Sparsistency of the Edge Lasso over Graphs

James Sharpnack
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
jsharpna@andrew.cmu.edu

Alessandro Rinaldo
Statistics Department
Carnegie Mellon University
Pittsburgh, PA 15213
arinaldo@cmu.edu

Aarti Singh
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
aartisingh@cmu.edu

Abstract

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 paper, we investigate *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 sparsistency. We examplify our results using simple graph-structured patterns, and demonstrate that in some cases fused lasso is sparsistent at very weak signal-to-noise ratios (scaling as $\sqrt{(\log n)/|A|}$, where n is the number of vertices in the graph and A is the smallest set of vertices with constant activation). 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 $\sqrt{\log n}$).

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

1 Introduction

In this paper, we consider the problem of correctly identifying the locations of a piece-wise constant signal over the vertices of a network from noisy observations in the high-dimensional setting. Specifically, for a given network $G = (V, E)$, we observe one realization of the random vector

$$y = \beta + \epsilon,$$

where $\beta \in \mathbb{R}^V$ and $\epsilon \sim N(0, \sigma^2 I)$, with σ^2 known. The vector β is piece-wise constant over elements of an unknown partition of the vertices V, where each element of the partition corresponds to a connected induced sub-graph of G. We seek to (1) perfectly identify such a partition and (2) determine the signs of all pairwise differences $\beta_v - \beta_{v'}$, where $(v, v') \in E$. These properties are known as *sparsistency*.

The motivations for studying such problem are multiple. From the practical standpoint, the localization of patterns in a network is an important task in a variety of applications, ranging from anomaly detection in sensor networks to disease outbreak detection, community extraction in social networks, identification of differentially expressed set of genes in microarray data analysis and virus spread detection in the Internet.

From a theoretical perspective, the problem described above is an instance of the *structured* normal means problem, a variant of the classical normal means problem in which the possible patterns of non-zero entries of the mean vector are prescribed as connected subgraph of a given graph. Existing analyses of the problems of signal detection and localization in the structured normal means setting are fairly recent: see, in particular, [3], [1], [2], [12], [4]. However, these primarily consider simple graph topologies such as linear [4], lattice [4, 2, 3] or tree [3, 12]. More general graph structures are considered in [11] and [1], however these do not guarantee exact recovery of the support of the pattern. The former focuses on ℓ_2 or hamming dis-

tance recovery under a probabilistic graphical model, while the latter focuses on testing. Furthermore, some of these works assume that the activation cluster size is known and some rely on procedures that are computationally infeasible for large graphs.

In this paper, we focus on the fused lasso originally proposed in [14] 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 [15, 7, 8] 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, is [9] which analyzes the linear graph topology, assuming that the signal-to-noise ratio (SNR) increases with graph size. In this paper, we investigate sparsistency of fused lasso for general graph structures and provide a more refined analysis of the SNR required for different graph structures. To emphasize this distinction, we call it Edge Lasso.

We will begin by introducing some mathematical terminology, concerning the oriented graph. Consider a undirected graph G defined by a set of vertices V and undirected edges E which are unordered pairs of vertices. We construct an **orientation** of G by defining a head $e^+ \in e$ and tail $e^- \in e$. The **incidence matrix** $D \in \mathbb{R}^{E \times V}$ for the oriented graph is the matrix whose $D_{e,v}$ entry is 1 if $v = e^+$, -1 if $v = e^-$ and 0 otherwise.

We suppose that there is a partitioning of the vertices, $\mathcal{A} = \{A_i\}_{i=1}^k$, defined by maximal subgraphs of constant signal. Henceforth, without loss of generality, we will assume that each $A \in \mathcal{A}$ is connected. For a vertex $v \in V$, denote $A(v) = A_i$ for $i \in \{1, ..., k\}$ such that $v \in A_i$. Hence, $\beta_v = \beta_w$ if and only if $A(v) = A(w)$. We also denote $\delta = \min_{e \in \mathcal{B}} |\beta_{e^+} - \beta_{e^-}|$ to be the minimal gap of the signal across elements of $\mathcal{A}$. We define the signal to noise ratio (SNR) as $\frac{\delta}{\sigma}$. Let $s = \text{sign}(D\beta)$ and $\mathcal{B} = \{e \in E : A(e^+) \neq A(e^-)\} = \text{support}(s)$. We also refer to the boundary of A as $\partial A = \{e : e \cap A \neq \emptyset \text{ and } e \cap V \backslash A \neq \emptyset\}$. Furthermore, we will define $-\mathcal{B} = E \backslash \mathcal{B}$ and $G_{-\mathcal{B}}$ denotes the graph with $V(G_{-\mathcal{B}}) = V$ and $E(G_{-\mathcal{B}}) = -\mathcal{B}$.

Remark 1.1. *Note that $\mathcal{A}$ comprises of the connected components of $G_{-\mathcal{B}}$ and $\mathcal{B}$ is the minimal set of edges with this property. In this way the set $\mathcal{B}$ induces $\mathcal{A}$, the structure of β.*

We will devote the remainder of the study to estimating $\mathcal{A}$, or equivalently estimating $\mathcal{B}$.

Definition 1.2. *Let $\hat{\beta}$ be an estimator of β. $\hat{\beta}$ is* ***sparsistent*** *if*

$$\lim_{n \to \infty} \mathbb{P}\{sign(\{D\hat{\beta}) = sign(D\beta)\} = 1$$

When there are tuning parameters to our estimators, we will say that the estimator is sparsistent if there exists any sequence of tuning parameters that give us sparsistency in accordance with the above definition.

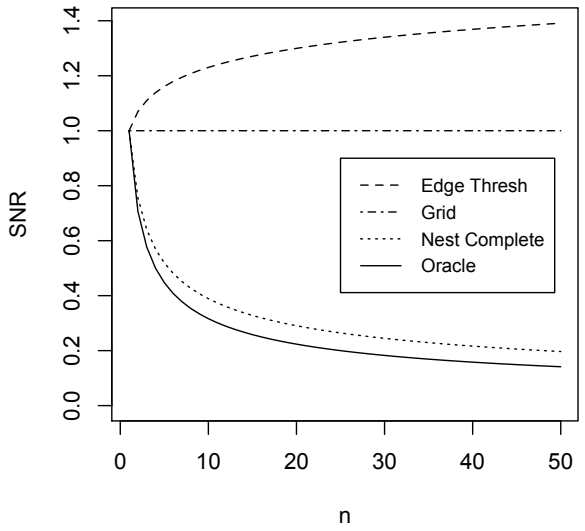

Figure 1: A qualitative summary of our sparsistency results for the SNR required by Edge thresholding, Edge lasso (for the 1-d, 2-d Grid and Nested Complete Graph), and an Oracle that has a priori knowledge of $\mathcal{A}$. (In the figure it is assumed that $|A|$ scales like n for all $A \in \mathcal{A}$.)

Our results provide conditions on the graph properties (motivated by algebraic graph theory) as well as the signal-to-noise ratio needed to ensure sparsistency for the edge lasso. We examplify our results using simple graph-structured patterns, and demonstrate that in some cases, such as the nested complete graph (defined later), edge lasso is sparsistent at very weak signal-to-noise ratios scaling as $\sqrt{(\log(n)/|A|}$, where $|A|$ denotes the smallest set of vertices with constant activation. This is close to the optimal performance expected since an oracle, that knows the partition, $\mathcal{A}$, a priori, will simply average the measurements in each $A \in \mathcal{A}$, thus boosting the SNR by $\sqrt{|A|}$. However, for the most common applications of the fused lasso, namely the 1 dimensional and 2 dimensional grids, edge lasso does not seem to perform much better than thresholding the difference of measurements at vertices which share an edge (which requires signal-to-noise ratio that scales as $\sqrt{\log n}$). In fact, we can show that edge lasso does not yield sparsistency for decreasing SNR for 1-d and 2-d grids. These findings are depicted in Figure 1.

1.1 Mathematical Preliminaries

We first introduce geometric properties of the incidence matrix, D. For a detailed exposition of these mathematical objects refer to [6]. For any oriented graph, the column space of D is $\text{row}(D^\top)$, and its orthogonal complement is $\text{null}(D^\top)$. 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 D, so by definition it forms a basis for $\text{row}(D^\top)$. Moreover $\chi(A) = \sum_{v \in A} D_v$ which we will use often. (We often subscript D with e and v interchangably 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 $\text{null}(D^\top)$ 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 $\mathcal{P}_{\text{null}(D^\top)}$ and $\mathcal{P}_{\text{row}(D^\top)}$.

Another common object of interest in algebraic graph theory is the unnormalized Laplacian matrix $L \in \mathbb{R}^{V \times V}$. If Δ is the diagonal degree matrix for the graph, and W its adjacency matrix, then $L = \Delta - W$. Also, let us denote the largest degree as $\Delta_{\max}$. Moreover, we see that the incidence matrix and the Laplacian are related by $L = D^\top D$. We also would like to note that the null space and row space $(\text{null}(D), \text{row}(D))$ is equal to the null and row space of L. The null space of D 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. Throughout this study we use Bachmann-Landau notation for asympotic statements. Namely, if $a_n / b_n \to 0$ then $a_n = o(b_n)$ and $b_n = \omega(a_n)$. To be clear we define $sign(0) = 0$. For a vector $z \in \mathbb{R}^E$ and a non-empty set of edges $\mathcal{B} \subset E$, we will denote with $z_{\mathcal{B}}$ the vector in $\mathbb{R}^E$ which agrees with z in the coordinates $\mathcal{B}$ and has zero entries in the coordinates in $\mathcal{B}^c$. Similarly, for a matrix $D \in \mathbb{R}^{V \times E}$, we will write $D_{\mathcal{B}}$ for the matrix D with the rows in $-\mathcal{B}$ replaced by zero vectors.

2 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 $\mathcal{B}$, and thus an estimate of $\mathcal{A}$. In this way, edge thresholding is meant to be the simplest estimator that might obtain sparsistency. Although we will be estimating $\mathcal{B}$ in this section notice that this is roughly equivalent to estimating β because we can average observations Y within connected components of $G_{-\mathcal{B}}$ to obtain a $\hat{\beta}$. The estimator is given by,

$$\hat{\mathcal{B}}_{th}(\tau) = \{e : |y_{e^+} - y_{e^-}| > \tau\}$$

We find that $\hat{\mathcal{B}}_{th}$ is the support of the solution to the dual problem,

$$\min_{z \in \mathbb{R}^E} \sum_{e \in E} (y_{e^+} - y_{e^-} - z_e)^2 + \lambda \|z\|_0$$

for $\lambda = \tau^2$. We now characterize necessary and sufficient conditions to obtain consistency of the estimator $\hat{\mathcal{B}}_{th}$, where consistency occurs if $\mathbb{P}\{\hat{\mathcal{B}}_{th} = \mathcal{B}\} \to 1$.

Theorem 2.1. *Suppose that* $\frac{|\mathcal{B}|}{|E|} \to 0$ *for simplicity.*

1. *If* $\frac{\delta}{\sigma} = \omega(\sqrt{\log |E|})$ *then* $\hat{\mathcal{B}}_{th}$ *is consistent for* $\mathcal{B}$.
2. *If* $\frac{\delta}{\sigma} = o(\sqrt{\log(|E| - |\mathcal{B}|)})$ *then* $\hat{\mathcal{B}}_{th}$ *is not consistent for* $\mathcal{B}$.

The proof of the theorem is given in the supplementary material.[10] We see immediately that the signal to noise ratio must be increasing like the log of the number of edges for $\hat{\mathcal{B}}_{th}$ to achieve consistency.

3 The Edge Lasso

In this section we will describe the edge lasso estimator, which arises as the solution to a generalized fused lasso problem as defined in [15] with the graph constraints specified by the matrix D. In particular, the edge lasso is the minimizer of the convex problem

$$\min_{\hat{\beta} \in \mathbb{R}^p} \frac{1}{2} \|y - \hat{\beta}\|_2^2 + \lambda \|D\hat{\beta}\|_1, \tag{1}$$

were $\lambda > 0$ is a tuning parameter. Thus, the edge lasso is the penalized least squares estimator of β with penalty term given by the ℓ_1 norm of the differences of measurements across edges in G. We denote the problem (1) the primal problem and its solution the *primal solution.* The fused primal problem directly estimates $\mathcal{A}$ through $\hat{\beta}$. We remark that the primal solution is *always* unique because the ℓ_2 loss is strictly convex and, just like edge thresholding, is invariant under changes in the orientation of the edges. Hence, the objective function in eq. (1) is strictly convex.

As shown in [15], the dual problem to (1) is given by

$$\min_{z\in\mathbb{R}^m} \frac{1}{2}\|y - \lambda D^\top z\|_2^2 \text{ such that } \|z\|_\infty \leq 1, \quad (2)$$

and any solution $\hat{z}$ to the dual problem results in the primal solution (see [15] for details)

$$\hat{\beta} = y - \lambda D^\top \hat{z} = \mathcal{P}_{\text{null}(D_{-\hat{\mathcal{B}}})}(y - \lambda D_{\hat{\mathcal{B}}}^\top \hat{z}_{\hat{\mathcal{B}}}). \quad (3)$$

where $\hat{\mathcal{B}} = \{e \in E : |\hat{z}_e| = 1\}$. 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 $\hat{\beta}$. [15]

When testing for sparsistency we use the following **primal dual witness** (PDW) construction, pioneered by [16], which results in a pair $(\hat{\beta}, \hat{z})$ 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 $\hat{z}_{\mathcal{B}} = \text{sign}(D\beta)$, which is equivalent to assuming the knowledge of both $\mathcal{B}$ and the sign differences. Using this knowledge, compute $\hat{\beta} = \mathcal{P}_{\text{null}(D_{-\mathcal{B}})}(y - \lambda D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}})$. The PDW steps are as follows.

1. Verify the complementary slackness condition $\text{sign}(D_{\mathcal{B}}\hat{\beta}) = \hat{z}_{\mathcal{B}}$.
2. Construct $\tilde{z}$ by solving the linear program

$$\min_{\tilde{z}\in Z} \|\tilde{z}\|_\infty \quad (4)$$

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

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

3. Construct the noisy dual by

$$\hat{z}_{-\mathcal{B}} = \frac{1}{\lambda} D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger \epsilon + \tilde{z}$$

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

Theorem 3.1. *If the PDW method passes for all large enough n then the solution to the dual program (2) is sparsistent.*

Before we can prove Theorem 3.1 we need the following lemma. This lemma will also be used when proving Proposition 5.5.

Lemma 3.2. *Suppose we are given the boundary set $\mathcal{B}$ with sign vector $\hat{z}_{\mathcal{B}} \in \{-1, 0, 1\}^E$ supported only over $\mathcal{B}$. Let $\hat{z}_{-\mathcal{B}}^\dagger = D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger (\frac{y}{\lambda} - D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}})$. Set $\hat{z}^\dagger = \hat{z}_{\mathcal{B}} + \hat{z}_{-\mathcal{B}}^\dagger$ and obtain the corresponding primal solution*

$$\hat{\beta} = y - \lambda D^\top \hat{z}^\dagger.$$

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

1. *Dual feasibility:* $\|z_{-\mathcal{B}}^\dagger + f\|_\infty \leq 1$
2. *Complementary slackness:* $\text{sign}(D_{\mathcal{B}}\hat{\beta}) \subseteq \hat{z}_{\mathcal{B}}$

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

Proof sketch. A detailed proof can be found in the supplementary material.[10] We first enumerate the KKT conditions and find that $\hat{z}^\dagger$ arises due to the zero-subgradient conditions leaving only the dual feasibility and complementary slackness to be satisfied. We have that for a subgradient γ of $\|z\|_\infty$ the KKT conditions are

1. (zero subgradient) $D(\lambda D^\top \hat{z} - y) + \gamma = 0$;
2. (dual feasibility) $\|\hat{z}\|_\infty \leq 1$;
3. (complementary slackness) $\gamma_i \geq 0$ if $\hat{z}_i = 1$, $\gamma_i \leq 0$ if $\hat{z}_i = -1$, and $\gamma = 0$ otherwise.

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

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

because over $-\mathcal{B}$, $\gamma_e = 0$. This is yields $\exists f \in \text{null}(D_{-\mathcal{B}}^\top), \hat{z}_{-\mathcal{B}} = \hat{z}_{-\mathcal{B}}^\dagger + f$. Using the SVD of $D_{-\mathcal{B}}$ we find that $(D_{-\mathcal{B}} D_{-\mathcal{B}}^\top)^\dagger D_{-\mathcal{B}} = D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger$. Furthermore defining $\gamma_{\mathcal{B}} = D_{\mathcal{B}}\hat{\beta}$ is necessary and sufficient for the remainder of the zero subgradient condition. Now the complementary slackness holds if and only if $\text{sign}(D_{\mathcal{B}}\hat{\beta}) \subseteq \hat{z}_{\mathcal{B}}$. □

Proof of Theorem 3.1. We will show that the conditions of Lemma 3.2 are satisfied if the PDW passes. By construction, if 1 passes then complementary slackness in Lemma 3.2 will be satisfied. By step 2 of the PDW construction, $D_{-\mathcal{B}}D_{-\mathcal{B}}^\top \tilde{z} = -D_{-\mathcal{B}}D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}}$ implies that $\hat{z} = -D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}} + f$, for some $f \in \text{null}(D_{-\mathcal{B}}^\top)$, again by the SVD decomposition of $L_{-\mathcal{B}}$. But we know that $L_{-\mathcal{B}}^\dagger \beta = 0$ because the Moore-Penrose pseudoinverse has zero action on any vectors that are constant over connected components of $G_{-\mathcal{B}}$. Therefore, $\frac{1}{\lambda}D_{-\mathcal{B}}^\top L_{-\mathcal{B}}^\dagger \epsilon = \frac{1}{\lambda}D_{-\mathcal{B}}^\top L_{-\mathcal{B}}^\dagger y$. Next, by step 3 we know that

$$\hat{z}_{-\mathcal{B}} = D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger(\frac{y}{\lambda} - D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}}) + f = \hat{z}_{-\mathcal{B}}^\dagger + f$$

If step 4 passes then dual feasibility holds. □

4 Noiseless Recovery

In this section we 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) For these graphs, consistent noisy recovery with the edge lasso is therefore hopeless. Secondly, and more importantly, as we will show later in Section 5, in order for noisy recovery to be possible at all, it is necessary that strict dual feasibility holds, i.e. $\|\hat{z}_{-\mathcal{B}}\|_\infty < c$ for some $c < 1$ for some dual solution $\hat{z}$ to the noiseless edge lasso problem.

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

$$\hat{\beta}_v - \beta_v = (\mathcal{P}_{\text{null}(D_{-\hat{\mathcal{B}}})}(\beta - \lambda D_{\hat{\mathcal{B}}}^\top \hat{z}_{\hat{\mathcal{B}}}))_v - \beta_v$$

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

Notice that $|\chi(\hat{A}(v))^\top \hat{z}_{\hat{\mathcal{B}}}| \leq |\partial \hat{A}(v)|$ because of the definition of the chararacteristic 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 4.1. *Let $\hat{z}$ be the result of the PDW method and notice that in the noiseless setting $\tilde{z} = \hat{z}_{-B}$. Then for some $A \in \mathcal{A}$ there exists a cut of A with shores $C, \bar{C}$ such that*

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

Proof. For the following theorem we will focus on a connected component A that contains an edge e such that $|\tilde{z}_e| = \|\tilde{z}\|_\infty$. Let $Q = \{e \in E(A) : |\tilde{z}_e| = \|\tilde{z}\|_\infty\}$ and denote $\zeta = \|\tilde{z}\|_\infty$. 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 $z' = \tilde{z} + \eta\chi(\phi)$ 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 $\tilde{z}$ with z'. We obtain a new edge vector that satisfies the zero subgradient condition because we only added elements of the $\text{null}(D_{-\mathcal{B}}^\top)$. Moreover $\|z'\|_\infty < \|\tilde{z}\|_\infty$, contradicting the fact that $\tilde{z}$ is the solution to eqn. (4) in PDW step (2). Hence, Q is a cut of A and for all $e \in Q$, $|\tilde{z}_e| = \|\tilde{z}\|_\infty$, and let one shore of the cut be C. Notice that $|Q| = |\partial C \cup \partial \bar{C}|$ and $\|\tilde{z}\|_\infty |\partial C \cup \partial \bar{C}| = |\chi(C)^\top \tilde{z}| = |\chi(C)^\top \hat{z}_{-B}|$. Now, Proposition 7.2 from the supplementary material [10] states that the zero subgradient condition is equivalent to,

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

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

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

□

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

Theorem 4.2. *Define the following notion of connectivity for each $A \in \mathcal{A}$,*

$$\rho(A) = \max_{C \subset A} \frac{|C|}{|\partial C \cap \partial \bar{C}|}\frac{|\partial \bar{C} \cap \partial A|}{|A|} \tag{5}$$

Then the result of the PDW method satisfies,

$$\|\hat{z}_{-\mathcal{B}}\|_\infty \leq 2 \max_{A \in \mathcal{A}} \rho(A)$$

Thus, the noiseless problem recovers the correct $\mathcal{A}$ if $\rho(\mathcal{A}) = \max_{A \in \mathcal{A}} \rho(A) < 1/2$.

Proof. Consider the C, A pair in the proof of Lemma 4.1. We need to show that, $\frac{1}{|\partial C \cap \partial \bar{C}|}|\frac{|C|}{|A|}\chi(\bar{C})^\top \hat{z} - \frac{|\bar{C}|}{|A|}\chi(C)^\top \hat{z}| \leq 2\rho(A)$. To this end, note that

$\frac{1}{|\partial C \cap \partial \bar{C}|}\left|\frac{|C|}{|A|}\chi(\bar{C})^\top \hat{z} - \frac{|\bar{C}|}{|A|}\chi(C)^\top \hat{z}\right|$ is smaller than

$$\frac{|\chi(\bar{C})^\top \hat{z}|}{|\partial C \cap \partial \bar{C}|}\frac{|C|}{|A|} + \frac{|\chi(C)^\top \hat{z}|}{|\partial C \cap \partial \bar{C}|}\left|\frac{|\bar{C}|}{|A|}\right.$$

which in turn is smaller than

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

□

See Figure 2 for an illustration of condition (5) in the previous theorem. Later we will describe a class of graphs which we call the nested complete graphs for which $\rho(A) = \frac{1}{|A|}$ for $A \in \mathcal{A}$.

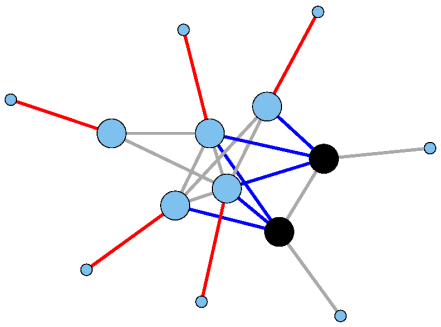

Figure 2: An example of the quantities in eq. (5) for a cut of set A depicted by the large vertices. The cut C are the black vertices, $\partial C \cap \partial \bar{C}$ are blue edges, and $\partial \bar{C} \cap \partial A$ are red edges. The RHS of eq. (5) for this cut is 5/21.

In our next result, we give necessary conditions for noiseless recovery, which require the $|\partial A|/|A|$ to be small enough and the out-degree to be greater than the in-degree for $v \in A$. Figure 3 shows an example of a graph for which this condition is violated and, therefore, noiseless recovery fails.

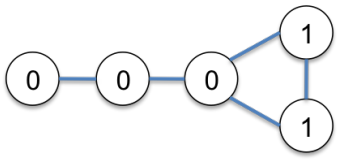

Figure 3: An example where for all $\lambda > 0$ the edge lasso does not recover the true $\mathcal{A}$. The 0 to the right is separated into its own element of the partition.

Corollary 4.3. *Consider a vertex $v \in A \in \mathcal{A}$. Let $\nu = |\{e \in -\mathcal{B} : v \in e\}|$ be the degree of v within $-\mathcal{B}$. If $\nu < |D_v^\top s| - \frac{|\partial A|}{|A|}$ then $\hat{\mathcal{B}} \neq \mathcal{B}$.*

Remark 4.4. *Under the conditions of Corollary 4.3, $\rho(A)$ is close to 1 for large $|A|$.*

Proof. First suppose $\hat{\mathcal{B}} = \mathcal{B}$ and $\tilde{z}_{\hat{\mathcal{B}}} = s$. Set $C = \{v\}$ in Proposition 7.2. Then $\nu \geq |D_v^\top \tilde{z}_{-\hat{\mathcal{B}}}|$ while by the proposition $|D_v^\top \tilde{z}_{-\hat{\mathcal{B}}}| = |\frac{1}{|A|}\chi(A)^\top \hat{z}_{\mathcal{B}} - D_v^\top \hat{z}_{\mathcal{B}}| \geq |D_v^\top s| - \frac{|\partial A|}{|A|}$. We immediately arrive at a contradiction: $\nu \geq |D_v^\top s| - \frac{|\partial A|}{|A|} > \nu$. □

5 Noisy Recovery

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

$$\frac{\delta}{\sigma} = \omega\left(\frac{1}{\sqrt{|A|}}\right) \text{ and } \lambda = o\left(\delta\frac{|A|}{|\partial A|}\right) \tag{6}$$

then step (1) of the PDW method passes with probability tending to 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 $D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger$. The intuition is that if we have some dual slack in the sense that $\|\tilde{z}\|_\infty$ is bounded away from 1 and if we bound the maximum of $|(D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger \epsilon)_e|$ then the PDW method will pass. We show that we can accurately describe $\|D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger\|_{2,\infty}$ 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 $\|D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger\|_{2,\infty}$.

Lemma 5.2. *Let $\tilde{z}$ be the result of step (2) of the PDW method. If $\|\tilde{z}_{-\mathcal{B}}\|_\infty < c$ for all large n for some $0 < c < 1$ and*

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

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

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

Theorem 5.3. *Suppose that the following conditions hold for all $A \in \mathcal{A}$,*

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

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

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

then the edge lasso is sparsistent.

Proof. Recall from Theorem 4.2 that $\|\tilde{z}\|_\infty \leq 2\max_{A\in\mathcal{A}}\rho(A)$ and so $\|\tilde{z}\|_\infty = o(1)$. The second condition implies the conditions of Lemma 5.2 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 n. The third condition completes the conditions of Lemma 5.1 and step (1) passes for large enough n. □

Thus far our conditions rely on the size of $\|D_{-\mathcal{B}}L^{\dagger}_{-B}\|_{2,\infty}$ 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. (The eigenvalues are denoted $\xi_v = \Xi_{v,v}$ and $\xi_v^\dagger$ is the pseudoinverse of the scalar - it is reciprocated if it is non-zero.)

Proposition 5.4. *Let the spectral decomposition of the Laplacian for $A \in \mathcal{A}$ be $L_A = U\Xi U^\top$ then $\|D_{-\mathcal{B}}L^{\dagger}_{-\mathcal{B}}\|_{2,\infty}$ is equal to*

$$\max_{A\in\mathcal{A}}\max_{e\in A}\sqrt{\sum_{v\in V}(U_{v,e^+}-U_{v,e^-})^2(\xi_v^2)^\dagger}$$

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

$$\text{and} \quad \|D_{-\mathcal{B}}L^{\dagger}_{-\mathcal{B}}\|_{2,\infty} \leq \max_{A\in\mathcal{A}}\sqrt{\sum_{v\in A}\eta_v^2(\xi_v^2)^\dagger}$$

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 5.5. *Suppose that $|\mathcal{A}|$ and that for some $A \in \mathcal{A}$ there exists $v \in A$ such that $|D^\top_{\mathcal{B},v}s| = |\{e \notin \mathcal{B} : v \in e\}|$. Let the maximum gap of β between elements of $\mathcal{A}$ be denoted $\delta_{\max}$. If $\frac{\delta_{\max}}{\sigma} = o(1)$ then edge lasso is not sparsistent.*

5.1 Examples

Our findings suggest that the 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 $\rho(\mathcal{A})$ to be large). 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 sparsistency, which is in sharp contrast to the performance of the oracle. (See Figure 1) We provide a topology called the nested complete graph that satisfies the sufficient conditions for sparsistency. These examples are meant to provide a blueprint for using the previous results to explore topologies for which the edge lasso is sparsistent.

5.2 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 $\{1,...,n\}$ 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 $\{1,...,n_0\}\times\{1,...,n_1\}$. 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 4)

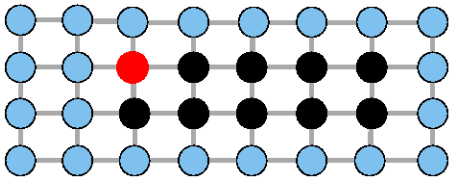

Figure 4: A 2D grid with $|\mathcal{A}| = 2$ depicted as union of black and red vertices. The red vertex is an example of a corner.

Corollary 5.6. *(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 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 sparsistent.

Proof. If the signal is non-trivial then there is a vertex $v \in A \in \mathcal{A}$ that is adjacent to ∂A. $|D_v^\top s| = 1$ which is the degree of v within A, so the conditions of Proposition 5.5 hold. The 2D case follows by considering the corner as v with $|D_v^\top s| = 2$. □

We see these typical mistakes in the 1D fused lasso in Figure 5. Here we observe a small incorrect element of $\hat{\mathcal{A}}$ at the boundary of a true element of $\mathcal{A}$. We also simulate (with 500 runs for each n) the tradeoff between the SNR and the probability of recovering $\mathcal{A}$ as

n increases. (Figure 6) The signal used in the simulations is a plateau where middle $n/3$ vertices have increased signal.

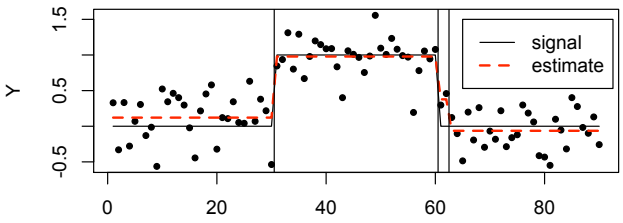

Figure 5: A typical mistake in the 1D fused lasso. The vertical lines indicate the beginning and end of estimated $\hat{\mathcal{A}}$.

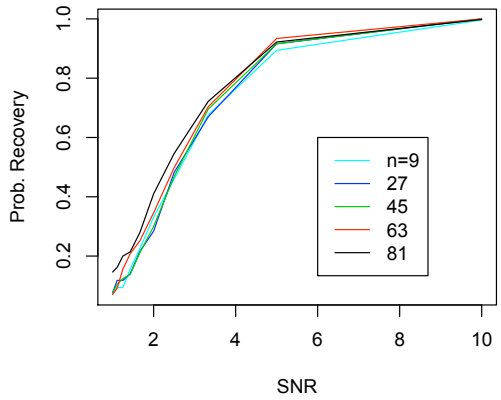

Figure 6: Simulated probability of recovery per SNR for a fused lasso for the plateau signal with $A \in \mathcal{A}$ of size $n/3$. There is little change in the tradeoff between SNR and probability of recovery as n increases.

5.2.1 Nested Complete Graph

We construct the nested complete graph from $p+1$ copies of the complete graph with p 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 7) We find that modulo factors that scale like the $\log n$, the sparsistency thresholds are the same as that of the oracle.

Corollary 5.7. *Suppose we construct the nested complete graph with p vertices in A and $p+1$ elements in the partition ($|A| = p$ and $|\mathcal{A}| = p+1$). If the SNR satisfies,*

$$\frac{\delta}{\sigma} = \omega(\frac{1}{\sqrt{p}}\sqrt{\log(p(p+1))})$$

Then the fused lasso is 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

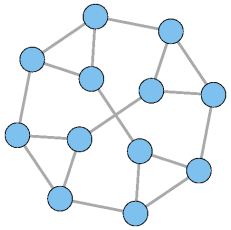

Figure 7: Nested complete graph with $p = 3$. $\mathcal{A}$ are the complete subgraphs of size 3.

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

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 5.4. 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 5.3 the result follows. □

6 Discussion

We have demonstrated that the performance of edge lasso depends critically on the structural properties of the graph. Edge lasso can achieve sparsistency for general graph structures under very weak signal-to-noise rations, however this happens under quite restrictive conditions of 5.3. For the 1D and 2D fused lasso, violating these conditions leads to inconsistency. Moreover, a typical case where we can demonstrate the conditions of Theorem 5.3 are satisfied (nested complete graph) is an example where we could have a priori identified the possible set $\mathcal{A}$ using graph cuts. In future work, we are investigating whether there are examples where edge lasso would dominate both edge thresholding and a priori graph cuts if exact recovery of $\mathcal{A}$ is desired. Another direction of work is to investigate whether *approximate* recovery guarantees can be made for edge lasso under measures such as the False Discovery Rate (FDR [5]).

Acknowledgements: This research is supported in part by AFOSR under grant FA9550-10-1-0382.

References

[1] L. Addario-Berry, N. Broutin, L. Devroye, , and G. Lugosi, *On combinatorial testing problems*, Annals of Statistics **38** (2010), 3063–3092.

[2] E. Arias-Castro, E. J. Candés, and A. Durand, *Detection of an anomalous cluster in a network*, Annals of Statistics **39** (2011), 278–304.

[3] E. Arias-Castro, E. J. Candés, H. Helgason, and O. Zeitouni, *Searching for a trail of evidence in a maze*, Annals of Statistics **36** (2007), 1726–1757.

[4] E. Arias-Castro, D. L. Donoho, and X. Huo, *Near-optimal detection of geometric objects by fast multiscale methods*, IEEE Transactions on Information Theory **51** (2005), no. 7, 2402–2425.

[5] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological) **57** (1995), no. 1, 289–300.

[6] C.D. Godsil and G. Royle, *Algebraic graph theory*, vol. 8, Springer New York, 2001.

[7] H. Hoefling, *A path algorithm for the fused lasso signal approximator*, Tech. report, October 2009.

[8] J. Liu, L. Yuan, and J. Ye, *An efficient algorithm for a class of fused lasso problems*, In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010.

[9] A. Rinaldo, *Properties and refinements of the fused lasso*, The Annals of Statistics **37** (2009), no. 5B, 2922–2952.

[10] J. Sharpnack, A. Rinaldo, and A. Singh, *Sparsistency of the edge lasso over graphs: Supplementary material*, `http://www.stat.cmu.edu/~jsharpna/ELsuppmat.pdf`, 2012.

[11] J. Sharpnack and A. Singh, *Identifying graph-structured activation patterns in networks*, Neural Information Processing Systems (NIPS), 2010.

[12] A. Singh, R. Nowak, and R. Calderbank, *Detecting weak but hierarchically-structured patterns in networks*, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 9, 2010, pp. 749–756.

[13] M. Talagrand, *The generic chaining*, Springer, 2005.

[14] 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** (2005), 91–108.

[15] Ryan J. Tibshirani and Jonathan Taylor, *The solution path of the generalized lasso*, (2010).

[16] M.J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using*, IEEE transactions on information theory **55** (2009), no. 5, 2183.

7 Supplementary Material

Throughout the supplementary material let $m = |E|$ and $n = |V|$.

Proof of thm 2.1. We will assume that the noise is Gaussian with variance 1, by making the gap $\delta' = \delta/\sigma$. Let us construct the statistics: $x_e = y_{e_h} - y_{e_t}$ for each edge e. Now we are interested in the estimator $\hat{\mathcal{B}} = \{e : |x_e| > \tau\}$. We can show (1) with Markov's inequality and Gaussian concentration,

$$\mathbb{P}\{\hat{\mathcal{B}} \neq \mathcal{B}\} = \mathbb{P}\{\inf_{e\in\mathcal{B}} |x_e| < \tau\} \cup \{\sup_{e\in -\mathcal{B}} |x_e| \geq \tau\} \quad (7)$$

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

The inequality works because within $\mathcal{B}$, x_e is normal with variance 2 and mean of magnitude at least δ'. Also, within $-\mathcal{B}$, x_e is normal with zero mean and variance 2. The RHS of (7) 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 $\tau = o(\mathbb{E}\sup_{e\in-\mathcal{B}} x_e)$ then certainly $\hat{\mathcal{B}}_{th}$ is not sparsistent. But certainly $\tau < \delta'$ is a necessary condition for sparsistency because there exists an $e \in \mathcal{B}$ such that $\mathbb{P}\{|x_e| \leq \delta'\} \geq C$ for some constant C. Thus, $\delta' = \Omega(\mathbb{E}\sup_{e\in-\mathcal{B}} x_e)$ is necessary for consistency.

Now we use the generic chaining methods developed by [13] to bound $\mathbb{E}\sup_{e\in-\mathcal{B}} x_e$ from below. Let $N_n = 2^{2^n}$ and define an admissible sequence to be a sequence of increasing partitions $\mathcal{H}_n$ such that $|\mathcal{H}_n| \leq N_n$. Let $\Delta(H_n(t))$ be the diameter of the cell containing t in the partition. Now define the following functional of metric space (T, d),

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

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

Theorem 7.1 ([13]). *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 x_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. □

Complete proof of Lemma 3.2. We will enumerate the KKT conditions and find that $\hat{z}^\dagger$ arises due to the zero-subgradient conditions leaving only the dual feasibility and complementary slackness to be satisfied. We introduce Lagrangian parameters γ_+, γ_- and use the following Lagrangian,

$$\frac{1}{2}\|y - \lambda D^\top \hat{z}\|_2^2 + \gamma_+^\top(\hat{z} - 1) + \gamma_-^\top(-\hat{z} - 1)$$

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

The following are the complete KKT conditions:

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

Now define $\gamma = \gamma_+ - \gamma_-$ the KKT conditions may be reduced to,

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

Notice that the existence of such a γ is necessary and sufficient for dual optimality due to convexity.

Consider the zero subgradient condition only over $-\hat{\mathcal{B}}$,

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

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

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

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

$$(D_{-\hat{\mathcal{B}}} D_{-\hat{\mathcal{B}}}^\top)^\dagger D_{-\hat{\mathcal{B}}} = U(\Lambda^\dagger)^2 U^\top U \Lambda V^\top$$
$$= U\Lambda^\dagger V^\top = U\Lambda V^\top V \Lambda^\dagger U^\top$$
$$= D_{-\hat{\mathcal{B}}} L_{-\hat{\mathcal{B}}}^\dagger$$

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

Proposition 7.2. *For dual solution $\hat{z}$ the zero-subgradient condition is satisfied if and only if for all $\hat{A} \in \hat{\mathcal{A}}$, and for all $C \subseteq \hat{A}$,*

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

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

$$D_{-\hat{\mathcal{B}},v}^\top \hat{z}_{-\hat{\mathcal{B}}} = \mathcal{P}_{\text{row}(D_{-\hat{\mathcal{B}}}^\top)}(-D_{\hat{\mathcal{B}},v}^\top \hat{z}_{\hat{\mathcal{B}}}) \quad (8)$$
$$= \tfrac{\chi(\hat{A})^\top \hat{z}_{\hat{\mathcal{B}}}}{|\hat{A}|} - D_{\hat{\mathcal{B}},v}^\top \hat{z}_{\hat{\mathcal{B}}} \quad (9)$$

Because $\hat{z}_{-\hat{\mathcal{B}}}$ is supported only over $-\hat{\mathcal{B}}$ we can rewrite $D_{-\hat{\mathcal{B}},v}^\top \hat{z}_{-\hat{\mathcal{B}}} = D_v^\top \hat{z}_{-\hat{\mathcal{B}}}$. Similarly, $D_{\hat{\mathcal{B}},v}^\top \hat{z}_{\hat{\mathcal{B}}} = D_v^\top \hat{z}_{\hat{\mathcal{B}}}$. Recall that $\sum_{v\in C} D_v^\top = \chi(C)^\top$ and the result follows by summing each side of eq. (8). The other direction follows immediately by setting $C = \{v\}$ and we have that $\chi(\{v\}) = D_v$. □

Proof of Lemma 5.1. In the PDW method we use $\hat{\mathcal{B}} = \mathcal{B}$ and $\hat{z}_{\mathcal{B}} = s$. Recall that the estimated signal is given by,

$$\hat{\beta}_v = (\mathcal{P}_{\text{null}(D_{-\mathcal{B}})}(y - \lambda D_{\mathcal{B}}^\top \hat{z}_{\mathcal{B}}))_v$$
$$= \beta_v + \frac{\sum_{w\in A(v)} \epsilon_w}{|A(v)|} - \lambda \frac{\chi(A(v))^\top \hat{z}_{\mathcal{B}}}{|A(v)|}$$

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

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

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

$$\frac{|\hat{\beta}_v - \hat{\beta}_w|}{\delta} \ge 1 - o(1)$$

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

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

The conditions of (6) imply that

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

□

Proof of Lemma 5.2. First we know that for each $e \in -\mathcal{B}$,

$$(D_{\mathcal{B}} L_{-\mathcal{B}}^\dagger \epsilon)_e \sim N(0, \sigma^2 \|D_e L_{-\mathcal{B}}^\dagger\|_2^2)$$

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

$$\|D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger \epsilon\|_\infty \le \sigma \|\!|D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger|\!\|_{2,\infty} \sqrt{2\log(\frac{|-\mathcal{B}|}{\gamma})}$$

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

$$\|\hat{z}\|_\infty \le \|\hat{z}\|_\infty + \frac{1}{\lambda}\|D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger \epsilon\|_\infty$$
$$\le \|\hat{z}\|_\infty + \frac{\sigma \|\!|D_{-\mathcal{B}} L_{-\mathcal{B}}^\dagger|\!\|_{2,\infty}}{\lambda}\sqrt{2\log(\frac{|-\mathcal{B}|}{\gamma})}$$

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

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

which occurs if

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

□

Proof of Proposition 5.4. Let $L_{-\mathcal{B}} = U\Xi U^\top$

$$\|\!|D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger|\!\|_{2,\infty} = \sup_{\|\delta\|_2\le 1} \|D_{-\mathcal{B}}U\xi^\dagger U^\top\delta\|_\infty$$

$$= \sup_{\|\delta\|_2\le 1} \|D_{-\mathcal{B}}U\xi^\dagger\delta\|_\infty = \sup_{\|\delta\|_2\le 1}\max_{e\in -\mathcal{B}} \|D_eU\xi^\dagger\delta\|_\infty$$

$$= \sup_{\|\delta\|_2\le 1}\max_{e\in -\mathcal{B}} |(U_{e_h}-U_{e_t})\xi^\dagger\delta|$$

$$= \max_{e\in -\mathcal{B}}\sup_{\|\delta\|_2\le 1} |(U_{e_h}-U_{e_t})\xi^\dagger\delta| = \max_{e\in -\mathcal{B}} \|(U_{e_h}-U_{e_t})\xi^\dagger\|_2$$

$$= \max_{e\in -\mathcal{B}} \sqrt{\sum_{v\in V}(U_{v,e_h}-U_{v,e_t})^2(\xi_v^2)^\dagger}$$

The above supremum is achieved for $\delta = \xi^\dagger(U_{e_h}-U_{e_t})^\top/\|(U_{e_h}-U_{e_t})\xi^\dagger\|_2$. Because $G_{-\mathcal{B}}$ is disconnected L_{-B} is block diagonal. Hence, 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

$$\|\!|D_{-\mathcal{B}}L_{-\mathcal{B}}^\dagger|\!\|_{2,\infty} = \max_{A\in\mathcal{A}}\max_{e\in A}\sqrt{\sum_{v\in V}(U_{v,e^+}-U_{v,e^-})^2(\xi_v^2)^\dagger}$$

Where the eigenvectors 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 $(U_{v,e^+}-U_{v,e^-})^2\le\eta^2$. But we have that η-Lipschitz in the shortest path distance occurs if and only if $\|DU_v\|_\infty\le\eta$. □

Proof of Proposition 5.5. Suppose that the solution to the primal program does recover $\mathcal{A}$ and s 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 $(\mathcal{P}_{\mathrm{null}(D_{-\hat{\mathcal{B}}})}(\beta-\lambda D_{\hat{\mathcal{B}}}^\top\hat z_{\hat{\mathcal{B}}}))_v$. So this mean is $\beta_v-\lambda\frac{\chi(A(v))^\top\hat z_{\hat{\mathcal{B}}}}{|A(v)|}$. So, if $\lambda\frac{\chi(A(v))^\top\hat z_{\hat{\mathcal{B}}}}{|A(v)|}>\delta_{\max}$ 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 3.2 implies

$$D_{-\hat{\mathcal{B}}}^\top\hat z_{-\hat{\mathcal{B}}} = D_{-\hat{\mathcal{B}}}^\top(D_{-\hat{\mathcal{B}}}D_{-\hat{\mathcal{B}}}^\top)^\dagger D_{-\hat{\mathcal{B}}}(\frac{y}{\lambda}-D_{\hat{\mathcal{B}}}^\top\hat z_{\hat{\mathcal{B}}})$$

$$= \mathcal{P}_{\mathrm{row}(D_{-\hat{\mathcal{B}}})}(\frac{y}{\lambda}-D_{\hat{\mathcal{B}}}^\top\hat z_{\hat{\mathcal{B}}})$$

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

$$D_{-\hat{\mathcal{B}},v}^\top\hat z_{-\hat{\mathcal{B}}} = -D_{\hat{\mathcal{B}},v}^\top\hat z_{\hat{\mathcal{B}}}+\frac{\chi(A(v))^\top\hat z_{\hat{\mathcal{B}}}}{|A(v)|}+$$

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

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 n 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 $|D_{-\hat{\mathcal{B}},v}^\top\hat z_{-\hat{\mathcal{B}}}|>|D_{\hat{\mathcal{B}},v}^\top\hat z_{\hat{\mathcal{B}}}|$. But by the condition there are not enough internal edges to dissipate $|D_{\hat{\mathcal{B}},v}^\top\hat z_{\hat{\mathcal{B}}}|$ without making $\|D_{-\hat{\mathcal{B}},v}^\top\hat z_{-\hat{\mathcal{B}}}\|_\infty>1$, and we arrive at our contradiction. □