Graph-based regularization for regression problems with highly-correlated designs

Yuan Li*, Benjamin Mark†, Garvesh Raskutti* and Rebecca Willett‡

Abstract:
Sparse models for high-dimensional linear regression and machine learning have received substantial attention over the past two decades. Model selection, or determining which features or covariates are the best explanatory variables, is critical to the interpretability of a learned model. Much of the current literature assumes that covariates are only mildly correlated. However, in modern applications ranging from functional MRI to genome-wide association studies, covariates are highly correlated and do not exhibit key properties (such as the restricted eigenvalue condition, RIP, or other related assumptions). This paper considers a high-dimensional regression setting in which a graph governs both correlations among the covariates and the similarity among regression coefficients. Using side information about the strength of correlations among features, we form a graph with edge weights corresponding to pairwise covariances. This graph is used to define a graph total variation regularizer that promotes similar weights for highly correlated features. The graph structure encapsulated by this regularizer helps precondition correlated features to yield provably accurate estimates. Graph-based similarity measures have led to successful regularization in the cases of the fused LASSO, edge LASSO, and graph trend filtering; however, using graph-based regularizers to develop theoretical guarantees for highly-correlated covariates has not been previously examined. This paper shows how our proposed graph-based regularization yields mean-squared error guarantees for a broad range of covariance graph structures and correlation strengths which in many cases are optimal by imposing additional structure on $\beta^\ast$ which encourages alignment with the covariance graph. Our proposed approach outperforms other state-of-the-art methods for highly-correlated design in a variety of experiments on simulated and real fMRI data.

1. Introduction

High-dimensional linear regression and inverse problems have received substantial attention over the past two decades (cf., Hastie, Tibshirani, and Wainwright (2015) for an overview). While there has been considerable theoretical and methodological development, applying these methods in real-world settings is more nuanced since variables or features are often highly correlated, while much of the existing theory and methodology is applicable when features are independent or satisfy weak correlation assumptions such as the restricted eigenvalue and other related conditions (cf., Bickel, Ritov, and Tsybakov (2009); van de Geer and Buhlmann (2009)). In this paper we develop an approach for parameter estimation in high-dimensional linear regression with highly-correlated designs.

More specifically, we consider observations of the form

$$y = X\beta^\ast + \epsilon \quad (1)$$

where $y \in \mathbb{R}^n$ is the response variable, $X \in \mathbb{R}^{n \times p}$ is the observation or design matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$ is Gaussian noise. Our goal is to estimate $\beta^\ast$ based on $(X, y)$ when $X$ potentially has highly-correlated columns and does not satisfy standard regularity assumptions. We consider the Gaussian linear model since it is more amenable to analysis but believe the ideas and results can be extended.

*Department of Statistics, University of Wisconsin-Madison. Raskutti and Li were supported by NSF DMS 1407028.
†Department of Mathematics, University of Wisconsin-Madison.
‡Department of Electrical and Computer Engineering, University of Wisconsin-Madison. Willett was supported by NIH Award 1 U54 AI117924-01 and NSF Award 1740707.
Highly-correlated or dependent features arise in many modern scientific problems. Consider, for instance, neural decoding using functional magnetic resonance imaging (fMRI) data. In this setting, we collect $n$ different scans of a patient and let $y_i$ indicate which stimulus was presented to the patient during the $i$th scan for $i = 1, \ldots, n$. The scan data is stored in $X$, with each row of $X$ corresponding to a different fMRI scan and each column corresponding to a different voxel (small localized region of the brain). Each design matrix element $X_{i,j}$ would represent a statistic of the blood-oxygen-level dependent (BOLD) signal in the $j$th voxel during the $i$th scan (cf., Michel, Gramfort, Varoquaux, Eger, and Thirion (2011)). The model in (1) reflects that information about the stimulus is represented (or encoded) by voxels in the brain corresponding to the support of $\beta^\star$. Because of the cost of fMRI experiments, it is typical to see on the order of one hundred scans and on the order of ten thousand voxels, resulting in a system that is under-determined by a factor of one hundred. In Section 5, we address a specific example involving fMRI. Similar challenges arise in genomics, such as genome-wide association studies (GWAS), where each column of $X$ corresponds to a different single nucleotide polymorphism (SNP) and each row corresponds to a different subject’s particular SNP realizations. In this application, the response variable $y$ reflects each patient’s disease state, and the goal is to determine which SNPs are potentially linked with a given disease (cf., Wu, Chen, Hastie, Sobel, and Lange (2009)). Again, the resulting system is typically underdetermined by a factor of about one hundred. In both settings, highly-correlated columns are unavoidable.

The key challenge associated with highly-correlated columns is that estimates of $\beta^\star$ become very sensitive to noise and, if columns are perfectly correlated, $\beta^\star$ may not be identifiable. For instance, if two voxels in the brain respond similarly to a stimulus in repeated trials, it becomes impossible to reliably determine which of the two voxels is potentially encoding information about the stimulus or if both are. Applying the classical LASSO (Tibshirani, 1996) estimator to such data would result in selecting the one voxel that is marginally more aligned with the stimulus. (That is, letting $X_j$ denote the $j$th column of $X$, if $X_j$ is highly correlated with $X_k$, then the LASSO would choose $\beta_j$ to be nonzero and $\beta_k$ to be zero valued when $\langle X_j, y \rangle > \langle X_k, y \rangle$; this test has low power and is sensitive to noise.)

1.1. Problem formulation and proposed estimator

First we define our model based on the standard linear model where data $(X^{(i)}, y^{(i)})_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$ are drawn i.i.d. according to

$$y^{(i)} = X^{(i)\top} \beta^\star + \epsilon^{(i)}, \text{ where } X^{(i)} \sim \mathcal{N}(0, \Sigma_{p \times p}) \text{ and } \epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2).$$

Let $y = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})^\top \in \mathbb{R}^n$, $X = [X^{(1)}, X^{(2)}, \ldots, X^{(n)}]^\top \in \mathbb{R}^{n \times p}$ and $\epsilon = (\epsilon^{(1)}, \epsilon^{(2)}, \ldots, \epsilon^{(n)})^\top \in \mathbb{R}^n$. Hence the linear model can be expressed in the standard matrix-vector form:

$$y = X \beta^\star + \epsilon.$$

Our goal is to estimates $\beta^\star$.

We assume $\Sigma$ is unknown and is estimated using either $X$ or side information; let $\hat{\Sigma}$ denote the estimate of the covariance matrix. Define $\hat{s}_{j,k} := \text{sign}(\hat{\Sigma}_{j,k})$. Based on the estimated covariance matrix $\hat{\Sigma}$, we consider the following estimator for $\beta^\star$:

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \|y - X \beta\|_2^2 + \lambda S \sum_{j,k} |\hat{\Sigma}_{j,k}|(\beta_j - \hat{s}_{j,k} \beta_k)^2 + \lambda_1 (\lambda_{TV} \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k} \beta_k| + \|\beta\|_1),$$

(2)

2
where \( \lambda_S, \lambda_1 \) and \( \lambda_{TV} \) are regularization parameters.

This estimator can be interpreted from a graph/network perspective by defining the covariance graph based on the covariance matrix \( \hat{\Sigma} \). Let \( G = (V, E, W) \) be an undirected weighted graph where \( V = \{1, 2, \ldots, p\} \) with edge weight \( w_{j,k} \) \((1 \leq j \neq k \leq p)\) associated with edge \((j, k) \in E\). The edge weights corresponding to \( W = (w_{j,k}) \) may be negative. Now we define our covariance graph. Let \( w_{j,k} = \hat{\Sigma}_{j,k} \), which denotes the \((j, k)\) entry of the covariance matrix \( \hat{\Sigma} \). Then \( E := \{(j, k) : w_{j,k} \neq 0, j \neq k\} \) and the entries of the weight matrix \( W \in \mathbb{R}^{p \times p} \) are \( W_{j,k} = w_{j,k} \). Given this graph, the regularization term \( \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k} \beta_k| \) is a measure of the graph total variation of the signal \( \beta \) with respect to the graph \( G \) and \( \sum_{j,k} (\beta_j - \hat{s}_{j,k} \beta_k)^2 \) corresponds to a graph Laplacian regularizer with respect to \( G \).

Further let \( \Gamma \) be the weighted edge incidence matrix associated with the graph \( G \). Specifically, we denote the set of edges in our graph as \((j_\ell, k_\ell)\) for \( \ell = 1, \ldots, m \) where \( m := |E| \) is the size of the edge set. Let

\[
\Gamma = \sum_{\ell=1}^m \Gamma_\ell, \quad \text{where} \quad \Gamma_\ell := |\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} u_\ell \left[ e_{j_\ell} - \text{sign}(\hat{\Sigma}_{j_\ell, k_\ell}) e_{k_\ell} \right]^\top \in \mathbb{R}^{m \times p},
\]

where \( u_\ell \in \mathbb{R}^m \) and \( e_\ell \in \mathbb{R}^p \) are the \( \ell \)-th canonical basis vectors (all zeros except for a one in the \( \ell \)-th element). Then the \( \ell \)-th row of \( \Gamma \) is

\[
|\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} \left[ e_{j_\ell} - \text{sign}(\hat{\Sigma}_{j_\ell, k_\ell}) e_{k_\ell} \right]^\top.
\]

Next suppose \( \lambda_1 > 0 \) and \( \lambda_{TV}, \lambda_S \geq 0 \). We define

\[
\tilde{X} = \tilde{X}_{\lambda_S} := \left[ \frac{X \sqrt{n \lambda_S \Gamma}}{\lambda_{TV} \Gamma} \right] \in \mathbb{R}^{(n+m) \times p}, \quad \tilde{y} := \left[ \frac{y}{0_{m \times 1}} \right] \in \mathbb{R}^{n+m}, \quad \text{and} \quad \tilde{\Gamma} := \left[ \frac{\lambda_{TV} \Gamma}{I_p \times p} \right] \in \mathbb{R}^{(m+p) \times p}.
\]

Using these definitions, we may write the estimator (2) equivalently as

\[
\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \left\| y - \tilde{X} \beta \right\|_2^2 + \lambda_S \left\| \Gamma \beta \right\|_2^2 + \lambda_1 \left( \lambda_{TV} \left\| \tilde{\Gamma} \beta \right\|_1 + \left\| \beta \right\|_1 \right)
\]

\[
= \arg \min_{\beta} \frac{1}{n} \left\| \tilde{y} - \tilde{X} \beta \right\|_2^2 + \lambda_1 \left\| \tilde{\Gamma} \beta \right\|_1.
\]

The three regularizers play the following roles:

- We refer to \( \left\| \Gamma \beta \right\|_1 = \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k} \beta_k| \) as the Laplacian smoothing penalty; Hebiri and van de Geer (2011) studied a variant of this regularizer with \( \hat{\Sigma}_{j,k} \) replaced with arbitrary non-negative weights. Because each term is squared, it helps to reduce the ill-conditionedness of \( X \) when columns are highly correlated, as reflected in our analysis.

- We refer to \( \left\| \Gamma \beta \right\|_1 = \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k} \beta_k| \) as the total variation penalty, as do Shuman, Narang, Frossard, Ortega, and Vandergheynst (2013); Wang, Sharpnack, Smola, and Tibshirani (2016); Sadhanala, Wang, and Tibshirani (2016); H"utter and Rigollet (2016); it is closely related to the edge LASSO penalty (Sharpnack, Singh, and Rinaldo, 2012). Note that these prior works consider general weighted graphs (instead of graphs defined by a covariance matrix \( \hat{\Sigma} \), as we do). This regularizer promotes estimates \( \hat{\beta} \) that are well-aligned with the graph structure; for instance, a group of nodes with large edge weights connecting them (i.e., a group of columns of \( X \) that are highly correlated) are more likely to be associated with coefficient estimates with similar values.

- We refer to \( \left\| \beta \right\|_1 \) as the sparsity regularizer. The combination of the sparsity regularizer and total variation penalty amount to the fused LASSO (Tibshirani, Saunders, Rosset, Zhu, and Knight, 2005; Tibshirani and Taylor, 2011).

The combined effect of the three regularization terms is to find estimates of \( \beta^* \) which are both a good fit to the data when the columns of \( X \) are highly correlated and well-aligned with the underlying graph.
1.2. Contributions

This paper addresses the question of how to estimate $\beta^*$ from observations in (1) when $X$ has highly-correlated columns. We propose a regularized regression approach in which the regularization function depends upon the covariance of $X$. For a fixed graph $G$, the proposed estimator is closely related to the previously-proposed fused LASSO (Tibshirani et al., 2005), generalized LASSO (Tibshirani and Taylor, 2011), edge LASSO (Sharpmack et al., 2012), network LASSO (Hallac, Leskovec, and Boyd, 2015), trend filtering (Wang et al., 2016), and total-variation regularization (Shuman et al., 2013; Hütter and Rigollet, 2016). In contrast to these past efforts, our focus is on settings in which columns of $X$ are highly correlated and these correlations inform the choice of graph $G$.

On the other hand there is a large body of work on highly dependent features; in Section 2 we provide a thorough comparison of our method with other related approaches. In this paper we make the following contributions:

- A novel estimator with corresponding finite-sample theoretical guarantees for highly-correlated design matrices $X$. General theoretical guarantees provide insight into the impact of the alignment of $\beta^*$ with the covariance graph, and properties of the covariance graph structure such as smallest and largest block-sizes and smallest non-zero eigenvalue.
- New guarantees for four specific covariance graph structures, a block complete graph, a chain graph, a lattice graph and a random d-regular graph. Our mean-squared error bounds match the optimal rates in the independent case, and also match previous error bounds (Figueiredo and Nowak, 2016) for the block complete covariance graph.
- A simulation study which shows that our method out-performs state-of-the-art alternatives such as the Cluster Representative LASSO (CRL, Bühlmann, Rütimann, van de Geer, and Zhang (2013)) and Ordered Weighted LASSO (OWL, Bogdan, van den Berg, Su, and Candès (2013)) in terms of mean squared error.
- A validation of our method on real fMRI data that demonstrates that our method out-performs state-of-the-art methods in terms of prediction error.

The remainder of this paper is organized as follows: In Section 2 we discuss existing work and results for this problem and its relationship to our estimator; in Section 3 we present our main theoretical results; in Section 4 we carry out a simulation study by comparing our methods to other state-of-the-art methods; in Section 5 we apply our method to a real fMRI dataset with comparison to other methods; proofs are provided in Section 7; we state our conclusions in Section 6.

2. Relationship to prior work

As mentioned earlier, variants of the proposed estimator have been widely studied in past works. We review these past works below.

2.1. Prior work: “Denoising” settings

Significant effort has been devoted to understanding estimators like (4) in the special case where $X = I$ – that is, in a “denoising” setting in which observations are direct measurements of the signal of interest, $\beta$. Variants of these estimators are often referred to as the edge or network LASSO (Sharpmack et al., 2012; Hallac et al., 2015), a special case of graph trend filtering (Wang et al., 2016), or graph total variation estimation (Shuman et al., 2013). In these settings, we assume a known graph or network $G = (V, E)$, where each edge connecting notes $v_j \in V$ and $v_k \in V$ is associated with a non-negative edge weight $w_{j,k}$. 
Zero-valued edge weights are equivalent to the absence of an edge. The edge LASSO estimator or graph trend filter has the form
\[
\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \| y - \beta \|^2_2 + \lambda_{TV} \sum_{1 \leq j < k \leq n} w_{j,k} | \beta_j - \beta_k |.
\]

In particular, Sharpnack et al. (2012) assume a known underlying graph and that the true generating parameter \( \beta^* \) is piece-wise constant over an unknown partition of the vertices and study the “sparsistency” of the edge LASSO estimator – essentially, the conditions under which the edge LASSO estimate and the generating \( \beta^* \) are supported on the same graph partition. Hallac et al. (2015) describe a “network LASSO” problem and propose a scalable Alternating Direction Method of Multipliers (ADMM) algorithm that can be implemented on distributed computers and guarantees global convergence even on large graphs.

The above ideas are also closely related to work on “graph signal processing” (Shuman et al. (2013) provide an excellent overview), in which classical concepts of frequency domain representations of signals, filtering, translation, and modulation are all generalized to signals observed on the vertices of graphs. In particular, the regularization function \( \sum_{1 \leq j < k \leq n} w_{j,k} | \beta_j - \beta_k | \) is referred to as the total variation of the signal with respect to the graph.

Wang et al. (2016) consider a generalization of graph total variation to higher-order measures of variation of signals for denoising piecewise-polynomial signals on graphs and derive squared error bounds for the estimates. Those bounds do not account for a sensing matrix \( X \) other than \( X = I \). Hüter and Rigollet (2016) also develop sharp oracle inequalities for the edge LASSO, with an emphasis on a 2d lattice graph used in image processing applications.

2.2. Prior work: Regression/inverse problem settings

The original work on the fused LASSO (Tibshirani et al., 2005) considers estimators of the form
\[
\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \| y - X \beta \|^2_2 + \lambda_{TV} \sum_{j=1}^{p-1} | \beta_j - \beta_{j+1} | + \lambda_{1} \| \beta \|_1.
\]

Their theory yielded asymptotic properties for fixed \( p \) and \( n \to \infty \). The generalized LASSO of Tibshirani and Taylor (2011); Liu, Yuan, and Ye (2013) consider the estimators of the form
\[
\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \| y - X \beta \|^2_2 + \lambda \| \Gamma \beta \|_1 \quad (6)
\]

for general \( X \) and \( \Gamma \); note that both the fused LASSO and the estimator in (4) can be written in this form. Tibshirani and Taylor (2011) consider the solution path of this estimator, a variety of optimization and computational considerations, and a degrees-of-freedom analysis. Understanding the degrees-of-freedom of an estimator is a crucial step towards sample complexity bounds emphasized in this paper, but the past work does not consider the role of correlations among columns of \( X \) relative to the choice of regularizer and the subsequent impact on performance. In particular, their analysis focuses on the setting in which \( \text{rank}(X) = p \), a condition we do not assume in our setting. Related work by Needell and Ward (2013b,a) consider the special case of the generalized LASSO of total variation regularization on a grid for image reconstruction problems. That analysis, while elegant, relies heavily upon the grid-like graph structure associated with pixels in images and does not generalize to the setting of this paper. In addition, they assume that the design matrix \( X \) satisfies the restricted isometry property (Candes and Tao, 2005), a condition similar to
the restricted eigenvalue condition described below that is generally not met when columns of \( X \) are highly correlated.

Estimators similar to (4) have been explored empirically in a variety of application settings; (cf., Grosenick, Klingenberg, Greer, Taylor, and Knutson (2009); Michel et al. (2011); Grosenick, Klingenberg, Knutson, and Taylor (2012); Shen, Huang, and Pan (2012); Xin, Kawahara, Wang, and Gao (2014); Xin, Kawahara, Wang, Hu, and Gao (2016); Jalali and Fazel (2013)). However, these studies are mainly empirical and do not provide performance guarantees that are the focus of this paper.

2.3. Prior work: Highly correlated design matrices

A key focus of our work is the setting in which columns of \( X \) may be highly correlated. Correlated columns arise frequently in applications, as discussed in the introduction. There are several approaches developed to deal with the high-dimensional linear regression problem with some highly correlated covariates. The Elastic Net estimator proposed by Zou and Hastie (2005) is

\[
\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_S \|\beta\|_2^2,
\]

which encourages a grouping effect, in which strongly-correlated predictors tend to be in or out of the support of the estimate together. Witten, Shojaie, and Zhang (2014) propose a Cluster Elastic Net estimator:

\[
\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_S \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \|X_i\beta_i - X_j\beta_j\|_2^2,
\]

where \( C_1, \ldots, C_k \) denotes a fixed partition of the \( p \) features into \( K \) groups. This estimator incorporates clustering information inferred from data to perform more accurate regression. The Elastic Corr-net proposed by El Anbari and Mkhadri (2014) proposes combining an \( l_1 \) penalty with a correlation based quadratic penalty from Tutz and Ulbricht (2009).

An alternative approach explored by Bühlmann et al. (2013), called Cluster Representative LASSO (CRL), clusters highly correlated columns of \( X \), chooses a single representative for each cluster, and regresses over the cluster centers. In our fMRI example from above, the two correlated voxels would be clustered together, and the resulting CRL estimate would indicate simply that one or more members of the clusters are good predictors of the stimulus. Bühlmann et al. (2013) also considered a Cluster Group LASSO (CGL) in which a group sparsity regularizer was used with the original design matrix \( X \) and the group structure was determined by a clustering of the columns of \( X \). These two-stage approaches (first cluster, then regress based on estimated clusters) admitted encouraging statistical guarantees and empirical performance. However, (i) they depend heavily upon our ability to find a good clustering of the columns of \( X \), where clusters must be disjoint or non-overlapping; (ii) clustering decisions are “hard” and do not reflect varying degrees of correlation among columns, and (iii) clusters are formed independently of the observed responses \( y \). Constraint (i) is particularly challenging to meet in real-world applications, including our motivating application of fMRI. If we wish to force non-overlapping clusters, then there is potentially more than one “good” candidate clustering of the data; in this case, \( y \) can potentially provide valuable insight into which candidate clustering is most relevant to the regression task, but the CRL and CGL estimators do not leverage \( y \) in the clustering phase and it is not immediately clear how one would make this extension. Grouping pursuit (Shen and Huang, 2010) explores clustering columns of \( X \) while leveraging \( y \) by using a non-convex variant of the fused LASSO.
Other approaches sidestep explicit estimation of correlations among columns of $X$. Early work on the adaptive LASSO by Zou (2006) illustrated the impact of adaptivity in the correlated design setting. Recent work on the Ordered Weighted LASSO (OWL) estimator (Bogdan et al., 2013) proposed an alternative weighted LASSO regularizer in which the weights depend on the order statistics of $\beta$; specifically,

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|^2_2 + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j|,$$

where $w_1 \geq w_2 \geq \ldots \geq w_p \geq 0$ and $|\beta_j|$ is the $j$th largest element in $\{||\beta_1||, ||\beta_2||, \ldots, ||\beta_p||\}$, their paper shows that this family of regularizers can be used for sparse linear regression with strongly correlated covariates.

A special case of OWL is the OSCAR estimator (Bondell and Reich, 2008), which takes the form

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|^2_2 + \lambda_{\text{TV}} \sum_{1 \leq j < k \leq p} |\beta_j| \vee |\beta_k| + \lambda_1 ||\beta||_1.$$

Figueiredo and Nowak (2016) demonstrated that when two columns of $X$ were identical, then OWL would assign the corresponding elements of $\beta$ equal values. OWL adaptively groups highly correlated columns of $X$ by assigning them equal weights whenever their correlation exceeds a critical value – the grouping does not need to be pre-computed and will depend on the value of $y$. However, we still do not fully understand the performance of the OWL for design matrices with more moderately correlated columns. Finally, unlike the CRL, CGL, and the proposed GTV estimator (4), the OWL has no mechanism for incorporating possible side information about the correlation among different columns of $X$.

A related approach is the clustered LASSO (She, 2010), which takes the form

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|^2_2 + \lambda_{\text{TV}} \sum_{1 \leq j < k \leq p} |\beta_j - \beta_k| + \lambda_1 ||\beta||_1.$$

In contrast to the fused LASSO, the clustered LASSO considers all pairwise differences of elements of $\beta$. She (2010) conducts a classical asymptotic analysis (fixed $p$ and $n \to \infty$) of the clustered LASSO and its generalization (6) and establishes consistency results that depend upon $\Sigma^{-1}$; note that $\Sigma$ will have very small singular values when columns of $X$ are highly correlated. A similar estimator called Pairwise Absolute Clustering and Sparsity (PACS) estimator is proposed by Sharma, Bondell, and Zhang (2013):

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|^2_2 + \lambda_1 \left\{ \sum_{j=1}^{p} w_j |\beta_j| + \sum_{1 \leq j < k \leq p} w_{jk(-)} |\beta_j - \beta_k| + \sum_{1 \leq j < k \leq p} w_{jk(+)} |\beta_k + \beta_j| \right\},$$

where the $w_j$s, $w_{jk(-)}$s and $w_{jk(+)}$s are non-negative weights chosen in penalization scheme. This PACS estimator depends heavily on the choice of non-negative weights. One suggestion offered by Sharma et al. (2013) in the setting of correlated variables is $w_j = 1$, $w_{jk(-)} = \sqrt{2(1 - r_{jk})}$ and $w_{jk(+)} = \sqrt{2(1 + r_{jk})}$ for $1 \leq j < k \leq p$, where $r_{jk}$ is the correlation between $(j, k)$th pair of covariates. Hebiri and van de Geer (2011) consider smooth S-LASSO estimators of the form

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|^2_2 + \lambda_S ||\Gamma\beta||^2_2 + \lambda_1 ||\beta||_1.$$

The first regularization term, unlike the total variation term in (4), is a quadratic penalty similar to what appears in the elastic net (7) (Zou and Hastie, 2005). The analyses by She (2010), Sharma et al. (2013) and
Hebiri and van de Geer (2011) do not consider settings in which $X$ and $\Gamma$ in (6) are related. Daye and Jeng (2009) propose a weighted fusion estimator,

$$
\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \frac{\lambda S}{p} \sum_{1 \leq j < k \leq p} w_{jk}(\beta_j - s_{jk}\beta_k)^2 + \lambda_1\|\beta\|_1,
$$

where $s_{jk} = \text{sign}(X_j^T X_k)$ and $w_{jk} \propto |X_j^T X_k|$; this approach is similar to Hebiri and van de Geer (2011) for a particular choice of weights. Daye and Jeng (2009) focus their analysis on grouping effects, sign consistency, and limiting distributions, but do not consider finite sample error bounds of the type developed in this paper. The Sparse Laplacian Shrinkage (SLS) estimator proposed by Huang, Ma, Li, and Zhang (2011) uses a minimum concave penalty (MCP) to replace the LASSO penalty in a weighted fusion estimator to reduce bias; specifically, the SLS estimator has the following form:

$$
\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \|y - X\beta\|_2^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, \gamma) + \lambda S \sum_{1 \leq i < j \leq p} w_{jk}(\beta_j - s_{jk}\beta_k)^2,
$$

where $\rho$ is the MCP with penalty parameter $\lambda_1$ and regularization parameter $\gamma$, the definitions of $w_{jk}$ and $s_{jk}$ are the same with weighted fusion estimator. Huang et al. (2011) provide analysis for sign consistency and selection properties of SLS estimator, but no finite sample error results are provided.

3. Assumptions and Main Results

We first introduce a set of assumptions needed for our main results. Throughout we use the induced matrix norm notation

$$
\|A\|_{p,q} = \sup_{x \neq 0} \frac{\|Ax\|_q}{\|x\|_p}.
$$

Specifically, note that $\|A\|_{1,2}$ is the maximum column norm of $A$ and $\|A\|_{op} = \|A\|_{2,2}$. For a symmetric positive semi-definite matrix $A$, let $\lambda_{\min}(A)$ denote its minimum eigenvalue and $\lambda_{\max}(A)$ denote its maximum eigenvalue.

The notation $X_n = O_P(a_n)$ means that the set of values $\frac{X_n}{a_n}$ is stochastically bounded. That is, for any $\epsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that

$$
P\left( \left| \frac{X_n}{a_n} \right| > M \right) < \epsilon, \forall n > N.
$$

**Assumption 3.1.** We assume that there exists an absolute constant $c_u > 0$ such that

$$
\lambda_{\max}(\Sigma) \leq c_u.
$$

**Remark 3.1.** This statement assumes that $\Sigma$ is normalized such that the largest eigenvalue of $\Sigma$ can be upper bounded by a positive constant.

**Assumption 3.2.** There exists an absolute constant $c_\ell > 0$ such that:

$$
c_\ell \leq \min_{1 \leq j \leq p} \sum_{k=1}^{p} |\Sigma_{j,k}|.
$$

**Remark 3.2.** Assumption 3.2 ensures the $\ell_1$ norm for each row/column is lower bounded by a constant. This assumption is much milder than assuming the minimum eigenvalue of $\Sigma$ is bounded away from 0. As an example, Assumption 3.2 is satisfied when every diagonal entry of $\Sigma$ is bounded below by $c_\ell$. 

8
**Assumption 3.3.** The estimated covariance matrix \( \hat{\Sigma} \) that is used to construct the matrix \( \Gamma \) satisfies

\[
\| \hat{\Sigma} - \Sigma \|_{1,1} = \max_{1 \leq j \leq p} \sum_{k=1}^{p} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}| \leq \frac{c_\ell \ell}{4},
\]

where \( c_\ell \) is defined in Assumption 3.2.

**Remark 3.3.** Assumption 3.3 states that we need a sufficiently accurate estimator \( \hat{\Sigma} \) for \( \Sigma \). If Assumption 3.3 is satisfied then we can use \( \hat{\Sigma} \) to construct \( \Gamma \) for our optimization problem stated in (5). In some cases, we estimate \( \Sigma \) using side information that is not based on \((X(i))_{i=1}^{n}\). If we need to estimate \( \Sigma \) based on \((X(i))_{i=1}^{n}\), there is a large literature on high-dimensional covariance estimation in high dimensions under different structural assumptions (cf., Bickel and Levina (2008b,a); Cai and Liu (2011); Cai, Zhao, and Zhou (2016); Donoho, Gavish, and Johnstone (2013); Baik and Silverstein (2006)). As an example, we consider estimators based on thresholding the sample covariance matrix under sparsity assumptions developed by Bickel and Levina (2008a).

To be specific, suppose the true covariance matrix \( \Sigma \) belongs to the following class:

\[
\Omega(q, c_0(p), M) = \left\{ \Sigma : \Sigma_{j,j} \leq M, \sum_{k=1}^{p} |\Sigma_{j,k}|^q \leq c_0(p), \text{for all } j \right\},
\]

where \( 0 \leq q < 1 \), \( c_0(p) \) is a constant that depends on \( p \) and \( M \) is an absolute constant. Then Bickel and Levina (2008a, Theorem 1) show that if we define the thresholded covariance matrix \( \hat{\Sigma}_{j,k} = S_{j,k} \mathbb{1}_{(|S_{j,k}| \geq t)} \) for all \( 1 \leq j, k \leq p \) where \( S \) is the sample covariance matrix and \( t = O \left( \frac{\sqrt{\log p}}{n} \right) \), then

\[
\| \hat{\Sigma} - \Sigma \|_{1,1} = O_P \left( c_0(p)M \left( \frac{\log p}{n} \right)^{1-q} \right),
\]

Though the original error bound result for \( \hat{\Sigma} - \Sigma \) in Bickel and Levina (2008a) was shown in operator norm, they bounded \( \| \hat{\Sigma} - \Sigma \|_{1,1} \) in the proof. In particular if \( q = 0 \) and \( c_0(p) \leq s \) denotes the sparsity level,

\[
\| \hat{\Sigma} - \Sigma \|_{1,1} = O_P \left( s \sqrt{\frac{\log p}{n}} \right),
\]

meaning if \( n = O(s^2 \log p) \), Assumption 3.3 is satisfied.

On the other hand, if the covariance matrix \( \Sigma \) is not sparse but rather block-structured, we can use an alternative bound developed in Bickel and Levina (2008a). If \( \Sigma \) has \( K \) identical blocks where each block has \( p/K \) elements, we can ensure Assumptions 3.1 and 3.2 are satisfied if \( \Sigma_{j,k} = O \left( \frac{K}{p} \right) \) for each non-zero \( \Sigma_{j,k} \). Then if we choose \( \hat{\Sigma} \) to be the sample covariance matrix, Bickel and Levina (2008a) prove that

\[
\max_{j,k} \left| \frac{\hat{\Sigma}_{j,k} - \Sigma_{j,k}}{K/p} \right| = O_P \left( \frac{\sqrt{\log p}}{n} \right) \tag{8}
\]

since now we have \( \frac{\Sigma_{j,k}}{K/p} = O(1) \) for \( 1 \leq j \leq p \). Thus by (8) we know that

\[
\max_{j,k} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}| = O_P \left( \frac{K \sqrt{\log p}}{p \sqrt{n}} \right) \tag{9}
\]
and

$$\|\hat{\Sigma} - \Sigma\|_{1,1} = \max_{1 \leq j \leq p} \sum_{k=1}^{p} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}| = O_P\left(\sqrt{\frac{\log p}{n}}\right)$$

by (9), so that when \( n = O(K^2 \log p) \) Assumption 3.3 is satisfied.

The performance of our estimator also depends upon the following two properties of the augmented edge incidence matrix \( \tilde{\Gamma} \) appearing in our regularizer:

**Definition 3.1** (Compatibility factor \( k_T \), Hütter and Rigollet (2016)). We define the compatibility factor \( k_T \) of matrix \( \tilde{\Gamma} \) for a set \( T \subset \{1, 2, \ldots, p, p + 1, \ldots, p + m\} \) as:

$$k_0 := 1, \quad k_T := \inf_{\beta \in \mathbb{R}^p} \frac{\sqrt{|T|}\|\beta\|_2}{\|((\tilde{\Gamma})_T)\|_1} \quad \text{for} \ T \neq \emptyset.$$ 

This compatibility factor \( k_T \) reflects the degree of compatibility of the \( \ell_1 \)-regularizer \( \|((\tilde{\Gamma})_T)\|_1 \) and the \( \ell_2 \)-error norm \( \|\beta\|_2 \) for a set \( T \). This compatibility factor appears explicitly in the bounds of our main theorem.

**Definition 3.2** (Inverse scaling factor \( \rho \), Hütter and Rigollet (2016)). Let \( S := \hat{\Gamma}^\dagger := [s_1, \ldots, s_{m+p}] \), where \( \hat{\Gamma}^\dagger \) is the Moore-Penrose pseudoinverse of the matrix \( \hat{\Gamma} \), and define the inverse scaling factor as:

$$\rho := \|\hat{\Gamma}^\dagger\|_{1,2} = \max_{j=1,2,\ldots,m+p} s_j.$$ 

**Remark 3.4.** Definitions 3.1 and 3.2 are first proposed in Hütter and Rigollet (2016), though the definition of \( \rho \) is based on \( \hat{\Gamma} \) rather than \( \Gamma \), as discussed in Remark 3.9. Later we will see that \( \rho \) and \( k_T \) are crucial for our main results. Note that

$$\frac{\rho}{k_T} := \sup_{\beta \in \mathbb{R}^p, \beta \neq 0} \frac{\|((\tilde{\Gamma})_T)\|_1\|\hat{\Gamma}^\dagger\|_{1,2}}{\sqrt{|T|}\|\beta\|_2} \leq \sup_{\beta \in \mathbb{R}^p, \beta \neq 0} \frac{\sqrt{|T|}\|\tilde{\Gamma}\|_2\|\hat{\Gamma}^\dagger\|_{1,2}}{\sqrt{|T|}\|\beta\|_2} = \|\tilde{\Gamma}\|_{op}\|\hat{\Gamma}^\dagger\|_{1,2}.$$ 

The quantity \( \frac{\rho}{k_T} \) is similar in flavour to the condition number of the matrix \( \tilde{\Gamma} \).

Finally, we define the weighted graph Laplacian \( L := \tilde{\Gamma}^\dagger \Gamma \). Spectral properties of \( L \), in particular block sizes and smallest non-zero eigenvalues in each block will play a crucial role in the mean-squared error bounds we derive.

**Theorem 1.** Suppose \( \lambda_1 > 0 \) and Assumptions 3.1 to 3.3 are satisfied. If

$$\lambda_1 \geq \max\left\{ 48\sqrt{\frac{c_u \rho^2 \sigma^2 \log p}{n}}, 8\lambda_S \|L\beta^*\|_\infty \right\},$$

then there exist absolute positive constants \( C_u \) and \( C_1 \) such that with probability at least \( 1 - \frac{C_1}{p} \) we have

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C_u \min_T \left\{ \frac{\lambda_1^2|T|}{k_T^2 \lambda_{\min}(\Sigma + \lambda_S L)} \right\} \left( \frac{\lambda_1 \|((\tilde{\Gamma})_T)\|_1 + \lambda_1^2\|((\tilde{\Gamma})_T)\|_2^2}{\lambda_{\min}(\Sigma + \lambda_S L)} \right)$$

provided \( \frac{\lambda_1^2|T|}{k_T^2 \lambda_{\min}(\Sigma + \lambda_S L)} \to 0 \) (i.e., that the estimator is consistent).
Remark 3.5. Here \( \lambda_{\min}(\Sigma + \lambda_S L) \) plays the role of the restricted eigenvalue constant (see Bickel et al. (2009) for more details about this condition). Recall that from the definition of \( L \), if we define the diagonal matrix \( D \in \mathbb{R}^{p \times p} \) where each diagonal entry is \( D_{jj} = \sum_{k=1}^{p} |\hat{\Sigma}_{j,k}|, \) \( 1 \leq j \leq p \), then
\[
\Sigma + L := \Sigma - \hat{\Sigma} + D.
\]
Hence if \( \Sigma \) and \( \hat{\Sigma} \) are “close” as is specified by Assumption 3.3, then \( \Sigma + L \) is “close” to a diagonal matrix which ensures that \( \lambda_{\min}(\Sigma + \lambda_S L) \) may be bounded away from 0, even if \( \lambda_{\min}(\Sigma) = 0 \). The following Lemma makes this statement precise:

**Lemma 1.** Suppose that Assumption 3.2 and 3.3 are satisfied and \( 0 \leq \lambda_S \leq 1 \). Then
\[
\lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)\lambda_{\min}(\Sigma) + \lambda_S c^\ell_2 / 2.
\]

Thus even if \( \lambda_{\min}(\Sigma) = 0 \), choosing \( \lambda_S \) bounded away from 0 results in a well-posed inverse problem. On the other hand, in the classical LASSO analysis where \( \lambda_{\min}(\Sigma) > 0 \), we can choose \( \lambda_S = 0 \).

Remark 3.6. \( \|L \beta^*\|_\infty \) can be seen as a measure of the misalignment of the signal \( \beta^* \) and the graph represented by the matrix \( \Gamma \). Note that we require \( \lambda_1 \geq 8 \lambda_S \|L \beta^*\|_\infty \). Hence there is a clear trade-off in the choice of \( \lambda_S \). Choosing \( \lambda_S \) close to 1 ensures \( \lambda_{\min}(\Sigma + \lambda_S L) \) is bounded away from 0 but incurs a cost that scales with \( \|L \beta^*\|_\infty \).

In general, if \( \lambda_{\min}(\Sigma) = 0 \), indicating high correlations, we require \( \|L \beta^*\|_\infty \approx 0 \) (i.e., \( \beta^* \) is well-aligned with \( L \)) in order to obtain consistent mean-squared error bounds. Note that analysis of OWL (Figueiredo and Nowak, 2016) assumes \( L \beta^* = 0 \) (perfect alignment). If \( \lambda_{\min}(\Sigma) = 0 \) and \( \|L \beta^*\|_\infty \) is bounded far away from 0, we encounter identifiability challenges which leads to an inconsistent estimator of \( \beta^* \), just like the classical LASSO.

Remark 3.7. A natural question to consider is how the mean-squared error bound would change if the graph Laplacian penalty \( \lambda_S \|\Gamma \beta\|_2^2 \) were replaced by \( \lambda_S \|\beta\|_2^2 \) as is used in the (Zou and Hastie, 2005). Going through the analysis, \( \lambda_{\min}(\Sigma + \lambda_S L) \) would be replaced by \( \lambda_{\min}(\Sigma + \lambda_S I_{p \times p}) \) and hence pre-conditioning is still achieved. However the important difference and why we prefer the graph Laplacian penalty is because using our analysis the condition \( \lambda_1 \geq 8 \lambda_S \|L \beta^*\|_\infty \) would be replaced by \( \lambda_1 \geq 8 \lambda_S \|\beta^*\|_\infty \). Hence if we were in the strictly sparse case and \( \lambda_{TV} = 0 \) we would recover the mean-squared error bound:
\[
\| \hat{\beta} - \beta^* \|_2^2 \leq \frac{\log p}{\lambda_{\min}(\Sigma + \lambda_S I_{p \times p})} \|\beta^*\|_0^2.
\]
Note that this exactly matches the mean-squared error bound in (11) in Hebiri and van de Geer (2011) by replacing \( \|\beta^*\|_2^2 \) with the bound \( \|\beta^*\|_0 \|\beta^*\|_\infty^2 \). (The estimator of Hebiri and van de Geer (2011) is a generalization of Elastic Net from Zou and Hastie (2005).) In general we can not expect \( \|\beta^*\|_\infty \) to be close to zero, but in the case where \( \beta^* \) is well-aligned with \( L \), we would expect \( \|L \beta^*\|_\infty \) to be close to zero which would achieve sharper bounds.

Now we turn our attention to quantifying \( k_T \) and \( \rho \) to provide a more interpretable bound. We first have the following lemma to bound \( k_T \):

**Lemma 2.** Suppose \( T = T_1 \cup T_2 \) with \( T_1 \subset \{p+1, p+2, \ldots, p+m\} \) and \( T_2 \subset \{1, 2, \ldots, p\} \). Then we have
\[
k_T^{-1} \leq \frac{\lambda_{TV} \sqrt{2\|\Sigma\|_1 |T_1| + \sqrt{|T_2|}}}{\sqrt{|T_1| + |T_2|}}.
\]
The proof for this lemma can be found in the Appendix.

**Remark 3.8.** The compatibility factor $k_T$ depends on the choice of support $T$. Usually $T$ will be chosen as $T = \text{Supp}(\hat{\Gamma} \beta)$ for some $\beta$; then $T_1 = \text{Supp}(\Gamma \beta)$ and $T_2 = \text{Supp}(\beta)$ and Lemma 2 can be reduced to

$$k_T^{-1} \leq \frac{\lambda_{TV} \sqrt{2 \| \hat{\Sigma} \|_{1,1} \| \Gamma \beta \|_0 + \sqrt{\| \beta \|_0^2}}}{\sqrt{\| \Gamma \beta \|_0 + \| \beta \|_0}}.$$ 

To provide an upper bound for $\rho$ we first define the following graph-based quantities. If $G$ has $K$ connected components where $1 \leq K \leq p$, $L$ is block-diagonal with $K$ blocks. Let $L_k$ denote the $k^{th}$ block of $L$, $B_k \subset \{1, 2, ..., p\}$ denote the nodes corresponding to the $k^{th}$ block, and $\mu_k$ denote the smallest non-zero eigenvalue of $L_k$.

**Lemma 3.** Let $G$ denote the graph associated with $\hat{\Sigma}$. Then

$$\rho^2 \leq \max_{1 \leq k \leq K} \left\{ \frac{1}{|B_k|} + \frac{2}{1 + \mu_k \lambda_{TV}^2} \right\},$$

where $K$ is the number of connected components in $G$; $|B_k|$ is the corresponding number of nodes in $B_k$; and $\mu_k$ is the smallest nonzero eigenvalue of the weighted Laplacian matrix for the $k^{th}$ connected component.

Remark 3.9. Note that the definition of $\rho$ for the edge-incidence matrix $\Gamma$ was considered in Hütter and Rigollet (2016) while we define $\rho$ for the modified matrix $\hat{\Gamma}$, which involves both $\lambda_{TV}$ and the identity matrix $I_{p \times p}$. Prior results in Hütter and Rigollet (2016) show that $\|\Gamma^\dagger\|_{1,2}^2 \leq \frac{1}{\mu_k}$, while Lemma 3 provides an upper bound for $\|\hat{\Gamma}^\dagger\|_{1,2}^2$. In many cases, $\|\hat{\Gamma}^\dagger\|_{1,2}^2 << \|\Gamma^\dagger\|_{1,2}^2$.

By combining results from Lemmas 2 and 3 we have the following theorem:

**Theorem 2.** Suppose Assumptions 3.1 to 3.3 are satisfied and we choose

$$\lambda_1 \geq 48 \sqrt{\frac{\sigma^2 C_u \log p}{n}} \max_{1 \leq k \leq K} \left( \frac{1}{|B_k|} + \frac{2}{1 + \mu_k \lambda_{TV}^2} \right) + 8 \lambda_S \| L \beta^* \|_\infty.$$ 

Then there exist absolute positive constants $C_1$ and $C$ such that

$$\| \hat{\beta} - \beta^* \|_2^2 \leq C \frac{\lambda_1^2 \| \beta^* \|_0 + \min(\lambda_1^2 \lambda_{TV}^2 \| \hat{\Sigma} \|_{1,1}, \| \Gamma \beta^* \|_0, \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1) \| \beta^* \|_0}{\min(\lambda_2^2 (\Sigma + \lambda_S L), \lambda_\min(\Sigma + \lambda_S L))},$$

with probability at least $1 - \frac{C_1}{p}$ provided $\frac{\lambda_1^2 \| \beta^* \|_0 + \lambda_1^2 \lambda_{TV}^2 \| \hat{\Sigma} \|_{1,1} \| \Gamma \beta^* \|_0}{\lambda_\min(\Sigma + \lambda_S L)} \to 0$ and $\lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1 \leq 1$.

The proof of Theorem 2 is provided in Section 7. The upper bound involves a minimum where one term depends on $\| \Gamma \beta^* \|_0$ and the other depends on $\| \Gamma \beta^* \|_1$ by using different choices of $T$. This minimum of two terms also appears in Hütter and Rigollet (2016). Theorem 2 captures the role of $\lambda_{TV}$ and its impact on the mean-squared error (MSE) bounds. As $\lambda_{TV}$ increases, $\| \beta^* \|_0$ contributes less to the MSE, while $\| \Gamma \beta^* \|_0$ or $\| \Gamma \beta^* \|_1$ contributes more. To see this, note that the lower bound on $\lambda_1$ decreases with $\lambda_{TV}$ and the first term in the MSE scales as $\lambda_1^2 \| \beta^* \|_0$. On the other hand the second term of the MSE scales as $\lambda_1^2 \lambda_{TV}^2 \| \hat{\Sigma} \|_{1,1} \| \Gamma \beta^* \|_0$ or $\lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1$ and the lower bound on $\lambda_1 \lambda_{TV}$ increases as $\lambda_{TV}$ increases.
3.1. Discussion of main results

If we are in the setting where $\lambda_{\min}(\Sigma) > C > 0$, which corresponds to the classical LASSO setting, we can set $\lambda_S = \lambda_{TV} = 0$. From Theorem 2 we can see that

$$\|\hat{\beta} - \beta^*\|^2 \leq \frac{\sigma^2 c_n \log p}{n} \|\beta^*\|_0,$$

which is consistent with classical LASSO results. On the other hand if $\lambda_{\min}(\Sigma) \approx 0$ (columns are highly correlated) and $\|L\beta^*\|_\infty \approx 0$ ($\beta^*$ is well-aligned with $L$), we can set $0 < \lambda_S \leq 1$ and $\lambda_{TV} = C \max_{1 \leq k \leq K} \sqrt{|B_k| \mu_k}$; then we obtain the bound

$$\|\hat{\beta} - \beta^*\|^2 \leq \lambda_1^2 \|\beta^*\|_0 + \min(\lambda_1^2 \lambda_{TV}^2 \|\hat{\Sigma}\|_{1,1}, \|\Gamma\beta^*\|_0, \lambda_1 \lambda_{TV} \|\Gamma\beta^*\|_1)$$

where $\lambda_1^2 = O(\max_{1 \leq k \leq K} \frac{\sigma^2 c_n \log p}{n|B_k|})$ and $\lambda_1^2 \lambda_{TV} = O(\max_{1 \leq k \leq K} \frac{|B_k|}{\mu_k} \max_{1 \leq k \leq K} \frac{\sigma^2 c_n \log p}{n|B_k|})$. The upper bound may be well below the classical LASSO bound in (10) when $\min_k |B_k| \gg 1$ and $\Gamma\beta^* \approx 0$.

As mentioned earlier, if $\lambda_{\min}(\Sigma) \approx 0$ (columns are highly correlated) but $\|L\beta^*\|_\infty > C > 0$ (bad alignment), our method cannot guarantee a consistent estimator for $\beta^*$; Cluster Representative LASSO and Ordered Weighted LASSO will also fail in this case. Identifiability assumptions arise, since if two columns of $X$ are nearly identical but the corresponding elements of $\beta^*$ are substantially different, no method will be able to accurately estimate parameter values in the absence of additional structure.

We now discuss the roles of the various parameters associated with the MSE bound.

**Role of $\lambda_S$** The smoothing penalty plays the role of a pre-conditioner where the trade-off is the addition of another term $\lambda_S \|\hat{L}\beta^*\|_\infty$. This can also be seen in the optimization problem (5) where $X$ is transformed to $\tilde{X}$, so even if the restricted eigenvalue condition is not satisfied for $X$, it is satisfied for $\tilde{X}$. What distinguishes our results from previous work using pre-conditioners for the LASSO (Jia, Rohe, et al., 2015; Wauthier, Jojic, and Jordan, 2013) is that prior work does not address the case where $\lambda_{\min}(\Sigma) = 0$, which is where the total variation penalty is important. See also Remark 3.7.

**Role of $\lambda_{TV}$** As mentioned earlier, the total variation penalty promotes estimates well-aligned with the graph. As $\lambda_{TV}$ increases, the sparsity parameter $\lambda_1$ decreases while $\lambda_1 \lambda_{TV}$ increases. By increasing $\lambda_{TV}$ we can also adapt to settings where $\beta^*$ is not sparse provided that $\Gamma\beta^*$ is sparse (see the examples of specific graph structures below).

**Graph-based quantities** Two important parameters of the covariance graph are $\mu_k$ (the smallest non-zero eigenvalue of a block) and $|B_k|$ (the block size). Clearly the larger $\mu_k$ and $|B_k|$, the lower the bound on $\lambda_1$ which potentially suggests lower mean-squared error. On the other hand, as we illustrate with specific examples later, larger $\mu_k$ typically indicates higher correlation between more covariates and larger $|B_k|$ corresponds to nodes being correlated, which means $\lambda_{\min}(\Sigma)$ is smaller.

3.2. Specific covariance graph structures

In this section we explore two different specific graph structures and discuss suitable choices of $\lambda_S, \lambda_1$ and $\lambda_{TV}$. For each graph structure we assume

$$\Sigma_{jj} = a > 0 \text{ for } 1 \leq j \leq p \quad \text{and} \quad \Sigma_{jk} = ar \; \forall(j, k) \in E \text{ for some } 0 < r \leq 1;$$

we refer to $r$ as the correlation coefficient. Note that here $a$ is a normalization parameter that we set to ensure such that Assumptions 3.1 and 3.2 are satisfied. We will talk about the specific choices of $a$ for each graph structure below. For interpretability and simplicity of exposition, we further assume that $\tilde{\Sigma} = \Sigma − \text{that is, that we have side information about the correlation graph.
3.2.1. Block covariance graph

We first consider a block complete graph $G$ that has $K$ connected components and each connected component is a complete graph with $\frac{p}{K}$ nodes. The corresponding covariance matrix $\Sigma$ (potentially after a suitable permutation of rows and columns) is block diagonal with $K$ blocks of size $\frac{p}{K} \times \frac{p}{K}$. Each of these blocks can be written as

$$ar\mathbb{1}_{p/K}\mathbb{1}_{p/K}^T + a(1-r)I_{p/K},$$

where $\mathbb{1}_{p/K}$ is the vector of $p/K$ ones.

![Graph structure and corresponding covariance matrix](image)

**Fig 1. Block complete graph with $p = 16$ and $K = 4$.**

We set $a = \frac{K}{p}$ to ensure that Assumptions 3.1 and 3.2 are satisfied. In the extreme case where $K = p$, we are in the independent case and the estimator reduces to the standard LASSO estimator; whereas for $K = 1$, we are in the fully-connected graph case.

The following lemma provides specific bounds on $\max_{1 \leq k \leq K} \frac{1}{|B_k|}, \mu_k, \rho, \lambda_{\min}(\Sigma)$:

**Lemma 4.** For a block complete graph with details described above, suppose that $\hat{\Sigma} = \Sigma$. Then we have

$$\max_{1 \leq k \leq K} \frac{1}{|B_k|} = \frac{K}{p},$$

$$\mu_k = r, \text{ for all } k,$$

$$\rho \leq \sqrt{\frac{K}{p} + \frac{2}{1 + r\lambda_{TV}^2}},$$

$$\lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)(1 - r)\frac{K}{p} + \lambda_S r.$$

The proof of Lemma 4 is deferred to the Appendix. Note that if $r = 1$ then $\lambda_{\min}(\Sigma) = 0$ but $\lambda_{\min}(\Sigma + \lambda_S L) \geq \lambda_S$. Using Lemma 4, we have the following mean-squared error bound for the block complete graph:

**Corollary 1.** Suppose that Assumptions 3.1 and 3.2 are satisfied and $\hat{\Sigma} = \Sigma$. If

$$\lambda_1 \geq 48\sqrt{\frac{\sigma^2 c_u \log p}{n} \left( \frac{K}{p} + \frac{2}{1 + r\lambda^2_{TV}} \right) + 8\lambda_S \| \beta^* \|_\infty},$$

then

$$\| \beta - \hat{\beta} \|^2 \leq \frac{1}{n} \mathbb{E} \| \mathbf{Y} \|^2.$$
and \( \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1 \leq 1 \). Then with probability at least \( 1 - \frac{C_1}{p} \)
\[
\| \hat{\beta} - \beta^* \|_2^2 \leq \frac{C(\lambda_1^2 \| \beta^* \|_0 + \min(\lambda_1^2 \lambda_{TV}^2 \| \Gamma \beta^* \|_0, \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1))}{\min\{(1 - \lambda_S)(1 - r) \frac{p}{p} + \lambda_S r, [(1 - \lambda_S)(1 - r) \frac{K}{p} + \lambda_S r]^2\}}
\] given the estimator is consistent, where \( C_1, C \) are absolute positive constants.

Note that for \( r \approx 1 \) and \( \Gamma \beta^* \approx 0 \) (near-perfect alignment which corresponds to the parameters in each block having the same values), if we choose \( \lambda_S > 0, \lambda_{TV}^2 \gtrsim \frac{p}{K}, \) and \( \lambda_1^2 = O\left(\frac{K \log p}{pn}\right) \), then
\[
\| \hat{\beta} - \beta^* \|_2^2 \leq \frac{K \log p}{n};
\]
that is, the MSE is not determined by the number of nonzeros in \( \beta^* \), but rather by \( K \), the number of clusters of nodes. A similar scaling was derived in Figueiredo and Nowak (2016) also under the assumption that \( \Gamma \beta^* \approx 0 \).

### 3.2.2. Chain covariance graph

The covariance matrix corresponding to the chain graph satisfies \( \Sigma_{jj} = 1 \) for all \( j \) and \( \Sigma_{jk} = r \) for all \((j,k) \in E \) where \( E = \{(1,2), (2,3), \ldots, (p-1,p)\} \). Assumptions 3.1 and 3.2 are clearly satisfied and requiring \( r \in (0, \frac{1}{2}) \) ensures \( \Sigma \) is positive semi-definite. Note that the chain graph is fully connected so \( K = 1 \) and \( B_1 = \{1, 2, \ldots, p\} \).

The following lemma provides bounds on \( \rho \) and \( \lambda_{\min}(\Sigma + \lambda_S L) \) for the chain covariance graph:

**Lemma 5.** For a chain graph with details described above, suppose that \( \hat{\Sigma} = \Sigma \). Then
\[
\rho \leq \frac{1}{r} \frac{2\pi}{\lambda_{TV} + 1}, \quad \lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)(1 - 2r) + \lambda_S.
\]

Using Lemma 5 we have the following corollary for the chain graph:
Corollary 2. Suppose Assumptions 3.1 and 3.2 are satisfied, and \( \hat{\Sigma} = \Sigma \). If we choose
\[
\lambda_1 > 48 \sqrt{\frac{\sigma^2 c_u \log p}{n} \left( \frac{1}{p} + \frac{2\pi}{r \lambda_{TV} + 1} \right) + 8 \lambda_S \| L \beta^* \|_\infty}
\]
and \( \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1 \leq 1 \), then with probability at least \( 1 - \frac{C_1}{p} \) we have
\[
\| \hat{\beta} - \beta^* \|_2^2 \leq C \left( \lambda_1 \| \beta^* \|_0 + \min \{ \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_0, \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1 \} \right)
\]
given the estimator is consistent, where \( C_1, C \) are absolute positive constants.

We consider an example where the alignment between the chain graph and \( \beta^* \) is strong but imperfect. Suppose that within \( \beta^* \) there are \( O(1) \) blocks which are active, and within each active block all the coefficient have identical magnitude. Further, suppose \( n \preceq p \). In this setting, \( \| \Gamma \beta^* \|_0, \| \Gamma \beta^* \|_1 \approx 1 \).

If we set \( \lambda_{TV} \approx \sqrt{\| \beta^* \|_0} \) and \( \lambda_S \approx 0 \) then the MSE bound says
\[
\text{MSE}_GTV \preceq \sqrt{\| \beta^* \|_0 \log p}
\]
which is stronger than the Lasso guarantee of
\[
\text{MSE}_\text{Lasso} \preceq \frac{\| \beta^* \|_0 \log p}{n}.
\]

3.2.3. Lattice covariance graph

We next consider a covariance structure corresponding to a lattice graph with \( p \) nodes (here \( p \) must be a perfect square). Both sides of such a lattice have length \( \sqrt{p} \) and the corresponding covariance matrix satisfies
\[
\Sigma_{j,k} = \begin{cases} 
1, & \text{if } j = k, \\
r, & \text{if } |j-k| = 1 \text{ and } \min(j,k) \neq 0 \text{ mod } \sqrt{p}, \\
r, & \text{if } |j-k| = \sqrt{p} \\
0, & \text{else}.
\end{cases}
\]

We require \( r \in (0, \frac{1}{4}) \) so that \( \Sigma \) is positive semi-definite. Clearly Assumptions 3.1 and 3.2 are satisfied for any \( r \in (0, \frac{1}{4}) \), and we note that the lattice graph is fully connected, so \( K = 1 \) and \( B_1 = \{1, 2, \ldots, p\} \). The following lemma gives bounds on \( \rho \) and \( \lambda_{\min}(\Sigma + \lambda_S L) \):

Lemma 6. For a lattice graph with details described above, suppose that \( \hat{\Sigma} = \Sigma \). Then
\[
\rho \leq \sqrt{\frac{1}{p} + \frac{5\pi \log(2 + r \lambda_{TV})}{r^2 \lambda_{TV}^2 + 1} + \frac{10\pi}{r \lambda_{TV} \sqrt{p} + 1}},
\]
\[
\lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)(1 - 4r) + \lambda_S.
\]

Using Lemma 6 we have the following corollary for the lattice graph:
Corollary 3. Suppose $\hat{\Sigma} = \Sigma$. If we choose

$$\lambda_1 > 48 \sqrt{\frac{\sigma^2 c_a \log p}{n}} \left( \frac{1}{p} + 5\pi \log(2 + r\lambda_{TV}) + \frac{10\pi}{r\lambda_{TV} \sqrt{\rho} + 1} \right) + 8\lambda_S \| L\beta^* \|_{\infty},$$

and $\lambda_1 \lambda_{TV} \| \Gamma\beta^* \|_1 \leq 1$, then with probability at least $1 - \frac{C_1}{p}$ we have

$$\| \hat{\beta} - \beta^* \|_2^2 \leq C \bigg( \lambda_1^2 \| \beta^* \|_0 + \min \{ \lambda_1^2 \lambda_{TV} \| \Gamma\beta^* \|_0, \lambda_1 \lambda_{TV} \| \Gamma\beta^* \|_1 \} \bigg) \min \{ (1 - \lambda_S)(1 - 4r) + \lambda_S, (1 - \lambda_S)(1 - 4r) + \lambda_S \},$$

given the estimator is consistent, where $C_1, C$ are absolute positive constants.

We again consider an example where the alignment between the graph and $\beta^*$ is strong but imperfect. Suppose that all the active nodes within a $\sqrt{p} \times \sqrt{p}$ lattice are contained in a $\sqrt{\| \beta^* \|_0} \times \sqrt{\| \beta^* \|_0}$ sublattice, and suppose all active nodes have equal magnitude. Then $\| \Gamma\beta^* \|_0, \| \Gamma\beta^* \|_1 \approx \sqrt{\| \beta^* \|_0}$.

We assume $n \geq p$ and we set $\lambda_{TV} \approx \sqrt{n}$, $\lambda_S \approx 0$ and $\lambda_1 \approx \frac{\log p}{n}$. Our MSE bound says

$$\text{MSE}_{\text{GTV}} \leq \lambda_1^2 \| \beta^* \|_0 + \lambda_1^2 \lambda_{TV} \| \Gamma\beta^* \|_0 \approx \frac{\| \beta^* \|_0 \log p}{n^2} + \frac{\sqrt{\| \beta^* \|_0 \log p}}{n} \approx \frac{\sqrt{\| \beta^* \|_0 \log p}}{n},$$

which is better than the Lasso guarantee of

$$\text{MSE}_{\text{Lasso}} \leq \frac{\| \beta^* \|_0 \log p}{n}.$$

Note that the MSE$_{\text{GTV}}$ bound from this example is identitical to the MSE$_{\text{GTV}}$ bound from the example considered in the chain graph section. On one hand, our bound on $\rho$ is stronger in the lattice graph case. This is consistent with Hütter and Rigollet (2016) even though we study the inverse scaling factor of a somewhat different matrix. However, this phenomenon is counterbalanced by the fact that it is easier to construct near perfect alignment between the chain graph and $\beta^*$ than between the lattice graph and $\beta^*$. With the chain graph, for any value of $\| \beta^* \|_0$ we can have $\| \Gamma\beta^* \|_0 \approx 1$. However, for the lattice graph it is impossible to give a general bound on $\| \Gamma\beta^* \|_0$ which is independent of $\| \beta^* \|_0$. The best possible alignment yields $\| \Gamma\beta^* \|_0 \approx \sqrt{\| \beta^* \|_0}$.

3.2.4. Random regular covariance graph

To get a sense of how $\rho$ scales in a more general setting, we consider a covariance matrix determined by a random d-regular graph. To do this, we construct an adjacency matrix $A$ corresponding to a randomly generated d-regular graph, and then let $\Sigma = I + rA$. For $r \in (0, \frac{1}{d})$, $\Sigma$ is positive semi-definite by diagonal dominance, and Assumptions 3.1 and 3.2 are satisfied. The following lemma provides bounds on $\max_{1 \leq k \leq K} \frac{1}{|B_k|}, \mu_k, \rho, \lambda_{\min}(\Sigma + \lambda_S L)$:

Lemma 7. Suppose that $\hat{\Sigma} = \Sigma$ and $d > 2$. For a random d-regular graph we have

$$|B_1| = p$$

with probability at least $1 - p^{-cd}$ for a constant $c_d$ depending only on $d$.

$$\mu_1 \geq \frac{d}{i}$$

with probability at least $1 - p^{-cd}$

$$\rho \leq \sqrt{\frac{1}{p} + \frac{14}{i + dr\lambda_{TV}^2}}$$

with probability at least $1 - p^{-cd}$

$$\lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)(1 - dr) + \lambda_S.$$
The crucial ingredient in Lemma 7 is a result from Friedman (2004) which establishes that the adjacency matrix of a random regular graph has a significant gap between its largest and second largest eigenvalues with high probability. Using Lemma 7 we obtain the following result:

**Corollary 4.** Suppose $\tilde{\Sigma} = \Sigma$. If we choose

$$\lambda_1 > 48 \sqrt{\frac{\sigma^2 c_n \log p}{n} \left( \frac{1}{p} + \frac{14}{7 + dr\lambda_{1TV}^2} \right)} + 8\lambda_S \|L\beta^*\|_{\infty}$$

and $\lambda_1 \lambda_{TV} \|\Gamma\beta^*\|_1 \leq 1$, then with probability at least $1 - \frac{C_1}{p}$ we have

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{C(\lambda_1^2 \|\beta^*\|_0 + \min\{\lambda_1^2 \lambda_{TV}^2 \|\Gamma\beta^*\|_0, \lambda_1 \lambda_{TV} \|\Gamma\beta^*\|_1\})}{\min\{[(1 - \lambda_S)(1 - dr) + \lambda_S], [(1 - \lambda_S)(1 + dr) + \lambda_S]^2\}}$$

given the estimator is consistent, where $C_1$, $C$ are positive constants depending only on $d$.

4. Simulation study

In this section we compare our proposed graph-based regularization method with other methods on the block, chain and lattice graphs considered in the corollaries above. The graphs and corresponding covariance structures are constructed as follows:

**Block Complete Graph** $\Sigma$ is block diagonal with $K$ blocks, each of size $\frac{p}{K} \times \frac{p}{K}$. Following the discussion in Section 3.2.1, all the diagonal elements are set to $\frac{K}{p}$ and all the off-diagonal elements in each block are set to $\frac{K}{p} r$ with $r \in (0, 1)$. Here $r$ is the correlation coefficient and will be set to different values in the experiments. Specifically, let

$$B = \frac{K}{p} \left( r \mathbb{1}_{\frac{p}{K}} \mathbb{1}_{\frac{p}{K}}^\top + (1 - r) I_{\frac{p}{K}} \right) \quad \text{and} \quad \Sigma = I_K \otimes B,$$

where $\otimes$ denotes the Kronecker product. To set the true coefficient vector $\beta^*$, we first randomly choose $\ell$ of the $K$ blocks to be “active blocks”. Then we set the elements in $\beta^*$ that correspond to the $\ell$ active blocks to be $\beta^*_j \sim \mathcal{N}(1, 0.01^2)$ when $i$ belongs to these $\ell$ active blocks and all other elements in $\beta^*$ to be zero (inactive). That is, let $S \in \{0, 1\}^p$ indicate the indices in active blocks (and hence the support of $\beta^*$); then

$$\beta^* \sim \mathcal{N}(S, 0.01^2 \text{diag}(S)).$$

**Chain Graph:** Following the discussion in Section 3.2.2, we set elements in the main diagonal of $\Sigma$ to be one, the first off-diagonal elements to be $r$ with $r \in (0, \frac{1}{2})$, and all the other elements to be zero; i.e.,

$$\Sigma_{j,k} = \begin{cases} 1, & \text{if } j = k, \\ r, & \text{if } |j - k| = 1, \\ 0, & \text{else}. \end{cases}$$

The corresponding true coefficient vector $\beta^*$ is set to have $\beta^*_j \sim \mathcal{N}(1, 0.01^2)$ for $1 \leq j \leq s$ and the remaining elements to be zero. That is, let $S \in \{0, 1\}^p$ have its first $s < p$ elements be one and the remaining be zero; then

$$\beta^* \sim \mathcal{N}(S, 0.01^2 \text{diag}(S)).$$
**Lattice Graph**  Following the discussion in Section 3.2.3, we construct $\Sigma$ as follows.

$$
\Sigma_{j,k} = \begin{cases} 
1, & \text{if } j = k, \\
r, & \text{if } |j - k| = 1 \text{ and } \min(j,k) \neq 0 \mod \sqrt{p}, \\
r, & \text{if } |j - k| = \sqrt{p} \\
0, & \text{else}.
\end{cases}
$$

The corresponding true coefficient vector $\beta^*$ with $s$ active elements is set to $\beta^*_j \sim N(1,0.01^2)$ if $j \leq \sqrt{s} \mod \sqrt{p}$ and $j \leq \sqrt{ps}$ and is set to $\beta^*_j = 0$ otherwise. This corresponds to an active $\sqrt{s} \times \sqrt{s}$ sublattice within the $\sqrt{p} \times \sqrt{p}$ lattice. The remaining elements outside of this sublattice are set to zero.

The data is generated according to $y = X\beta^* + \epsilon$ with $X \in \mathbb{R}^{n \times p}$ and $\epsilon \in \mathbb{R}^n$. Each row of $X$ is independently generated from $\mathcal{N}(0, \Sigma_{p \times p})$ and $\epsilon$ is generated from $\mathcal{N}(0, \sigma^2 I_{n \times n})$ with $\sigma = 0.01$. Additionally, we generate $X_{\text{ind}} \in \mathbb{R}^{1000 \times p}$ with each row of $X_{\text{ind}}$ independently generated from $\mathcal{N}(0, \Sigma_{p \times p})$. This $X_{\text{ind}}$ provides side information that can be used to improve estimates of $\Sigma$. This $X_{\text{ind}}$ can be used for covariance estimation (GTV) or clustering (CRL) before parameter estimation.

We show how our proposed graph-based regularization scheme compares to existing state-of-the-art methods in terms of mean-squared error $\text{MSE} = \|\hat{\beta} - \beta^*\|_2^2$. For all methods, tuning parameters are chosen based on five-fold cross-validation. In particular we consider the following estimation procedures:

**GTV-Esti (Our method)**  Graph-based total variation (GTV) method using original design matrix $X \in \mathbb{R}^{n \times p}$ for both covariance matrix estimation and parameter estimation. To implement GTV-Esti, we first use $X$ to compute the estimated covariance matrix, $\hat{\Sigma}$, using hard thresholding of the sample covariance matrix with a threshold is chosen by cross validation (see Bickel and Levina (2008a) for more details). We construct the edge incidence matrix $\Gamma$ based on $\hat{\Sigma}$ and then estimate $\hat{\beta}$ using (5).

**GTV-Indep (Our method)**  This approach is equivalent to GTV-Esti (above), except that the side information $X_{\text{ind}}$ is used to compute the estimated covariance matrix $\hat{\Sigma}$.

**GTV-Indep (Our method)**  Cluster Representative LASSO (CRL) method of Bühlmann et al. (2013) using $X$ for both covariate clustering and parameter estimation. To implement CRL-Esti, we first use $X$ for covariate clustering using canonical correlations in $X$ (see Buhlmann et al. (2013, Algorithm 1) for more details), then the Cluster Representative LASSO is implemented based on the clusters.

**CRL-Indep**  This approach is equivalent to CRL-Esti (above), except that the side information $X_{\text{ind}}$ is used to improve clustering of the covariates. That is, we run CRL as before, but based on the canonical correlations computed from $X_{\text{ind}}$.

**LASSO**  Standard LASSO (Tibshirani, 1996).

**Elastic Net**  Method from (Zou and Hastie, 2005) which includes both an $l_1$ and an $l_2$ penalty term in order to encourage grouping strongly correlated predictors.

**OWL**  Ordered Weighted LASSO (Bogdan et al., 2013). We set the weights for OWL corresponding to the OSCAR regularizer (Bondell and Reich, 2008), i.e., $w_i = \lambda_1 + \lambda_2(p - i)$ with $1 \leq i \leq p$ and $\lambda_1, \lambda_2 \geq 0$.

We want to investigate how the mean-squared error (MSE) changes with number of observations $n$, number of covariates $p$, support size of the true coefficient vector $s = ||\beta^*||_0$, and the correlation coefficient $r$ between covariates. The results are summarized in Figures 3 and 4. We show the median MSE of 100 trials.
and we add error bars with the standard deviation (of the median) estimated using the bootstrap method with 500 resamplings on the 100 MSEs. In Figure 3 we see that over the different graph structures and values of \( p, n \), GTV-Esti usually has lower MSE than CRL-Esti, OWL, Elastic Net and LASSO; if we have additional side information we can achieve better results by using GTV-Indep or CRL-Indep. We can also see that the MSE decreases as \( n \) increases and MSE increases as \( p \) or \( s \) increases, which is consistent with our theoretical results.

In Figure 4b we see that the block graph GTV-Esti, CRL-Esti and OWL have lower MSE when \( r \) increases; this phenomenon is reasonable because when \( r \) increases it will be easier for these methods to capture true cluster information from \( X \) and then the results will be better. For the chain graph in Figure 4d we can see GTV-Esti has lower MSE when \( r \) increases but the MSE for CRL-Esti, OWL, Elastic Net and LASSO have no significant changes. In particular, note that the CRL requires one to cluster features independently of label information in \( y \), and for the chain graph there are no clear clusters. In contrast, GTV is essentially clustering features based on both the covariance structure in \( \hat{\Sigma} \) and the label information in \( y \), so that a group of active elements in \( \beta^* \) on a subset of the chain can be correctly identified as a cluster.

We next test how the error scales with \( \| \Gamma \beta^* \|_0 \) and \( \| \Gamma \beta^* \|_1 \). In Figure 5a we take a chain graph with \( p = 280 \) nodes and let the first \( s = 80 \) nodes be active. For the active nodes we set \( \beta_j^* \sim \mathcal{N}(1, \sigma^2) \) for varying values of \( \sigma \). In other words we change the value of \( \| \Gamma \beta^* \|_1 \) while holding \( \| \Gamma \beta^* \|_0 \) constant. We see that GTV is reasonably robust to increases in \( \| \Gamma \beta^* \|_1 \) and still performs well with high levels of noise within the active block.

In Figure 5b we again look at a chain graph with \( p = 280 \) nodes and \( s = 80 \) active nodes, but this time we break up the active nodes into distinct blocks. Each active node is chosen from \( \mathcal{N}(1, .01^2) \). We measure MSE a function of the number of distinct blocks the active nodes are divided into. In other words, this setting measures robustness to \( l_0 \) misalignment as opposed to \( l_1 \) misalignment. We see that GTV performs well even when \( \| \Gamma \beta^* \|_0 \) is reasonably large, again suggesting that our methods are robust to moderate amounts of misalignment between the graph and \( \beta^* \).

We also visual comparison of the support recovery for the seven different methods. The results are shown in Figures 6 and 7. For both cases we see that GTV methods (both with and without side information) outperform the other methods in terms of recovering the true structure of \( \beta^* \). Figure 6 shows that in the block graph setting, the LASSO, Elastic Net and CRL-Esti struggle to recover the signal structure and OWL produces slightly noisier estimates than GTV; the other three methods successfully recover the true signal structure. From Figure 7, we can see in the chain graph setting that only GTV-Indep and GTV-Esti achieve satisfactory results for recovering true signal \( \beta^* \). We can see that except for LASSO, Elastic Net and CRL-Esti, the methods can approximately recover the true signal structure.

5. Application to real fMRI data

In this section we describe an application of the proposed GTV methodology to neural decoding with fMRI data. The fMRI data we use was provided by the Postle Lab at UW-Madison (Lewis-Peacock and Postle, 2008). The data contains fMRI measurements for ten different participants. Each participant takes part in an fMRI experiment with ninety trials, where for each trial the participant views a picture of either a human face or something with no human face present. fMRI blood-oxygen-level dependent (BOLD) signals corresponding to these ninety trials are recorded when the participant is looking at these pictures, and for each voxel we compute how well the measured bold signal fits a signal template corresponding to the expected hemodynamic response to a visual stimulus. We model each individual separately. A parameter associated with each (trial,voxel) pair is stored in a matrix \( X \in \mathbb{R}^{90 \times p} \), where \( p \) is the number of voxels in the brain recorded for the individual. The values of \( p \) for the ten participants are: \{7303, 5179, 4176, 5651, 5847, 8082, 4101, 6600, 9111, 7193\}. 


FIG 3. MSE for varying covariance graph structures and values of \( n \) and \( p \). Median of 100 trials are shown, and error bars denote the standard deviation of the median estimated using the bootstrap method with 500 resamplings on the 100 mean-squared errors. GTV-Esti yields lower MSEs than other methods for a broad range of \( n, p \).
Fig 4. MSE for varying covariance graph structures and values of s and r. Median of 100 trials are shown, and error bars denote the standard deviation of the median estimated using the bootstrap method with 500 resamplings on the 100 mean-squared errors. GTV-Esti yields lower MSEs than other methods for a broad range of s, r.

Fig 5. Chain graph (p=280, n=100, s=80 and r=.4). On left, all active nodes are contained in one continuous block, and active nodes are chosen from $\mathcal{N}(1, \sigma^2)$. On right, active nodes are separated into an increasing number of distinct block, and all active nodes are chosen from $\mathcal{N}(1, .01^2)$. Plots demonstrate that GTV performs well with moderate amounts of misalignment between the graph and $\beta^*$. Median of 100 trials are shown, and error bars denote the standard deviation of the median estimated using the bootstrap method with 500 resamplings on the 100 mean-squared errors.
Fig 6. Comparison of seven different estimator’s behavior for block graph setting with simulation parameters $p = 280, n = 100, s = 84, r = 0.6$.

Fig 7. Comparison of seven different estimator’s behavior for chain graph setting with simulation parameters $p = 280, n = 100, s = 80, r = 0.45$. 
We normalize this fMRI data matrix $X$ such that each column has unit Euclidean norm before analysis. For each participant we also have the response variable $y \in \mathbb{R}^{90}$ which records whether each picture is a “face” picture for each trial: $y_i = 1$ if picture $i$ is “face” picture and $y_i = 0$ otherwise. Although the responses are binary, we use least-squares loss based on the original objective function. For each participant, the Postle Lab also provides a time series version of the fMRI measurements $X_{TS} \in \mathbb{R}^{1170 \times p}$. This $X_{TS}$ matrix contains the entire BOLD signal for the entirety of the fMRI experiment.

5.1. Implementation for GTV

To implement our GTV method on the fMRI dataset, we use $X_{TS}$ to estimate the covariance matrix. We first normalize $X_{TS}$ such that each column has unit Euclidean norm. Then we estimate the covariance matrix by using a diagonal plus low-rank matrix decomposition (Saunderson, Chandrasekaran, Parrilo, and Willsky (2012), with connections to the spiked covariance model Donoho et al. (2013); Baik and Silverstein (2006)) which seemed to fit this data; we denote this estimator $\hat{\Sigma}$. The resulting covariance estimate $\hat{\Sigma}$ is not necessarily sparse. If the estimated covariance matrix is dense, there are $O(p^2)$ terms in (2), affecting the speed of computation for both the smoothing penalty and the total variation penalty. For faster computation, instead of using the entire estimated covariance matrix for the graph-based regularizers, we obtain a skeleton graph for $G$ and define $\hat{\Sigma}$ with respect to this graph, which reduces the number of terms in (2) to $O(p)$. Our idea is to find a skeleton graph with maximum possible total absolute edge weight. Thus we obtain a skeleton tree graph $G_T$ through the following three steps: (1) Define $\hat{\Sigma}_{Pos} \in \mathbb{R}^{p \times p}$ with $\hat{\Sigma}_{Pos,j,k} = |\hat{\Sigma}_{j,k}|$ for $1 \leq j, k \leq p$; (2) Find a maximum spanning tree $G_{MST}$ based on this matrix $\hat{\Sigma}_{Pos}$ by implementing Prim’s algorithm (Prim, 1957) based on a greedy search; (3) For each edge $(j, k)$ in the maximum spanning tree $G_{MST}$, replace the edge weight $\hat{\Sigma}_{Pos,j,k}$ by $\hat{\Sigma}_{j,k}$. We call this new tree graph with updated edge weights $G_T$. Then the smoothing and total variation penalties based on $G_T$ can be implemented efficiently.

5.2. Results

We compare our GTV method with Ordered Weighted LASSO, Cluster Representative LASSO, Elastic Net and the standard LASSO method. According to Lewis-Peacock and Postle (2008), for each participant the order of showing ninety pictures is random, thus we use fMRI data corresponding to the first sixty trials as the training set $X_{train}$ and fMRI data corresponding to the remaining thirty trials as the testing set $X_{test}$. For all five models tuning parameters were set using five-fold cross validation. For GTV, Lasso, CRL and Elastic Net we use a default set of choices for $\lambda_1$ provided by the glmnet package, while for OWL we use the Atlas package with a custom set of $\lambda_1$. The classification accuracy of these five methods on the testing set are summarized in Figure 8, which shows that GTV has highest classification accuracy among all methods for eight participants out of ten. CRL performs similarly to the LASSO since in the clustering step CRL does not create many big clusters for covariates. OWL has better performance than CRL, Elastic Net and LASSO in terms of classification accuracy and is closest in accuracy to GTV.

6. Conclusion

This paper describes a new graph-based regularization method for high-dimensional regression with highly-correlated designs. The structure of the estimator leverages ideas behind the Elastic Net (Zou and Hastie, 2005), the Fused LASSO (Tibshirani et al., 2005), the edge LASSO (Sharpnack et al., 2012), trend filtering on graphs (Wang et al., 2016), and graph total variation (Shuman et al., 2013; Hütter and Rigollet, 2016). Unlike these past works, this paper explores how graph-based regularization can mitigate the effects of
highly-correlated covariates. Under our model, the graph corresponding to the covariance structure of the covariates also provides prior information about the similarities among elements in the regression weights. Thus this graph allows us to effectively pre-condition our design matrix and regularize regression weights to promote alignment with the covariance structure of the problem. We are able to provide mean-squared error bounds in settings where covariates are highly dependent, provided there is alignment between the \( \beta^* \) and graph. We also demonstrate in both simulations and an fMRI application superior performance of our method compared to OWL, Lasso, Elastic Net and CRL. The proposed framework allows us to leverage correlation structure jointly with the response variable \( y \), in contrast to previous work that depended upon clustering covariates independent of the responses. In settings where there exist very strong clusters (like the block graph studied above), clustering with and without responses yield similar results. However, when correlations are too weak to reveal strong clusters and yet too strong for the LASSO alone to be effective (like with the chain and lattice graphs studied above), the implicit response-based clustering associated with our method can yield significant performance benefits. The results in this paper suggest several exciting avenues for future exploration, including more refined performance bounds for additional classes of graphs, extensions to logistic regression and other linear models, and more extensive evaluations on real-world data.

7. Proofs

7.1. Proof of Theorem 1

Much of our analysis follows standard steps for analysis of regularized M-estimators (cf., Bickel et al. (2009); Negahban, Ravikumar, Wainwright, and Yu (2012); van de Geer (2000)), but we face two additional challenges not present in these works. First, since the regularization penalty in Equation (5) is \( \| \hat{\Gamma}\beta \|_1 \) rather than \( \| \beta \|_1 \) we need to deal with error terms involving \( \tilde{X}\hat{\Gamma}^\dagger \) instead of \( \tilde{X} \). To address this we incorporate techniques from Hütter and Rigollet (2016) and Raskutti and Yuan (2015). Second, we need to establish a restricted eigenvalue condition for \( \tilde{X} \) rather than \( X \). We incorporate techniques from Raskutti, Wainwright, and Yu (2010) in order to accomplish this.

Based on the optimization problem (5), by the definition of \( \hat{\beta} \) and the basic inequality,

\[
\frac{1}{n} \| \tilde{y} - \tilde{X} \hat{\beta} \|_2^2 + \lambda_1 \| \hat{\Gamma} \hat{\beta} \|_1 \leq \frac{1}{n} \| \tilde{y} - \tilde{X} \beta^* \|_2^2 + \lambda_1 \| \hat{\Gamma} \beta^* \|_1.
\]
By simple re-arrangement,
\[
\frac{1}{n} \| \bar{X} (\hat{\beta} - \beta^*) \|_2^2 \leq \frac{2}{n} (\bar{y} - \bar{X} \beta^*)^\top \bar{X} (\hat{\beta} - \beta^*) + \lambda_1 (\| \tilde{\Gamma} \beta^* \|_1 - \| \tilde{\Gamma} \beta \|_1).
\]

For the remainder of the proof let \( \Delta := \hat{\beta} - \beta^* \). Then
\[
\frac{1}{n} \| \tilde{X} \Delta \|_2^2 \leq \frac{2}{n} (\bar{y} - \bar{X} \beta^*)^\top \bar{X} \Delta + \lambda_1 (\| \tilde{\Gamma} \beta^* \|_1 - \| \tilde{\Gamma} \beta \|_1).
\]

First we control the term \((\bar{y} - \bar{X} \beta^*)^\top \tilde{X} \Delta\). Using basic algebra,
\[
(\bar{y} - \bar{X} \beta^*)^\top \tilde{X} \Delta = \epsilon^\top X \Delta - n \lambda_S \beta^* \Gamma^\top \Gamma \Delta.
\]

Since \( \tilde{\Gamma}^\dagger = I_{p \times p} \), where \( \tilde{\Gamma}^\dagger \) is the pseudo-inverse of \( \tilde{\Gamma} \). Therefore
\[
\epsilon^\top X \Delta \leq \| (X \tilde{\Gamma}^\dagger)^\top \epsilon \|_\infty \| \tilde{\Gamma} \Delta \|_1.
\]

We next bound \( n \lambda_S \beta^* \Gamma^\top \Gamma \Delta \) by
\[
n \lambda_S \beta^* \Gamma^\top \Gamma \Delta \leq n \lambda_S \| \Gamma^\top \beta^* \|_\infty \| \Delta \|_1
\]

\[
\leq n \lambda_S \| L \beta^* \|_\infty \| \tilde{\Gamma} \Delta \|_1
\]

\[
\leq n \frac{\lambda_1}{2} \| \tilde{\Gamma} \Delta \|_1,
\]

where the last inequality follows from the constraint that \( \lambda_1 \geq 8 \lambda_S \| L \beta^* \|_\infty \). Now recall the constraint \( \lambda_1 \geq 48 \rho \sigma \sqrt{\frac{c_u \log p}{n}} \), the following lemma shows that with high probability we have \( \lambda_1 \geq \frac{8}{n} \| (X \tilde{\Gamma}^\dagger)^\top \epsilon \|_\infty \).

**Lemma 8.** Suppose we have \( \lambda_1 \geq 48 \rho \sigma \sqrt{\frac{c_u \log p}{n}} \). Then with probability at least \( 1 - \frac{C_1}{p} \),
\[
\lambda_1 \geq \frac{8}{n} \| (X \tilde{\Gamma}^\dagger)^\top \epsilon \|_\infty
\]

for absolute constant \( C_1 > 0 \).

The proof of Lemma 8 is deferred to the Appendix. Combining the constraints for \( \lambda_1 \) with the inequalities above,
\[
\frac{2}{n} (\bar{y} - \bar{X} \beta^*)^\top \tilde{X} \Delta \leq \lambda_1 \frac{1}{4} \| \tilde{\Gamma} \Delta \|_1 + \frac{1}{4} \| \tilde{\Gamma} \Delta \|_1 = \frac{\lambda_1}{2} \| \tilde{\Gamma} \Delta \|_1.
\]

Putting these pieces together we have
\[
\frac{1}{n} \| \tilde{X} \Delta \|_2^2 \leq \frac{\lambda_1}{2} (\| \tilde{\Gamma} \Delta \|_1 + 2 \| \tilde{\Gamma} \beta^* \|_1 - 2 \| \tilde{\Gamma} \beta \|_1).
\]

Furthermore by the triangle inequality and the fact that \( \frac{1}{n} \| \tilde{X} \Delta \|_2^2 \geq 0 \) we have
\[
0 \leq \| \tilde{\Gamma} (\hat{\beta} - \beta^*) \|_1 + 2 \| \tilde{\Gamma} \beta^* \|_1 - 2 \| \tilde{\Gamma} \beta \|_1 \leq 3 \| (\tilde{\Gamma} \Delta)_{\mathcal{T}^\circ} \|_1 + 4 \| (\tilde{\Gamma} \beta^*)_{\mathcal{T}^\circ} \|_1.
\]

Therefore \( \Delta \) lies in the translated cone
\[
\mathcal{C} := \{ v : \| (\tilde{\Gamma} v)_{\mathcal{T}^\circ} \|_1 \leq 3 \| (\tilde{\Gamma} v)_{\mathcal{T}^\circ} \|_1 + 4 \| (\tilde{\Gamma} \beta^*)_{\mathcal{T}^\circ} \|_1 \}.
\]
Moreover by the definition of $k_T$ we have

$$\|(\tilde{\Gamma} \Delta)_T\|_1 \leq \frac{\sqrt{|T|}\|\Delta\|_2}{k_T};$$

from (11) we have

$$\frac{1}{2n}\|\tilde{X} \Delta\|_2^2 \leq \lambda_1\|((\tilde{\Gamma} \beta^*)_T)_c\|_1 + \frac{3\lambda_1}{4k_T}\sqrt{|T|}\|\Delta\|_2.$$

(13)

### 7.1.1. Restricted Eigenvalue Condition for $\tilde{X}$

From (11) and (12) we need to lower bound

$$\frac{\|\tilde{X} \Delta\|_2^2}{n} = \Delta^\top \left(\frac{X^\top X}{n} + \lambda_S L\right) \Delta,$$

for all $\Delta$ belonging to the cone $C$ defined in (12). The result is stated in the following lemma:

**Lemma 9.** For all $\Delta$ belonging to the cone defined in (12) if we have

$$\lambda_1 \leq c_2 \sqrt{\frac{\lambda_{\min}(\Sigma + \lambda_S L)}{|T|} k_T},$$

(14)

then

$$\Delta^\top \left(\frac{X^\top X}{n} + \lambda_S L\right) \Delta \geq c_1 \lambda_{\min}(\Sigma + \lambda_S L)\|\Delta\|_2^2 - c_3 \lambda_1^2\|(\tilde{\Gamma} \beta^*)_T\|_1^2$$

holds with probability at least $1 - c_4 \exp(-c_5 n)$, where $c_i > 0$ for $i = 1, ..., 5$ are positive constants.

The proof for this lemma is is based on a technique used in Raskuti et al. (2010).

### 7.1.2. Final Part for Proof

From (13) and (15),

$$c_1 \lambda_{\min}(\Sigma + \lambda_S L)\|\Delta\|_2^2 - c_3 \lambda_1^2\|(\tilde{\Gamma} \beta^*)_T\|_1^2 \leq 2\lambda_1\|(\tilde{\Gamma} \beta^*)_T\|_1 + \frac{3\lambda_1}{2k_T}\sqrt{|T|}\|\Delta\|_2,$$

which is a quadratic inequality involving $\|\Delta\|_2$ as follows:

$$a\|\Delta\|_2^2 - b\|\Delta\|_2 - c \leq 0$$

with

$$a = 1,$$

$$b = \frac{3\lambda_1\sqrt{|T|}}{2c_1 k_T \lambda_{\min}(\Sigma + \lambda_S L)},$$

$$c = \frac{1}{c_1 \lambda_{\min}(\Sigma + \lambda_S L)}(2\lambda_1\|(\tilde{\Gamma} \beta^*)_T\|_1 + c_3 \lambda_1^2\|(\tilde{\Gamma} \beta^*)_T\|_1^2).$$
By solving this quadratic inequality,
\[ \|\Delta\|_2^2 \leq 4 \max\{b^2, |c|\}. \]

Therefore these exists a positive constant \( C_u \) such that
\[ \|\hat{\beta} - \beta^*\|_2^2 \leq C_u \max \left\{ \frac{\lambda^2_1|T|}{k_T^2 \lambda^2_{\min}(\Sigma + \lambda_s L)}, \frac{\lambda_1 |(\hat{\Gamma}\beta^*)_T||_1 + \lambda^2_T |(\hat{\Gamma}\beta^*)_T|_1^2}{\lambda_{\min}(\Sigma + \lambda_s L)} \right\}. \]

Note that the above inequality is true for all \( T \), thus
\[ \|\hat{\beta} - \beta^*\|_2^2 \leq C_u \min_{T} \max \left\{ \frac{\lambda^2_1|T|}{k_T^2 \lambda^2_{\min}(\Sigma + \lambda_s L)}, \frac{\lambda_1 |(\hat{\Gamma}\beta^*)_T||_1 + \lambda^2_T |(\hat{\Gamma}\beta^*)_T|_1^2}{\lambda_{\min}(\Sigma + \lambda_s L)} \right\}. \]

This completes the proof.

### 7.2. Proof of Theorem 2

The upper bound result \( \|\hat{\beta} - \beta^*\|_2^2 \) stated in Theorem 1 holds for all choices of \( T \). If we choose \( T = \text{supp}(\hat{\Gamma}\beta^*) \) then \( |(\hat{\Gamma}\beta^*)_T|_1 = 0 \) and by Lemma 2,
\[ k_T^{-1} \leq \frac{\lambda_{TV} \sqrt{2||\hat{\Sigma}||_{1,1}||\Gamma\beta^*||_0 + \sqrt{||\beta^*||_0}}}{\sqrt{||\Gamma\beta^*||_0 + ||\beta^*||_0}}. \]

Then by Theorem 1 we have
\[ \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{2C_u}{\lambda^2_{\min}(\Sigma + \lambda_s L)} (\lambda^2_1 ||\beta^*||_0 + 2\lambda^2_T \lambda_{TV} ||\hat{\Sigma}||_{1,1}||\Gamma\beta^*||_0). \quad (16) \]

On the other hand if we choose \( T = \text{supp}(\beta^*), \ |(\hat{\Gamma}\beta^*)_T||_1 = \lambda_{TV} ||\Gamma\beta^*||_1 \) and by Lemma 2, \( k_T^{-1} \leq 1 \). Thus if \( \lambda_1 \lambda_{TV} ||\Gamma\beta^*||_1 \leq 1 \) by Theorem 1
\[ \|\hat{\beta} - \beta^*\|_2^2 \leq C_u \left( \frac{\lambda^2_1 ||\beta^*||_0}{\lambda^2_{\min}(\Sigma + \lambda_s L)} + \frac{2\lambda_1 \lambda_{TV} ||\Gamma\beta^*||_1}{\lambda_{\min}(\Sigma + \lambda_s L)} \right). \quad (17) \]

Theorem 2 follows by combining (16) and (17) and taking the minimum over these two choices of \( T \).

### 7.3. Proof of Lemma 1

First note that
\[ \lambda_{\min}(\Sigma + \lambda_s L) = \lambda_{\min}((1 - \lambda_s)\Sigma + \lambda_s(\Sigma + L)) \geq (1 - \lambda_s)\lambda_{\min}(\Sigma) + \lambda_s \lambda_{\min}(\Sigma + L) \]

where the second inequality follows from Weyl’s inequality. For the remainder of the proof, we bound \( \lambda_{\min}(\Sigma + L) \). Recall that
\[ \Sigma + L = \Sigma - \hat{\Sigma} + D \quad (18) \]
where $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix with

$$D_{jj} = \sum_{k=1}^{p} |\hat{\Sigma}_{j,k}|, \ 1 \leq j \leq p.$$ 

Then

$$\lambda_{\text{min}}(\Sigma + L) = \lambda_{\text{min}}(\Sigma - \hat{\Sigma} + D) \geq \lambda_{\text{min}}(\Sigma - \hat{\Sigma}) + \lambda_{\text{min}}(D)$$

by Weyl’s inequality. Since

$$\lambda_{\text{min}}(\Sigma - \hat{\Sigma}) = -\lambda_{\text{max}}(\hat{\Sigma} - \Sigma) \geq -\|\Sigma - \hat{\Sigma}\|_{op} \geq -\|\Sigma - \hat{\Sigma}\|_{1,1}.$$ 

Hence

$$\lambda_{\text{min}}(\Sigma + L) \geq \lambda_{\text{min}}(D) - \|\Sigma - \hat{\Sigma}\|_{1,1} \geq -\min_{j} \sum_{k=1}^{p} |\hat{\Sigma}_{j,k}| - \frac{c_\ell}{4} (\text{by Assumption 3.3})$$

$$\geq \min_{j} \left[ \sum_{k=1}^{p} |\Sigma_{j,k}| - \sum_{k=1}^{p} |\Sigma_{j,k} - \hat{\Sigma}_{j,k}| - \frac{c_\ell}{4} \right]$$

$$\geq \min_{j} \sum_{k=1}^{p} |\Sigma_{j,k}| - \max_{j} \sum_{k=1}^{p} |\Sigma_{j,k} - \hat{\Sigma}_{j,k}| - \frac{c_\ell}{4}$$

$$\geq c_\ell - \frac{c_\ell}{4} - \frac{c_\ell}{4} = \frac{c_\ell}{2} (\text{by Assumptions 3.2 and 3.3}).$$

### 7.4. Proof of Lemma 2

By the definition of $k_T$ we have

$$\sqrt{|T|}k_T^{-1} = \sup_{\beta} \frac{\| (\hat{\Gamma}\beta) T \|_1}{\| \beta \|_2}$$

$$= \sup_{\beta: \| \beta \|_2 = 1} \| (\hat{\Gamma}\beta) T \|_1$$

$$= \sup_{\beta: \| \beta \|_2 = 1} \lambda_{TV} \| (\Gamma\beta) T_1 \|_1 + \| \beta T_2 \|_1$$

$$\leq \sup_{\beta: \| \beta \|_2 = 1} \lambda_{TV} \| (\Gamma\beta) T_1 \|_1 + \sqrt{|T_2|} \| \beta \|_2$$

$$\leq \sup_{\beta: \| \beta \|_2 = 1} \lambda_{TV} \| (\Gamma\beta) T_1 \|_1 + \sqrt{|T_2|}.$$
Next we will bound the term \( \| (\Gamma \beta)_{T_1} \|_1 \). First note that
\[
\| (\Gamma \beta)_{T_1} \|_1 \leq \sqrt{|T_1| (\Gamma \beta)_{T_1} \|_2 \leq |T_1| \sum_{j,k \in E \cap T_1} |\Sigma_{j,k}| |\beta_j - \text{sign}(\Sigma_{j,k})\beta_k|^2 \leq |T_1| \sum_{j,k \in E \cap T_1} \left| \frac{\Sigma_{j,k}}{\beta_k} \right| (2|\beta_j|^2 + 2|\beta_k|^2) \leq |T_1| \sum_{j,k \in E \cap T_1} \left( \sum_{k: (j,k) \in E \cap T_1} 2|\Sigma_{j,k}| \right) |\beta_j|^2 \leq \sqrt{|T_1|} \max_{1 \leq j \leq p} \left[ \left( \sum_{k: (j,k) \in E \cap T_1} 2|\Sigma_{j,k}| \right) \right] \sqrt{\sum_{j=1}^p |\beta_j|^2} \leq \sqrt{|T_1|} \sqrt{2\|\Sigma\|_1,1}. \]
Thus
\[
k_T^{-1} \leq \frac{\lambda_{TV} \sqrt{2\|\Sigma\|_1,1|T_1| + \sqrt{|T_2|}}}{\sqrt{|T_1| + |T_2|}}, \]
which completes the proof.

**7.5. Proof of Lemma 3**

Note that \( \Gamma \) is the edge incidence matrix and \( L = \Gamma^\top \Gamma \) is the weighted graph Laplacian matrix. Let the singular value decomposition for \( \Gamma \) be \( \Gamma = U_{m \times p} D_{p \times p} V_{p \times p}^\top \). Next recall that \( \tilde{\Gamma} = \begin{bmatrix} \lambda_{TV} \Gamma \\ I \end{bmatrix} \), then we have
\[
\tilde{\Gamma} = \begin{bmatrix} \lambda_{TV} \Gamma^\top + I & \lambda_{TV} \Gamma^\top \\ I & I \end{bmatrix}^{-1} \begin{bmatrix} \lambda_{TV} \Gamma \\ I \end{bmatrix} = \begin{bmatrix} \lambda_{TV} \Gamma V D^2 V^\top + I & \lambda_{TV} \Gamma V D U^\top \\ V (\lambda_{TV} D^2 + I)^{-1} V^\top & I \end{bmatrix} = \begin{bmatrix} \lambda_{TV} \Gamma U V D^2 + I & \lambda_{TV} \Gamma U V D U^\top \\ V (\lambda_{TV} D^2 + I)^{-1} V^\top & I \end{bmatrix} = \begin{bmatrix} A \lambda_{TV} \Gamma \Gamma \Gamma + I & A \lambda_{TV} \Gamma \Gamma \Gamma^\top B \\ B \lambda_{TV} \Gamma \Gamma \Gamma + I & B \lambda_{TV} \Gamma \Gamma \Gamma^\top B \end{bmatrix} \]

From the definition of \( \rho \) we can see that the maximum diagonal entry of \( (\tilde{\Gamma})^\top \tilde{\Gamma} \) will just be \( \rho^2 \). Since
\[
(\tilde{\Gamma})^\top \tilde{\Gamma} = \begin{bmatrix} A^\top A & A^\top B \\ B^\top A & B^\top B \end{bmatrix}, \]
we need to find the maximum diagonal values for matrices \( A^\top A \) and \( B^\top B \).

Suppose there are \( K \) connected components in the associated graph \( G \). Thus the weighted graph Laplacian matrix \( L \) is block diagonal, as is the matrix \( V \) (after appropriate permutation of rows and columns), with each block corresponding to a different connected components. That is, each of the \( K \) connected components of the graph has its own weighted graph Laplacian \( L_k = V_k D_k^2 V_k^\top \), for \( k = 1, \ldots, K \) and the diagonal blocks of \( V \) are the \( V_k \)s. Let \( \mu_k \) be the minimum nonzero eigenvalue of \( L_k \). Let \( B_k \) be the subset of vertices in the \( k \)-th connected component and \( |B_k| \) be the number of vertices in that component, and let \( k(i) \) denote
which block contains vertex $i$. Now let $v_i^T$ be the $i^{th}$ row of $V$, $u_i^T$ be the $i^{th}$ row of $U$, and note that $v_i$ is only supported on $B_{k(i)}$. Further note that the first (upper left) element of the $k$-th diagonal block of $V$ is $1/\sqrt{|B_k|}$ if the minimum eigenvalue of $L_k$ is 0. Then we have:

$$B^T B = V (\lambda_{TV}^2 D^2 + I)^{-2} V^T,$$

and then the maximum diagonal element for $B^T B$ can be upper bounded as:

$$\text{max} \text{ diag} (B^T B) = \max_{i \in \{1, \ldots, p\}} v_i^T (\lambda_{TV}^2 D^2 + I)^{-2} v_i$$

$$= \max_{i \in \{1, \ldots, p\}} \max_{j \in B_{k(i)}} \sum_{j} v_{i,j}^2 \left( \frac{1}{\lambda_{TV}^2 D_{j,j} + 1} \right)^2$$

$$\leq \max_{i \in \{1, \ldots, p\}} \left\{ \frac{1}{|B_{k(i)}|} + \sum_{j \in B_{k(i)}} \frac{v_{i,j}^2}{D_{j,j}^2} \right\}$$

$$\leq \max_{i \in \{1, \ldots, p\}} \left\{ \frac{1}{|B_{k(i)}|} + \sum_{j \in B_{k(i)}} \frac{v_{i,j}^2}{D_{j,j}^2} \right\}$$

(19)

On the other hand we note that

$$A^T A = U \lambda_{TV}^2 D^2 (\lambda_{TV}^2 D^2 + I)^{-2} U^T,$$

similarly the maximum diagonal element for $A^T A$ can be upper bounded as:

$$\text{max} \text{ diag} (A^T A) = \max_{i \in \{1, \ldots, m\}} u_i^T \lambda_{TV}^2 D^2 (\lambda_{TV}^2 D^2 + I)^{-2} u_i$$

$$= \max_{i \in \{1, \ldots, m\}} \sum_{j=1}^p \frac{u_{i,j}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1}$$

$$= \max_{i \in \{1, \ldots, m\}} \sum_{j=1}^p \frac{u_{i,j}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1}$$

(20)

Then by combining the results above we have

$$\rho^2 \leq \max_{1 \leq k \leq K} \left\{ \frac{1}{|B_k|} + \frac{1}{\lambda_{TV}^2 \mu_k + 1} + \frac{1}{\lambda_{TV}^2 \mu_k + 1} \right\}$$

$$\leq \max_{1 \leq k \leq K} \left\{ \frac{1}{|B_k|} + \frac{2}{\lambda_{TV}^2 \mu_k + 1} \right\}.$$

This completes the proof of Lemma 3.
7.6. Proof of Lemma 4

By the definition of the block complete graph in Section 3.2.1 we can see that $|B_k| = \frac{p}{K}$ for $1 \leq k \leq K$ thus we have $\max_{1 \leq k \leq K} \frac{1}{|B_k|} = \frac{K}{p}$. Note that $\mu_k$ is defined to be the smallest non-zero eigenvalue of weighted graph Laplacian matrix for the $k$th complete graph. It is known that the smallest non-zero eigenvalue for un-weighted Laplacian matrix for complete graph is the number of nodes (see Hütter and Rigollet (2016, Section 4.1)). Thus, applying appropriate normalization $\mu_k = ar|B_k| = ar\frac{p}{K} = r$ since $a = \frac{K}{p}$. Hence $\mu_k = r$ for $1 \leq k \leq K$. Also note that $\lambda_{\min}(\Sigma) = a(1 - r)$, so we have

$$\lambda_{\min}(\Sigma + \lambda S L) = \lambda_{\min}[(1 - \lambda S)\Sigma + \lambda S(\Sigma + L)]$$

$$\geq (1 - \lambda S)\lambda_{\min}(\Sigma) + \lambda S\lambda_{\min}(\Sigma + L)$$

$$= (1 - \lambda S)a(1 - r) + \lambda S[a + ar\frac{p}{K} - 1)] \quad \text{(by (18) and } \hat{\Sigma} = \Sigma)$$

$$\geq (1 - \lambda S)(1 - r)\frac{K}{p} + \lambda_S r \quad \text{(by using } a = \frac{K}{p}).$$

This completes the proof of Lemma 4.

7.7. Proof of Lemma 5

Let $\Gamma$ be the edge incidence matrix for the chain graph and let $\Gamma = UDV^T$ denote the SVD of $\Gamma$. Note that the chain graph has one connected component, so in the language of Lemma 3 we have $|B_1| = p$. From Equations (20) and (21) in the proof of Lemma 3 it follows that

$$\rho^2 \leq \max \left( \frac{1}{p} + \sum_{j:D_{j,j} > 0} \frac{v_{j,i}^2}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2}, \max_{i \in \{1, \ldots, p-1\}} \sum_{j:D_{j,j} > 0} \frac{u_{j,i}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1} \right).$$

First note that if $\lambda_{TV} = 0$ then $\rho^2 \leq \frac{1}{p} + 1$ and our bound is satisfied, so for the remainder of the proof we assume $\lambda_{TV} > 0$.

**Right singular vectors:** We first bound

$$\frac{1}{p} + \sum_{j:D_{j,j} > 0} \frac{v_{j,i}^2}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2}. \quad \text{(21)}$$

The right singular vectors corresponding to the nonzero singular values are the normalized eigenvectors of the Laplacian matrix which (see Hütter and Rigollet (2016, Section B.2)) are of the form

$$v_{j,i} = \frac{\sqrt{2}}{p} \cos \left( \frac{(i + 1/2)j\pi}{p} \right)$$

so in particular, $v_{j,i}^2 \leq \frac{2}{p}$ for all $i, j$. Thus Equation (21) is

$$\leq \frac{1}{p} + \frac{2}{p} \sum_{j:D_{j,j} > 0} \frac{1}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2}.$$
The \( \frac{D_{j,j}^2}{r^2} \) are the nonzero eigenvalues of the unweighted Laplacian matrix for the path graph which are also given in Hütter and Rigollet (2016, Section B.2) as \( \sigma_j = 2 - 2 \cos \left( \frac{i \pi}{p} \right) \) for \( j = 1, \ldots, p - 1 \). We have \( 2 - 2 \cos \left( \frac{i \pi}{p} \right) \geq \frac{j^2}{p^2} \) for \( 1 \leq k \leq p - 1 \) so this is

\[
\leq \frac{1}{p} + \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{\left( \frac{r^2 \lambda_{TV}^2}{\lambda_{TV}^2} + 1 \right)^2}
= \frac{1}{p} + 2p^3 \sum_{j=1}^{p-1} \frac{1}{\left( \frac{r^2 \lambda_{TV}^2}{\lambda_{TV}^2} + p^2 \right)^2}
= \frac{1}{p} + \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{\left( \frac{r^2 \lambda_{TV}^2}{\lambda_{TV}^2} + p^2 \right)^2}.
\]

Because \( f(j) = \frac{1}{(j^2 + (r \lambda_{TV})^2)^2} \) is monotonically decreasing on \( \mathbb{R}^+ \) we get that this is

\[
\leq \frac{1}{p} + \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{\left( \frac{r^4 \lambda_{TV}^2}{\lambda_{TV}^2} \right)^2} dx
= \frac{1}{p} + \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{\left( \frac{r^4 \lambda_{TV}^2}{\lambda_{TV}^2} \right)^2} = \frac{1}{p} + \frac{\pi}{2r \lambda_{TV}}.
\]

(22)

**Left singular vectors** We next focus on bounding

\[
\sum_{j : D_{j,j}^2 > 0} \frac{u_{j,i}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1}.
\]

The \( u_j \) are the normalized eigenvectors of \( \Gamma \Gamma^T \). A computation shows that

\[
\Gamma \Gamma^T_{i,j} = \begin{cases} 
2 & \text{if } i = j \\
-1 & \text{if } |i - j| = 1 \\
0 & \text{otherwise}
\end{cases}
\]

Strang (2007, Section 1.5), gives \( p - 1 \) orthonormal eigenvectors \( u_j \) of \( \Gamma \Gamma^T \) which are of the form \( u_{j,i} = \sqrt{\frac{2}{p}} \sin (\frac{i \pi}{p}) \). In particular \( u_{j,i}^2 \leq \frac{2}{p} \) so Equation (23) is

\[
\leq \frac{2}{p} \sum_{j : D_{j,j}^2 > 0} \frac{1}{\lambda_{TV}^2 D_{j,j}^2 + 1}.
\]

(24)

As before, we have that the \( \frac{D_{j,j}^2}{r^2} \) are the nonzero eigenvalues of the unweighted Laplacian of the path graph, so they are of the form \( 2 - 2 \cos \left( \frac{i \pi}{p} \right) \) for \( j = 1, \ldots, p - 1 \) and since \( 2 - 2 \cos \left( \frac{i \pi}{p} \right) \geq \frac{j^2}{p^2} \) for \( j = 1, \ldots, p - 1 \)
we get that this is
\[
\leq \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{r^2 \lambda_{TV}^2 j^2 + 1} = 2p \sum_{j=1}^{p-1} \frac{1}{p^2 + r^2 \lambda_{TV}^2 j^2} = \frac{2p}{r^2 \lambda_{TV}^2} \sum_{j=1}^{p-1} \frac{1}{j^2 + \left(\frac{p}{r \lambda_{TV}}\right)^2}.
\]

Since \( f(j) = \frac{1}{j^2 + \left(\frac{p}{r \lambda_{TV}}\right)^2} \) is monotonically decreasing on \( \mathbb{R}^+ \) we have that this is
\[
\leq \frac{2p}{r^2 \lambda_{TV}^2} \int_{x=0}^{\infty} \frac{1}{x^2 + \left(\frac{p}{r \lambda_{TV}}\right)^2} dx
\]
\[
= \frac{2p}{r^2 \lambda_{TV}^2} \frac{r \lambda_{TV}}{p} \left. \arctan \left( \frac{r \lambda_{TV} x}{p} \right) \right|_{x=0}^{x=\infty} = \frac{2 \pi}{r \lambda_{TV}}.
\]

Moreover, since \( u_{i,j}^2 \) and \( v_{i,j}^2 \) are bounded by \( \frac{2}{p} \) we immediately have the bound
\[
\rho^2 \leq \frac{1}{p} + 2.
\]

Combining this with Equations (22) and (25) we conclude that
\[
\rho^2 \leq \min \left( \frac{1}{p} + 2, \max \left( \frac{\pi}{r \lambda_{TV}}, \frac{1}{p} + \frac{\pi}{2 r \lambda_{TV}} \right) \right) \leq \frac{1}{p} + \frac{2 \pi}{r \lambda_{TV} + 1}
\]
as claimed.

For the final part of the proof, note that \( \lambda_{\min}(\Sigma) = a[1 + 2 r \cos(\frac{p}{p+1} \pi)] \) (see Noschese, Pasquini, and Reichel (2013, Section 2)), so we have
\[
\lambda_{\min}(\Sigma + \lambda S L) = \lambda_{\min}[(1 - \lambda S)\Sigma + \lambda S(\Sigma + L)]
\]
\[
\geq (1 - \lambda S)\lambda_{\min}(\Sigma) + \lambda S \lambda_{\min}(\Sigma + L)
\]
\[
= (1 - \lambda S)a[1 + 2 r \cos(\frac{p}{p+1} \pi)] + \lambda S a(1 + r) \quad \text{(by (18) and } \hat{\Sigma} = \Sigma)
\]
\[
\geq (1 - \lambda S)[1 + 2 r \cos(\frac{p}{p+1} \pi)] + \lambda S \quad \text{(by using } a = 1)
\]
\[
\geq (1 - \lambda S)(1 - 2r) + \lambda S.
\]

This completes the proof of Lemma 5.

**7.8. Proof of Lemma 6**

Note that the lattice graph has one connected component, so in the language of Lemma 3 we have \(|B_1| = p\). Then from Equations (20) and (21) in the proof of Lemma 3 it follows that
\[
\rho^2 \leq \max \left( \frac{1}{p} + \sum_{i \in \{1, \ldots, p\}} \frac{u_{i,i}^2}{\lambda_{TV}^2 D_{i,i}^2 + 1}, \frac{1}{p} \sum_{j : D_{j,j} > 0} \frac{u_{j,j}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1}, \max_{i : D_{i,i} > 0} \frac{u_{i,i}^2}{\lambda_{TV}^2 D_{i,i}^2 + 1} \right).
\]
First note that if \( \lambda_{TV} = 0 \) then \( \rho^2 \leq \frac{1}{p} + 1 \) and our bound is satisfied, so for the remainder of the proof we assume \( \lambda_{TV} > 0 \).

**Right singular vectors** We first bound

\[
\frac{1}{p} + \sum_{j:D_{j,j} > 0} \frac{v_{j,i}^2}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2}.
\]

The \( v_j \) correspond to the normalized eigenvectors of the unweighted Laplacian for the Lattice graph. We denote the Laplacian \( L_{\text{Lat}} \). Let \( L_{\sqrt{p}} \) denote the unweighted Laplacian for the path graph with \( \sqrt{p} \) nodes. Since the Lattice graph is the direct product of two copies of the path graph, we have

\[ L_{\text{Lat}} = L_{\sqrt{p}} \otimes I_{\sqrt{p}} + I_{\sqrt{p}} \otimes L_{\sqrt{p}}. \]

Let \( \{w_k\}_{j=1}^{\sqrt{p}} \) denote the normalized eigenvectors of \( L_{\sqrt{p}} \) and \( \sigma_k \) the corresponding eigenvalues. Then

\[ L_{\text{Lat}}(w_k \otimes w_l) = L_{\sqrt{p}} w_k \otimes I_{\sqrt{p}} w_l + I_{\sqrt{p}} w_k \otimes L_{\sqrt{p}} w_l = \sigma_k w_k \otimes w_l + w_k \otimes \sigma_l w_l = (\sigma_k + \sigma_l)(w_k \otimes w_l). \]

The tensor product of unit vectors is also a unit vector, so \( \|w_k \otimes w_l\|_2 = 1 \) and \( \{w_k \otimes w_l\}_{k,l=1}^{\sqrt{p}} \) are the normalized eigenvectors \( v_j \) of \( L_{\text{Lat}} \). The \( w_k \) were given in the proof of the path graph case as \( w_{k,m} = \sqrt{\frac{2}{\sqrt{p}}} \cos \left( \frac{\pi m + 1/2}{\sqrt{p}} k \right) \) so in particular we have \( v_{j,i}^2 \leq \frac{4}{p} \). Therefore Equation (26) is

\[
\leq \frac{1}{p} + \frac{4}{p} \sum_{j:D_{j,j} > 0} \frac{1}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2}.
\]

We have \( \frac{D_{j,j}^2}{r^2} = \lambda_j \) where \( \lambda_j \) denotes the \( j \)th eigenvalue of \( L_{\text{Lat}} \). We concluded above that the eigenvalues of \( L_{\text{Lat}} \) are of the form \( \sigma_k + \sigma_l \) where \( \{\sigma_k\}_{k=0}^{\sqrt{p}-1} \) are the eigenvalues of \( L_{\sqrt{p}} \). From the path graph proof, we know these are of the form

\[ \sigma_k + \sigma_l = 4 - 2 \cos \left( \frac{\pi k}{\sqrt{p}} \right) - 2 \cos \left( \frac{\pi l}{\sqrt{p}} \right) \geq \frac{k^2 + l^2}{p} \]

for \( k, l = 0, \ldots, \sqrt{p} - 1 \). Thus

\[
\frac{1}{p} + \frac{4}{p} \sum_{j:D_{j,j} > 0} \frac{1}{(\lambda_{TV}^2 D_{j,j}^2 + 1)^2} \leq \frac{1}{p} + \frac{4}{p} \sum_{k=0}^{\sqrt{p}} \sum_{l=0}^{\sqrt{p}} \frac{1}{(r^2 \lambda_{TV}^2 k^2 + l^2 + p)^2}.
\]

Algebraic rearrangement gives that this is

\[
= \frac{1}{p} + \frac{4p}{r^4 \lambda_{TV}^2} \sum_{k=1}^{\sqrt{p}} \sum_{l=1}^{\sqrt{p}} \frac{1}{(k^2 + l^2 + \frac{p}{r^2 \lambda_{TV}^2})^2} + \frac{8p}{r^4 \lambda_{TV}^2} \sum_{k=1}^{\sqrt{p}} \frac{1}{(k^2 + \frac{p}{r^2 \lambda_{TV}^2})^2}.
\]

35
The above functions are monotonically decreasing in \( k \) and \( l \) for \( k, l \geq 0 \) and so we can say this is

\[
\leq \frac{1}{p} + \frac{4p}{r^4 \lambda^4_{TV}} \int_{x=0}^{\infty} \int_{y=0}^{\infty} \frac{1}{(x^2 + y^2 + \frac{p}{r^2 \lambda^2_{TV}})^2} dydx + \frac{8p}{r^4 \lambda^4_{TV}} \int_{x=0}^{\infty} \frac{1}{(x^2 + \frac{p}{r^2 \lambda^2_{TV}})^2} dx
\]

\[
= \frac{1}{p} + \frac{4p}{r^4 \lambda^4_{TV}} \frac{\pi r^2 \lambda^2_{TV}}{4p} + \frac{8p}{r^4 \lambda^4_{TV}} \frac{\pi r^3 \lambda^3_{TV}}{4p^{3/2}} = \frac{1}{p} + \frac{4\pi}{r^2 \lambda^2_{TV}} + \frac{8\pi}{r \lambda_{TV} \sqrt{p}}.
\]

(27)

**Left singular vectors**

We next focus on bounding \( \sum_{j:D_{j,j}^2 > 0} u^2_{j,i} \lambda^2_{TV} D_{j,j}^2 + 1 \).

The \( u_j \) are the normalized eigenvectors of \( \Gamma \Gamma^T \). The eigenvectors of this matrix are nontrivial to derive, but Wang et al. (2016) finds them in their proof of Corollary 8. Moreover, they show that after normalizing the eigenvectors, each entry is bounded by \( \sqrt{\frac{4}{p}} \). In particular, we have \( u^2_{j,i} \leq \frac{4}{p} \) for all \( i, j \) and so Equation (28) is

\[
\leq \frac{4}{p} \sum_{j:D_{j,j}^2 > 0} \frac{1}{\lambda^2_{TV} D_{j,j}^2 + 1}.
\]

(28)

As in the right singular vector case, the \( \frac{D_{j,j}^2}{r^2} \) are the eigenvalues of the unweighted Laplacian for the lattice graph, so they are of the form

\[
\sigma_k + \sigma_l = 4 - 2 \cos\left(\frac{\pi k}{\sqrt{p}}\right) - 2 \cos\left(\frac{\pi l}{\sqrt{p}}\right) \geq \frac{k^2 + l^2}{p}
\]

for \( k, l = 0, \ldots, \sqrt{p} - 1 \). Thus

\[
\frac{4}{p} \sum_{j:D_{j,j}^2 > 0} \frac{1}{\lambda^2_{TV} D_{j,j}^2 + 1} \leq \frac{4}{p} \sum_{k=0}^{\sqrt{p}} \sum_{l=0}^{\sqrt{p}} \frac{1}{r^2 \lambda^2_{TV} (k^2 + l^2) + \frac{p}{r^2 \lambda^2_{TV}}}.
\]

Algebraic manipulation gives that this is

\[
= \frac{4}{r^2 \lambda^2_{TV}} \sum_{k=1}^{\sqrt{p}} \sum_{l=1}^{\sqrt{p}} \frac{1}{r^2 \lambda^2_{TV} (k^2 + l^2) + \frac{p}{r^2 \lambda^2_{TV}}} + 8 \sum_{k=1}^{\sqrt{p}} \frac{1}{r^2 \lambda^2_{TV} k^2 + p}
\]

\[
= \frac{4}{r^2 \lambda^2_{TV}} \sum_{k=1}^{\sqrt{p}} \sum_{l=1}^{\sqrt{p}} \frac{1}{k^2 + l^2 + \frac{p}{r^2 \lambda^2_{TV}}} + \frac{8}{r^2 \lambda^2_{TV}} \sum_{k=1}^{\sqrt{p}} \frac{1}{k^2 + \frac{p}{r^2 \lambda^2_{TV}}}.
\]

And now we use an integral comparison as before to conclude that this is

\[
\leq \frac{4}{r^2 \lambda^2_{TV}} \int_{x=0}^{\sqrt{p}} \int_{y=0}^{\infty} \frac{1}{x^2 + y^2 + \frac{p}{r^2 \lambda^2_{TV}}} dydx + \frac{8}{r^2 \lambda^2_{TV}} \int_{x=0}^{\infty} \frac{1}{x^2 + \frac{p}{r^2 \lambda^2_{TV}}} dx
\]

\[
= \frac{2\pi}{r^2 \lambda^2_{TV}} \int_{x=0}^{\sqrt{p}} \frac{1}{\sqrt{x^2 + \frac{p}{r^2 \lambda^2_{TV}}}} dx + \frac{8}{r^2 \lambda^2_{TV}} \frac{\pi r \lambda_{TV}}{2\sqrt{p}}.
\]

36
We compute this last integral explicitly as
\[
\int_{x=0}^{\sqrt{p}} \frac{1}{\sqrt{x^2 + \frac{p}{r^2\lambda_{TV}^2}}} \, dx = \frac{\pi}{2} \log(\sqrt{\frac{p}{r^2\lambda_{TV}^2} + x^2 + x}) \bigg|_{x=0}^{x=\sqrt{p}} = \frac{\pi}{2} \log \left( \frac{\sqrt{\frac{p}{r^2\lambda_{TV}^2} + p + \sqrt{p}}}{\sqrt{\frac{p}{r^2\lambda_{TV}^2}}} \right).
\]

Some additional algebra, along with the fact that \(\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}\) for \(a, b \geq 0\) gives that this is
\[
\leq \frac{\pi}{2} \log(2 + \sqrt{r\lambda_{TV}}).
\]

Overall we’ve concluded that Equation (29) is
\[
\leq \frac{\pi^2 \log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2} + \frac{8\pi}{r\lambda_{TV}\sqrt{p}}.
\]

Moreover, since \(u^2_{i,j}\) and \(v^2_{i,j}\) are bounded by \(\frac{4}{p}\) we immediately have the bound
\[
\rho^2 \leq 5.
\]

Combining this with Equations (27) and (29) we conclude that
\[
\rho^2 \leq \min \left( 5, \frac{1}{p} + \frac{4\pi \log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2} + \frac{8\pi}{r\lambda_{TV}\sqrt{p}} \right) \leq \frac{1}{p} + \frac{5\pi \log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2 + 1} + \frac{10\pi}{r\lambda_{TV}\sqrt{p} + 1}
\]
as claimed.

For the final part of the proof recall that \(r \in (0, \frac{1}{4})\). Thus \(\Sigma\) is diagonally dominant with \(\Sigma_{i,i} - \sum_{j \neq i} \Sigma_{i,j} \geq 1 - 4r > 0\) for all \(i\) and therefore \(\lambda_{\min}(\Sigma) \geq 1 - 4r\). This implies that
\[
\lambda_{\min}(\Sigma + \lambda_{S}L) = \lambda_{\min}[(1 - \lambda_{S})\Sigma + \lambda_{S}(\Sigma + L)]
\geq (1 - \lambda_{S})\lambda_{\min}(\Sigma) + \lambda_{S}\lambda_{\min}(\Sigma + L)
= (1 - \lambda_{S})(1 - 4r) + \lambda_{S}(1 + 2r)\) (by (18) and \(\hat{\Sigma} = \Sigma\))
\geq (1 - \lambda_{S})(1 - 4r) + \lambda_{S}.
\]

This completes the proof of Lemma 6.

7.9. Proof of Lemma 7

We rely heavily on Theorem 1.3 in Friedman (2004) which we summarize as follows. One can construct a random d-regular graph by randomly choosing \(d\) perfect matchings on the \(p\) vertices. However, the resulting graph is not necessarily simple: it can have multiple edges.

Let \(\lambda_i(A)\) denote the \(i\)th largest eigenvalue of a matrix \(A\) and fix \(\epsilon > 0\). Theorem 1.3 in Friedman (2004) states that the adjacency matrix \(A\) corresponding to a graph constructed as in the previous paragraph will satisfy \(\lambda_2(A) \leq 2\sqrt{d - 1} - \epsilon\) with probability at least \(1 - c_\epsilon p^{-c_d}\) for constants \(c_\epsilon, c_d\) depending only on \(\epsilon\) and \(d\) respectively.

We aim to sample uniformly from the set of all simple d-regular graphs. Kim and Vu (2006) notes that we can do this by repeatedly choosing regular graphs according to the perfect matching procedure above until
we pick one which is simple. Moreover, Kim and Vu (2006) says that the probability of choosing a simple graph via this procedure can be bounded below by a constant depending only on $d$ but independent of $p$. Thus, the spectral gap result above will also hold for a randomly chosen simple random $d$-regular graph.

Observe that each connected component of the graph gives rise to an eigenvector of $A$ having all ones on the connected component and all zeros elsewhere. This eigenvector has eigenvalue $d$ and we have $d > 2\sqrt{d - 1}$ for $d > 2$. Thus this result implies that $|B_1| = p$ with probability at least $1 - c_d p^{-c_d}$.

The weighted Laplacian $L$ can be written as $L = rdI - rA$ where $A$ denotes the unweighted adjacency matrix of the graph. We have $\lambda_i(L) = rd - r\lambda_{p+1-i}(A)$ for $i = 1, \ldots, p$. Then $\lambda_p(L)$, the smallest eigenvalue of the Laplacian, can be written as $\lambda_p(L) = rd - r\lambda_1(A)$. Since the all ones vector is an eigenvector of $A$ with eigenvalue $d$, and $d$ is the largest possible eigenvalue of $A$, we have $\lambda_1(A) = d$. Thus $\lambda_p(L) = 0$.

By the spectral gap result from Friedman (2004) we have $\lambda_2(A) \leq 2\sqrt{d - 1} - \epsilon$ with high probability. For simplicity we use that $2\sqrt{d - 1} - \epsilon \leq \frac{\partial d}{7}$ for sufficiently small $\epsilon$ and $d > 2$. Therefore we can state the result from Friedman (2004) as guaranteeing $\lambda_2(A) \leq \frac{\partial d}{7}$ with probability at least $1 - p^{-c_d}$ for a constant $c_d$ depending only on $d$. Thus $\lambda_{p-1}(L)$, the smallest nonzero eigenvalue of the laplacian, satisfies

$$\lambda_{p-1}(L) = rd - r\lambda_2(A) \geq \frac{dr}{7}$$

with probability at least $1 - p^{-c_d}$. We can now apply Lemma 3 to get the bound on $\rho$.

Finally, the result on $\lambda_{\min}(\Sigma + \lambda_s L)$ follows exactly as in the proof of Lemma 6.

**7.10. Proof of Lemma 8**

We will use two classical Lemmas for Gaussian processes Anderson (1984); Slepian (1962) to prove our results.

**Lemma 10** (Anderson’s comparison inequality). Let $X$ and $Y$ be zero-mean Gaussian random vectors with covariance $\Sigma_X$ and $\Sigma_Y$ respectively. If $\Sigma_Y - \Sigma_X$ is positive semi-definite then for any convex symmetric set $C$,

$$P(X \in C) \leq P(Y \in C).$$

**Lemma 11** (Slepian’s Lemma). Let $\{G_s, s \in S\}$ and $\{H_s, s \in S\}$ be two centered Gaussian processes defined over the same index set $S$. Suppose that both processes are almost surely bounded. For each $s, t \in S$, if $E[(G_s - G_t)^2] \leq E[(H_s - H_t)^2]$, then $E[\sup_{s \in S} G_s] \leq E[\sup_{s \in S} H_s]$. Further if $E(G_s^2) = E(H_s^2)$ for all $s \in S$, then

$$P\{\sup_{s \in S} G_s > x\} \leq P\{\sup_{s \in S} H_s > x\},$$

for all $x > 0$.

$$(X^\dagger Y)^\epsilon = \sum_{i=1}^n (\tilde{F}^\dagger, \epsilon_iX^{(i)})$$

and $\text{Cov}(X) = \Sigma \leq \lambda_{\max}(\Sigma)I_{p \times p}$. Then by Assumption 3.1 $\lambda_{\max}(\Sigma) \leq c_u$, and if we use Lemma 10, for any $x > 0$

$$P\{\sup_{i=1}^n (\tilde{F}^\dagger, \epsilon_iX^{(i)}) \leq x\} \geq P\{\sup_{i=1}^n (\tilde{F}^\dagger, \epsilon_i g_i) \leq x\},$$

where $X^{(i)}$ is the $i$th row of matrix $X$ and $\{g_i : i = 1, \ldots, n\}$ is i.i.d. standard normal Gaussian vectors with $g_i \in \mathbb{R}^p$. Now let $G \in \mathbb{R}^p$ be an i.i.d. standard norm Gaussian vector and define the zero-mean Gaussian
Thus we have shown with high probability that \( \sup P\{\sqrt{c_u}\sum_{i=1}^{n}\langle \hat{\Gamma}^\dagger, \epsilon_i g_i \rangle \leq x \} \geq P\{\sup \sigma \sqrt{n c_u}(\hat{\Gamma}^\dagger, G) \leq x \} \).

Further, using known results on Gaussian maxima (Boucheron, Lugosi, and Massart (2013, Theorem 2.5)), \( \sup(\hat{\Gamma}^\dagger, G) \leq 3\rho \sqrt{\log(m + p)} \) with probability at least \( 1 - \frac{C_1}{p} \) for some absolute constant \( C_1 > 0 \). By choosing \( x = 3\sigma \rho \sqrt{n c_u \log(m + p)} \),
\[
P\{\sup_{i=1}^{n} \langle \hat{\Gamma}^\dagger, \epsilon_i X^{(i)} \rangle \leq x \} \geq P\{\sup \sigma \sqrt{n c_u}(\hat{\Gamma}^\dagger, G) \leq x \} \geq 1 - \frac{C_1}{p}.
\]

Thus we have shown with high probability that \( \| (X \hat{\Gamma}^\dagger)^\top \epsilon \|_{\infty} \leq 3\sigma \rho \sqrt{n c_u \log(m + p)} \). Since \( m \) is the number of edges, \( m \leq \frac{p(p-1)}{2} \), thus with probability at least \( 1 - \frac{C_1}{p} \) we have that \( \| (X \hat{\Gamma}^\dagger)^\top \epsilon \|_{\infty} \leq 6\sigma \rho \sqrt{n c_u \log p} \). This completes the proof.

### 7.11. Proof of Lemma 9

The proof of Lemma 9 involves two parts.

**Part 1:** We first show that the following inequality
\[
\frac{\| X \Delta \|_2}{\sqrt{n}} \geq \frac{1}{4} \| \Sigma^{1/2} \Delta \|_2 - \frac{9\lambda_1}{\sigma} \| \hat{\Gamma} \Delta \|_1
\]
holds with probability at least \( 1 - c_4 \exp(-c_5 n) \) by using similar techniques to those used to prove Theorem 1 in Raskutti et al. (2010).

First note that it is sufficient to show (30) holds with \( \| \Sigma^{1/2} \Delta \|_2 = 1 \). The reason is as follows: if \( \| \Sigma^{1/2} \Delta \|_2 = 0 \) we can see that (30) holds trivially; otherwise when \( \| \Sigma^{1/2} \Delta \|_2 > 0 \) we can define \( \hat{\Delta} = \frac{\Delta}{\| \Sigma^{1/2} \Delta \|_2} \) then we have \( \| \Sigma^{1/2} \hat{\Delta} \|_2 = 1 \). Since (30) is invariant with respect to the scale of \( \Delta \), if it holds for \( \hat{\Delta} \), it also holds for \( \Delta \). Thus in the following proof we just assume that \( \| \Sigma^{1/2} \Delta \|_2 = 1 \). To show (30) with \( \| \Sigma^{1/2} \Delta \|_2 = 1 \) holds there are three main steps:

1. Since we want to lower bound \( \| X \Delta \|_2 \) in terms of \( \| \Sigma^{1/2} \Delta \|_2 \) and \( \| \hat{\Gamma} \Delta \|_1 \), we define the set \( V(r) := \{ \Delta \in \mathbb{R}^p \mid \| \Sigma^{1/2} \Delta \|_2 = 1, \| \hat{\Gamma} \Delta \|_1 \leq r \} \) for a fixed radius \( r \). Note that we are only concerned with choices of \( r \) such the set \( V(r) \) is non-empty. Our first step is to give an upper bound for \( \mathbb{E}[M(r, X)] \), where \( M(r, X) \) is defined as:
\[
M(r, X) := 1 - \inf_{\Delta \in V(r)} \frac{\| X \Delta \|_2}{\sqrt{n}} = \sup_{\Delta \in V(r)} \left\{ 1 - \frac{\| X \Delta \|_2}{\sqrt{n}} \right\}.
\]

2. The second step is to use concentration inequalities to show that with high probability for each fixed \( r > 0 \), the random quantity \( M(r, X) \) is sharply concentrated around \( \mathbb{E}[M(r, X)] \).

3. The third step is to use a peeling argument to show that the analysis holds uniformly over all possible values of \( r \) with high probability, then we can show that (30) holds with high probability.

In the following proof we only provide details for proving step (1) since our proof for step (2) and (3) will be identical to those in Raskutti et al. (2010). For step (1) we prove the following lemma:
Lemma 12. For any radius $r > 0$ such that $V(r)$ is non-empty, we have

$$\mathbb{E}[M(r, X)] \leq \frac{1}{4} + 3r \frac{\lambda_1}{\sigma}.$$  

Proof. Define the Euclidean sphere of radius 1 to be $S^{n-1} = \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$. Then $\|X\Delta\|_2 = \sup_{u \in S^{n-1}} u^\top X \Delta$. In order to write the quantity $M(r, X)$ in a form that is easier to analyze, we define $Y_{u, \Delta} := u^\top X \Delta$ for each pair $(u, \Delta) \in S^{n-1} \times V(r)$. Then we have

$$- \inf_{\Delta \in V(r)} \|X \Delta\|_2 = - \inf_{\Delta \in V(r)} \sup_{u \in S^{n-1}} u^\top X \Delta = \sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u, \Delta}.$$

Next we will use a Gaussian comparison inequality to upper bound the expected value of the quantity $\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u, \Delta}$. Here we use a form of Gordon’s inequality that is stated in Davidson and Szarek (2001) for our analysis. Suppose that $\{Y_{u, \Delta}, (u, \Delta) \in U \times V\}$ and $\{Z_{u, \Delta}, (u, \Delta) \in U \times V\}$ are two zero-mean Gaussian processes on $U \times V$. We denote $\sigma(\cdot)$ to be the standard deviation of a random variable. Using Gordon’s inequality, if

$$\sigma(Y_{u, \Delta} - Y_{u', \Delta'}) \leq \sigma(Z_{u, \Delta} - Z_{u', \Delta'}), \quad \forall \ (u, \Delta) \text{ and } (u', \Delta') \in U \times V,$$

and this inequality holds with equality when $\Delta = \Delta'$, then

$$\mathbb{E}[\sup_{\Delta \in V} \inf_{u \in U} Y_{u, \Delta}] \leq \mathbb{E}[\sup_{\Delta \in V} \inf_{u \in U} Z_{u, \Delta}].$$

Now we consider the zero-mean Gaussian process $Z_{u, \Delta}$ with $(u, \Delta) \in S^{n-1} \times V(r)$ as follows:

$$Z_{u, \Delta} = g^\top u + h^\top \Sigma^{1/2} \Delta,$$

where $g \sim \mathcal{N}(0, I_{n \times n})$ and $h \sim \mathcal{N}(0, I_{p \times p})$. It follows that (see Raskutti et al. (2010) for more details)

$$\sigma(Y_{u, \Delta} - Y_{u', \Delta'}) \leq \sigma(Z_{u, \Delta} - Z_{u', \Delta'}), \quad \forall \ (u, \Delta) \text{ and } (u', \Delta') \in S^{n-1} \times V(r),$$

and the equality holds when $\Delta = \Delta'$. Thus we can apply Gordon’s inequality to conclude that

$$\mathbb{E}[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u, \Delta}] \leq \mathbb{E}[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Z_{u, \Delta}].$$

Next we bound the term $\mathbb{E}[\sup_{\Delta \in V(r)} h^\top \Sigma^{1/2} \Delta]$ using the following lemma:

Lemma 13. Suppose we have $\lambda_1 \geq 48 \rho \sigma \sqrt{c_n \log p / n}$, then we have that

$$\lambda_1 \geq 8 \frac{\sigma}{\sqrt{n}} \mathbb{E}[\|\Sigma^{1/2} \hat{\Gamma}^\top h\|_\infty].$$

with high probability at least $1 - \frac{c}{p}$ for some absolute constant $c > 0$. 

40
The proof for this lemma will be provided shortly. Thus
\[
E[\sup_{\Delta \in \mathcal{V}(r)} | h^T \Sigma^{1/2} \Delta |] = E[\sup_{\Delta \in \mathcal{V}(r)} | h^T \Sigma^{1/2} \hat{\Sigma}^T \hat{\Delta} |]
\leq E[\sup_{\Delta \in \mathcal{V}(r)} \| h^T \Sigma^{1/2} \hat{\Sigma}^T \|_\infty \| \hat{\Delta} \|_1]
\leq E[\| (\Sigma^{1/2} \hat{\Sigma}^T)^T h \|_\infty] r
\leq 3 \frac{\lambda_1}{\sigma} \sqrt{n},
\]
where the last inequality holds with high probability from Lemma 13. Also by standard $\chi^2$ tail bounds (Ledoux and Talagrand, 1991) when $n \geq 10$ we have $E[\| g \|_2] \geq \frac{3}{4} \sqrt{n}$. By combining these pieces together
\[
E[- \inf_{\Delta \in \mathcal{V}(r)} \| X \Delta \|_2] \leq - \frac{3}{4} \sqrt{n} + 3 \frac{\lambda_1}{\sigma} \sqrt{n}.
\]
Thus by dividing by $\sqrt{n}$ and adding 1 to both sides we have
\[
E[M(r, X)] = E[1 - \inf_{\Delta \in \mathcal{V}(r)} \frac{\| X \Delta \|_2}{\sqrt{n}}] \leq \frac{1}{4} + 3 \frac{\lambda_1}{\sigma}.
\]

Part 2: Next we can go to second part of the proof. From (11) and (12) we know that
\[
\| \hat{\Delta} \|_1 \leq 4 \| (\hat{\Delta})_T \|_1 + 4 \| (\hat{\Sigma}^* \beta^*)^{Tc} \|_1
\leq \frac{4 \sqrt{|T| \| \Delta \|_2}}{k_T} + 4 \| (\hat{\Sigma}^* \beta^*)^{Tc} \|_1.
\]
Then
\[
\frac{\| X \Delta \|_2}{\sqrt{n}} \geq \frac{1}{4} \| \Sigma^{1/2} \Delta \|_2 - 9 \frac{\lambda_1}{\sigma} \left( 4 \| (\hat{\Sigma}^* \beta^*)^{Tc} \|_1 + \frac{4 \sqrt{|T| \| \Delta \|_2}}{k_T} \right).
\]
Thus there exist constants $c', c'' > 0$ such that
\[
\Delta^T \left( \frac{X^T X}{n} + \lambda S L \right) \Delta \geq c' \Delta^T (\Sigma + \lambda S L) \Delta - c'' \lambda_1^2 \left( \| (\hat{\Sigma}^* \beta^*)^{Tc} \|_1^2 + \frac{|T| \| \Delta \|_2^2}{k_T^2} \right).
\]
Since
\[
\Delta^T (\Sigma + \lambda S L) \Delta \geq \lambda_{\min}(\Sigma + \lambda S L) \| \Delta \|_2^2,
\]
then when $\lambda_1$ satisfies (14) for some constant $c_2 > 0$,
\[
\Delta^T \left( \frac{X^T X}{n} + \lambda S L \right) \Delta \geq c_1 \lambda_{\min}(\Sigma + \lambda S L) \| \Delta \|_2^2 - c_3 \lambda_1^2 \| (\hat{\Sigma}^* \beta^*)^{Tc} \|_1^2,
\]
for absolute constants $c_1, c_3 > 0$. 

41
7.12. Proof of Lemma 13

Here we use similar techniques to the proof of Lemma 8. First note that $\Sigma \preceq c_u I_{p \times p}$ and by using Lemma 10 we have for any $y > 0$ the following inequality

$$P\{\sup[(\Sigma^{1/2}\widehat{\Gamma}^\dagger)\top h] \leq y\} = P\{\sup(\widehat{\Gamma}^\dagger, \Sigma^{1/2}h) \leq y\} \geq P\{\sup(\widehat{\Gamma}^\dagger, h) \leq \frac{y}{\sqrt{c_u}}\}.$$ 

Since $h \sim \mathcal{N}(0, I_{p \times p})$ then also by known results on Gaussian maxima (Boucheron et al. (2013, Theorem 2.5)) we have $\sup(\widehat{\Gamma}^\dagger, h) \leq 3\rho\sqrt{\log(m + p)}$ with probability at least $1 - \frac{c}{p}$ for some constant $c > 0$. Then we can choose $y = 3\rho\sqrt{c_u \log(m + p)}$ and

$$P\{\sup[(\Sigma^{1/2}\widehat{\Gamma}^\dagger)\top h] \leq y\} \geq P\{\sup(\widehat{\Gamma}^\dagger, h) \leq \frac{y}{\sqrt{c_u}}\} \geq 1 - \frac{c}{p}.$$ 

Thus with high probability $E[\| (\Sigma^{1/2}\widehat{\Gamma}^\dagger)\top h \|_\infty] \leq 3\rho\sqrt{c_u \log(m + p)}$, then using the fact that $m \leq \frac{p(p-1)}{2}$, $E[\| (\Sigma^{1/2}\widehat{\Gamma}^\dagger)\top h \|_\infty] \leq 6\rho\sqrt{c_u \log p}$ holds with probability at least $1 - \frac{c}{p}$. This completes the proof of Lemma 13.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Baik, J. and J. W. Silverstein (2006, July). Eigenvalues of large sample covariance matrices of spiked populations models. Journal of Multivariate Analysis 97(6), 1382–1408.

Bickel, P. and E. Levina (2008a). Covariance estimation by thresholding. The Annals of Statistics 36, 2577–2604.

Bickel, P. and E. Levina (2008b). Regularized estimation of large covariance matrices. The Annals of Statistics 36, 199–227.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37(4), 1705–1732.

Bogdan, M., E. van den Berg, W. Su, and E. Candes (2013). Statistical estimation and testing via the ordered $\ell_1$ norm. Technical Report arXiv:1310.1969.

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64(1), 115–123.

Boucheron, S., G. Lugosi, and P. Massart (2013). Concentration inequalities: A nonasymptotic theory of independence. Oxford university press.

Bühlmann, P., P. Rütimann, S. van de Geer, and C. Zhang (2013). Correlated variables in regression: clustering and sparse estimation. Journal of Statistical Planning and Inference 143(11), 1835–1858.

Cai, T. T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association 106(494), 672–684.

Cai, T. T., R. Zhao, and H. H. Zhou (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electron. J. Statist. 10(1), 1–59.

Candes, E. J. and T. Tao (2005). Decoding by linear programming. IEEE transactions on information theory 51(12), 4203–4215.

Davidson, K. R. and S. J. Szarek (2001). Local operator theory, random matrices, and Banach spaces. In Handbook of Banach Spaces, Volume 1, pp. 317–336. Amsterdam, NL: Elsevier.
Daye, Z. and X. Jeng (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis* 53(4), 1284–1298.

Donoho, D. L., M. Gavish, and I. M. Johnstone (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*.

El Anbari, M. and A. Mkhadri (2014). Penalized regression combining the $\ell_1$ norm and a correlation based penalty. *Sankhya* 76(1), 82102.

Figueiredo, M. and R. Nowak (2016). Ordered weighted $\ell_1$ regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 930–938.

Friedman, J. (2004). A proof of alon’s second eigenvalue conjecture. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 720–724.

Grosenick, L., B. Klingenberg, S. Greer, J. Taylor, and B. Knutson (2009). Whole-brain sparse penalized discriminant analysis for predicting choice. *NeuroImage* 47, S58.

Grosenick, L., B. Klingenberg, B. Knutson, and J. Taylor (2012). Interpretable whole-brain prediction analysis with graphnet. *NeuroImage* 72, 304–321.

Hallac, D., J. Leskovec, and S. Boyd (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 387–396. ACM.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hebiri, M. and S. van de Geer (2011). The smooth-lasso and other $\ell_1+\ell_2$-penalized methods. *Electronic Journal of Statistics* 5, 1184–1226.

Huang, J., S. Ma, H. Li, and C.-H. Zhang (2011, 08). The sparse laplacian shrinkage estimator for high-dimensional regression. *Ann. Statist. 39*(4), 2021–2046.

Hütter, J. and P. Rigollet (2016). Optimal rates for total variation denoising. *arXiv preprint arXiv:1603.09388*.

Jalali, A. and M. Fazel (2013). A convex method for learning d-valued models. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pp. 1123–1126. IEEE.

Jia, J., K. Rohe, et al. (2015). Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics* 9(1), 1150–1172.

Kim, J. and J. Vu (2006). Generating random regular graphs. *Combinatorica* (26), 683708.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag.

Lewis-Peacock, J. A. and B. R. Postle (2008). Temporary activation of long-term memory supports working memory. *Journal of Neuroscience* 28(35), 8765–8771.

Liu, J., L. Yuan, and J. Ye (2013). Dictionary lasso: Guaranteed sparse recovery under linear transformation. *arXiv preprint arXiv:1305.0047*.

Michel, V., A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion (2011). Total variation regularization for fmri-based prediction of behavior. *IEEE transactions on medical imaging* 30(7), 1328–1340.

Needell, D. and R. Ward (2013a). Near-optimal compressed sensing guarantees for total variation minimization. *IEEE Transactions on Image Processing* 22(10), 3941–3949.

Needell, D. and R. Ward (2013b). Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences* 6(2), 1035–1058.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science* 27(4), 538–557.

Noschese, S., L. Pasquini, and L. Reichel (2013). Tridiagonal toeplitz matrices: properties and novel appli-
Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Labs Technical Journal* 36(6), 1389–1401.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research* 11, 2241–2259.

Raskutti, G. and M. Yuan (2015). Convex regularization for high-dimensional tensor regression. *arXiv preprint arXiv:1512.01215*.

Sadhanala, V., Y. Wang, and R. Tibshirani (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pp. 3513–3521.

Saunderson, J., V. Chandrasekaran, P. Parrilo, and A. Willsky (2012). Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications* 33(4), 1395–1416.

Sharma, D. B., H. D. Bondell, and H. H. Zhang (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics* 22(2), 319–340.

Sharpnack, J., A. Singh, and A. Rinaldo (2012). Sparsistency of the edge lasso over graphs. In *Artificial Intelligence and Statistics*, pp. 1028–1036.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics* 4, 1055–1096.

Shen, X. and H.-C. Huang (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* 105(490), 727–739.

Shen, X., H.-C. Huang, and W. Pan (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika* 99(4), 899–914.

Shuman, D. I., S. K. Narang, P. Frossard, A. Ortega, and P. Vandergeynst (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* 30(3), 83–98.

Slepian, D. (1962). The one-sided barrier problem for gaussian noise. *Bell System Tech. J* 41, 463–501.

Strang, G. (2007). *Computational Science and Engineering*. Wellesley, MA: Wellesley-Cambridge Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1), 91–108.

Tibshirani, R. and J. Taylor (2011, 06). The solution path of the generalized lasso. *The Annals of Statistics* 39(3), 1335–1371.

Tutz, G. and J. Ulbricht (2009). Penalized regression with correlation-based penalty. *Statistics and Computing* 19(3), 239253.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

van de Geer, S. and P. Buhlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* 3, 1360–1392.

Wang, Y., J. Sharpnack, A. Smola, and R. Tibshirani (2016). Trend filtering on graphs. *Journal of Machine Learning Research* 17(105), 1–41.

Wauthier, F. L., N. Jojic, and M. I. Jordan (2013). A comparative framework for preconditioned lasso algorithms. In *Advances in Neural Information Processing Systems* 26, pp. 1061–1069. Curran Associates, Inc.

Witten, D., A. Shojaie, and F. Zhang (2014, 2). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* 56(1), 112–122.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009). Genome-wide association analysis by lasso
penalized logistic regression. *Bioinformatics* 25(6), 714–721.

Xin, B., Y. Kawahara, Y. Wang, and W. Gao (2014). Efficient generalized fused lasso and its application to the diagnosis of alzheimer’s disease. In *AAAI*, pp. 2163–2169. Citeseer.

Xin, B., Y. Kawahara, Y. Wang, L. Hu, and W. Gao (2016). Efficient generalized fused lasso and its applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7(4), 60.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B* (67(2)), 301–320.