

[58] R. Merris. Laplacian graph eigenvectors. *Linear algebra and its applications*, 278(1):221–236, 1998. 4.6.3

[59] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967. 4.6.3

[60] J. Molitierno, M. Neumann, and B. Shader. Tight bounds on the algebraic connectivity of a balanced binary tree. *Electronic Journal of Linear Algebra*, 6:62–71, 2000. 4.6.1

[61] J. Pearl and M. Tarsi. Structuring causal trees. *Journal of Complexity*, 2(1):60–77, 1986. 5.2.1

[62] H. Racke. Minimizing congestion in general networks. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 43–52. IEEE, 2002. 5.6

[63] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. 2006. 1, 1.2

[64] A. Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009. 1.2

[65] O. Rojo. The spectrum of the laplacian matrix of a balanced binary tree. *Linear algebra and its applications*, 349(1):203–219, 2002. 4.6.1

[66] O. Rojo and R. Soto. The spectra of the adjacency matrix and laplacian matrix for some balanced trees. *Linear algebra and its applications*, 403:97–117, 2005. 4.6.1

[67] L. L. Scharf and B. Friedlander. Matched sub-space detectors. *Signal Processing, IEEE Transactions on*, 42(8):2146–2157, 1994. 1.2, 4.3.1

[68] J. Sharpnack, A. Krishnamurthy, and A. Singh. Detecting activations over graphs using spanning tree wavelet bases. *Arxiv preprint arXiv:1206.0937*, 2012. 5.5.2, 5.5.3

[69] J. Sharpnack, A. Rinaldo, and A. Singh. Changepoint detection over graphs with the spectral scan statistic. *Arxiv preprint arXiv:1206.0773*, 2012. 4.3

[70] J. Sharpnack, A. Rinaldo, and A. Singh. Sparsistency of the edge lasso over graphs. *AIStats (JMLR WCP)*, 22:1028–1036, 2012. 3

[71] J. Sharpnack and A. Singh. Identifying graph-structured activation patterns in networks. In *Proceedings of Neural Information Processing Systems, NIPS*, 2010. 2, 3, 4.6.1, 4.6.2

[72] D. Shmoys. Cut problems and their application to divide-and-conquer. *Approximation algorithms for NP-hard problems*, pages 192–235, 1997. 4.3.1

[73] A. Singh, R. Nowak, and R. Calderbank. Detecting weak but hierarchically-structured patterns in networks. *Arxiv preprint arXiv:1003.0205*, 2010. 1.2, 2.2.1, 4.6.1

[74] A. Smola and R. Kondor. Kernels and regularization on graphs. *Learning theory and kernel machines*, pages 144–158, 2003. 1.2

[75] M. Talagrand. *The generic chaining*. Springer, 2005. 3.1, 11

[76] P. Tetali. Random walks and the effective resistance of networks. *Journal of Theoretical Probability*, 4(1):101–109, 1991. 62

[77] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B*, 67:91–108, 2005. 1.2

[78] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. 05 2010. 1.1.2, 1.2, 3.2, 3.2, 3.2, 3.2

[79] A. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009. 1.2

[80] S. Vishwanathan, K. Borgwardt, I. Kondor, and N. Schraudolph. Graph kernels. *Arxiv preprint arXiv:0807.0093*, 2008. 1.2

[81] U. Von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large graphs are often misleading. *ReCALL*, 2010. 5.5.2, 64, 5.5.3, 65

[82] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using. *IEEE transactions on information theory*, 55(5):2183, 2009. 3.2.1

[83] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of American Mathematical Society*, 54:426–482, 1943. 4.3.1

[84] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. 2.3

[85] C. Zhan, G. Chen, and L. F. Yeung. On the distribution of laplacian eigenvalues versus node degrees in complex networks. *Physica A*, 389:1779–1788, 2010. 8

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA15213

Carnegie Mellon