Published in final edited form as:

SIAM J Math Data Sci. 2020 ; 2(2): 480–504. doi:10.1137/19M1287365.

Graph-based regularization for regression problems with alignment and highly-correlated designs[*]

Yuan Li[†,‡], **Benjamin Mark**[†,§], **Garvesh Raskutti**[‡], **Rebecca Willett**[¶], **Hyebin Song**[‡], **David Neiman**[‡]

[‡]Department of Statistics, University of Wisconsin-Madison

[§]Department of Mathematics, University of Wisconsin-Madison

[¶]Departments of Statistics and Computer Science, University of Chicago

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 many modern applications covariates are highly correlated and do not exhibit key properties (such as the restricted eigenvalue condition, restricted isometry property, or other related assumptions). This work considers a high-dimensional regression setting in which a graph governs both correlations among the covariates and the similarity among regression coefficients – meaning there is *alignment* between the covariates and 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 correlated features. This work shows how the proposed graph-based regularization yields mean-squared error guarantees for a broad range of covariance graph structures. These guarantees are optimal for many specific covariance graphs, including block and lattice graphs. Our proposed approach outperforms other methods for highly-correlated design in a variety of experiments on synthetic data and real biochemistry data.

1. Introduction

High-dimensional linear regression and inverse problems have received substantial attention over the past two decades (see Hastie et al. (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 (see Candes and Tao (2007); Bickel et al. (2009); van de Geer and Buhlmann

[*]Submitted to the editors DATE.

bmark2@wisc.edu.

[†]These authors contributed equally to the manuscript.

(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

$$y = X\beta* + \epsilon \tag{1.1}$$

where $y \in \mathbb{R}^n$ is the response variable, $X \in \mathbb{R}^{n \times p}$ is the observation or *design* matrix, and $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_{n \times n}\right)$ is Gaussian noise. Our goal is to estimate β^* based on (X, y) when X potentially has highly-correlated columns and does not satisfy standard regularity assumptions. Specifically, we define $\Sigma: = \frac{1}{n}\mathbb{E}\left[X^\top X\right]$ and consider settings where the minimum eigenvalue of Σ may be zero-valued or arbitrarily close to zero. We consider a Gaussian linear model for simplicity of exposition but our ideas and results can be extended to other settings. In Appendix K we discuss an extension to logistic regression.

Highly-correlated or dependent features arise in many modern scientific problems, including the study of enzyme thermostability (detailed in Section 1.1), genome wide association studies (GWAS) (Wu et al., 2009; Viallon et al., 2016), neuroscience (Caoa et al., 2018), climate data (Barnston and Smith, 1996; Geisler et al., 1985; DelSole and Banerjee, 2017; Mamalakis et al., 2018), and topic modeling.

As we discuss and expand upon in Section 1.4, there is alarge body of work addressing the problem of high-dimensional regression under highly correlated design (*e.g.*, Bühlmann et al. (2013); Zou and Hastie (2005)). The key challenge associated with highly-correlated columns is that estimates of β^* become very sensitive to noise and, if columns are perfectly correlated, $\beta*$ may not be identifiable, which means additional assumptions are required on β^*.

On the other hand, for many applications such as those mentioned above, there is known structure among β^* since groups of covariates often exhibit similar influence on the response. There is also a large body of work studying the high-dimensional linear model under additional assumptions on β^* including group structure (*e.g.*, Shen and Huang (2010); P. Zhao and Yu (2009)), graph structure (*e.g.*, Sharpnack et al. (2012); Hallac et al. (2015); Marial and Yu (2013); Wang et al. (2016)), and others.

In this work, we consider a case of highly correlated designs with additional structure on β^*. We use side information to generate a covariance graph and then use an *alignment* condition to ensure a corresponding graph structure on β^*. The alignment condition resolves the lack of identifiability by incorporating side information about the covariance. Importantly, we develop novel theoretical guarantees for our procedure under this alignment condition.

1.1. Motivating application: Biochemistry

In this section we apply the proposed graph total variation (GTV) methodology to an application in biochemistry, specifically protein analysis. In particular we focus on a specific protein of great interest, the cytochrome P450 enzyme, which is an important protein in a number of environments. More specifically, cytochrome P450 proteins are versatile biocatalysts which have been heavily employed for production of pharmaceutical products and synthesis of other useful compounds (Guengerich, 2002). Additionally, thermostable proteins have great industrial importance since they can withstand tough industrial process conditions (Niehaus et al., 1999). We aim to understand how 3-D structural properties of proteins are related to the thermostability of the proteins.

The dataset we use is a P450 chimeric protein dataset generated by the Romero Lab at UW-Madison*. The dataset contains thermostability measurements and features encoding the amino acid sequences and describing structural properties of 242 chimeric P450 proteins. The chimeric proteins in the dataset are created by recombining fragments of the genes of the three wild-type P450s (parent proteins) for eight blocks (Li et al., 2007). Since the amino acid sequences for the parent proteins are known, the amino acid sequence for a chimeric protein can be obtained from the recombination information for each block which parent the gene fragment is inherited from. From the amino acid sequence information, 50 features describing the structural properties of each protein were estimated by modeling 3-D structures of the chimeric enzymes via the Rosetta biomolecular modeling suite (Alford et al., 2017). A full description of the 50 structural features is provided in Table 2 in the Appendix. As our goal is to understand the relationship between the structure and thermostability of the proteins, we use a linear model where the design matrix $X \in \mathbb{R}^{n \times p}$ consists of the structure features and the response variable $y \in \mathbb{R}^n$ contains the thermostability measurements for $n = 242$ and $p = 50$.

Importantly, many of the structural features are known to be highly correlated and we use side information to estimate the covariance structure between the structural features. The side information consists of the amino acid sequences for the P450 chimeric proteins. We use the sequence, structure, and function paradigm for protein design in which a protein sequence determines the structure of the protein and the structure determines the function of the protein. In particular, we exploit the sequence-structure relationship to obtain a good estimate of the covariance matrix of the structural features. The combination of highly correlated features and side information to estimate the covariance matrix makes this problem a natural fit for out GTV methodology. More details on the estimation of the covariance and the application are provided in Section 4.

1.2. Problem formulation and proposed estimator

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

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

[*] Raw data is available at https://github.com/Jerry-Duan/Structural-features

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

$$y = X\beta^{*} + \epsilon .$$

Our goal is to estimate β^{*}. We are particularly interested in a setting where the columns of X may be highly correlated (*i.e.*, $\lambda_{\min}(\Sigma) \approx 0$), but β^{*} is well-aligned with the covariance structure (*i.e.*, correlated features have similar weights in β^{*}).

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

$$\begin{aligned}\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\|y - X\beta\|_{2}^{2} &+ \lambda_{S} \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right| \left(\beta_{j} - \hat{s}_{j,k}\beta_{k}\right)^{2} \\ &+ \lambda_{1}\left(\lambda_{\text{TV}} \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right|^{1/2} \left|\beta_{j} - \hat{s}_{j,k}\beta_{k}\right| + \|\beta\|_{1}\right),\end{aligned} \tag{1.2}$$

where λ_{S}, λ_{1} and λ_{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 $\sum_{j,k} \left|\hat{\Sigma}_{j,k}\right|^{1/2} \left|\beta_{j} - \hat{s}_{j,k}\beta_{k}\right|$ is a measure of the *graph total variation* of the signal β with respect to the graph G and $\sum_{j,k} \left|\hat{\Sigma}_{j,k}\right| \left(\beta_{j} - \hat{s}_{j,k}\beta_{k}\right)^{2}$ corresponds to a *graph Laplacian regularizer* with respect to G.

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

$$\Gamma = \sum_{\ell = 1}^{m} \Gamma_{\ell}, \qquad \text{where} \qquad \Gamma_{\ell}: = \left|\hat{\Sigma}_{j_{\ell}, k_{\ell}}\right|^{1/2} u_{\ell}\left[e_{j_{\ell}} - \operatorname{sign}\left(\hat{\Sigma}_{j_{\ell}, k_{\ell}}\right) e_{k_{\ell}}\right]^{\top} \in \mathbb{R}^{m \times p}, \tag{1.3}$$

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

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

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

$$\widetilde{X} = \widetilde{X}_{\lambda_S} := \begin{bmatrix} X \\ \sqrt{n\lambda_S}\Gamma \end{bmatrix} \in \mathbb{R}^{(n+m)\times p},\ \tilde{y} := \begin{bmatrix} y \\ \mathbf{0}_{m\times 1} \end{bmatrix} \in \mathbb{R}^{n+m},\ \text{and } \widetilde{\Gamma} := \begin{bmatrix} \lambda_{\mathrm{TV}}\Gamma \\ I_{p\times p} \end{bmatrix} \in \mathbb{R}^{(m+p)\times p}.$$

Using these definitions, we may write the estimator (1.2) equivalently as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda_S\|\Gamma\beta\|_2^2 + \lambda_1(\lambda_{\mathrm{TV}}\|\Gamma\beta\|_1 + \|\beta\|_1) \tag{1.4}$$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\|\tilde{y} - \widetilde{X}\beta\|_2^2 + \lambda_1\|\widetilde{\Gamma}\beta\|_1. \tag{1.5}$$

The three regularizers play the following roles:

- We refer to $\|\Gamma\beta\|_2^2 = \sum_{j,k}\left|\widehat{\Sigma}_{j,k}\right|\left(\beta_j - \hat{s}_{j,k}\beta_k\right)^2$ as the **Laplacian smoothing penalty**; Hebiri and van de Geer (2011) studied a variant of this regularizer with $\widehat{\Sigma}_{j,k}$ 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 $\|\Gamma\beta\|_1 = \sum_{j,k}\left|\widehat{\Sigma}_{j,k}\right|^{1/2}\left|\beta_j - \hat{s}_{j,k}\beta_k\right|$ as the **total variation penalty**, as do Shuman et al. (2013); Wang et al. (2016); Sadhanala et al. (2016); Hütter and Rigollet (2016); it is closely related to the edge LASSO penalty (Sharpnack et al., 2012). Note that these prior works consider general weighted graphs (instead of graphs defined by a covariance matrix $\widehat{\Sigma}$, 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 $\|\beta\|_1$ as the **sparsity regularizer**. The combination of the sparsity regularizer and total variation penalty amount to the fused LASSO (Tibshirani et al., 2005; Tibshirani and Taylor, 2011).

The combined effect of the three regularization terms is to find estimates of β^* which are both a good fit to the data when the columns of X are highly correlated and well-aligned with the underlying graph. This alignment structure may be desirable in a number of settings. Note that both the Laplacian smoothing and total variation penalties promote this alignment structure. We believe GTV will perform similarly on synthetic data with only one penalty included, but whether theoretical results can be derived to a variant of GTV only one of the penalties is an open question.

1.3. Contributions

This paper addresses the question of how to estimate β^* from observations in (1.1) *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 (Sharpnack et al., 2012), network LASSO (Hallac et al., 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 1.4 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 for mean-squared error (i.e. $\|\hat{\beta} - \beta^*\|_2^2$) which provide insight into the impact of the alignment of β^* with the covariance graph, and properties of the covariance graph structure such as smallest and largest block-sizes and smallest non-zero eigenvalue.
- New mean-squared error guarantees for three specific covariance graph structures, a block complete graph, a chain graph, and a lattice graph. Our error bounds match the optimal rates in the independent case where Σ is a diagonal matrix, and also match the optimal rates for the block and lattice covariance graphs.
- A simulation study which shows that our method out-performs state-of-the-art alternatives such as the Cluster Representative LASSO (CRL, Bühlmann et al. (2013)) and Ordered Weighted LASSO (OWL, Bogdan et al. (2013)) in terms of mean squared error in a variety of settings.
- A validation of our method on real biochemistry data that demonstrates the adavantages of GTV.

The remainder of this paper is organized as follows: In Section 1.4 we discuss existing work and results for this problem and its relationship to our estimator; in Section 2 we present our main theoretical results for mean-squared error; in Section 3 we carry out a simulation study by comparing our methods to other state-of-the-art methods; in Section 4 we apply our method to a real biochemistry dataset with comparison to other methods; we state our conclusions in Section 5; proofs are provided in the Appendix.

1.4. Prior work

There is a large body of work related to our proposed estimator. Significant effort has been devoted to understanding estimators like (1.4) in the special case where $X = I$ – that is, in a "denoising" setting in which observations are direct measurements of the signal of interest,

β. Variants of these estimators are often referred to as the edge or network LASSO (Sharpnack 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). 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. Hütter 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.

In the high-dimensional regression setting, our approach may be viewed as a generalization of the classical *fused LASSO* (Tibshirani et al., 2005), where instead of promoting alignment between features with adjacent indices, we instead promote alignment of features that are neighbors in a graph. Specifically, the *generalized LASSO* of Tibshirani and Taylor (2011); Liu et al. (2013) consider the estimators of the form

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\Gamma\beta\|_1 \tag{1.6}$$

for general X and Γ; note that both the fused LASSO and the estimator in (1.4) can be written in this form.

The works Caoa et al. (2018) and Viallon et al. (2016) use the generalized LASSO to mitigate correlation effects similar to the approach described in this work, but *without theoretical support*. Caoa et al. (2018) aims to predict Alzheimers disease outcomes using MRI measures as features. The authors use prior knowledge of correlations between MRI features to construct a regularizer which promotes alignment between correlated features. Viallon et al. (2016) seeks to predict outcomes in cancer patients based on gene expression data. The authors leverage side information of gene regulatory networks and promote alignment between adjacent vertices in the network. This work provides theoretical justification for the approaches described in those papers.

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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda_{\mathrm{TV}} \sum_{1 \le j < k \le p} \left|\beta_j - \beta_k\right| + \lambda_1\|\beta\|_1 .$$

In contrast to the fused LASSO, the clustered LASSO considers *all* pairwise differences of elements of β. She (2010) conducts a classical asymptotic analysis (fixed p and $n \to \infty$) of the clustered LASSO and its generalization (1.6) and establishes consistency results that depend upon Σ^{-1}.

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.

A key focus of our work is the setting in which columns of X may be highly correlated. 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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_S\|\beta\|_2^2, \tag{1.7}$$

which encourages a grouping effect, in which strongly-correlated predictors tend to be in or out of the support of the estimate together. Witten et al. (2014) propose a *Cluster Elastic Net* estimator which 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. 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). We examine the performance of CRL in this paper. *Grouping pursuit* (Shen and Huang, 2010) explores clustering columns of X while leveraging y by using a non-convex variant of the fused LASSO.

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 β; specifically,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} w_j|\beta|_{[j]},$$

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). Figueiredo and Nowak (2016) demonstrated that when two columns of X were *identical*, then OWL would assign the corresponding elements of β 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.

An estimator called *Pairwise Absolute Clustering and Sparsity (PACS)* estimator is proposed by Sharma et al. (2013). Hebiri and van de Geer (2011) consider smooth *S-LASSO* estimators of the form

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

The first regularization term, unlike the total variation term in (1.4), is a quadratic penalty similar to what appears in the elastic net (1.7) (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 Γ in (1.6) are related. A similar approach to Hebiri and van de Geer (2011) is the *weighted fusion estimator* proposed by Daye and Jeng (2009). 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 et al. (2011) uses a *minimum concave penalty (MCP)* to replace the LASSO penalty in a weighted fusion estimator to reduce bias.

2. 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

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

Assumption 2.1

We assume that there exists an absolute constant $c_u > 0$ such that

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

Remark 2.1

This statement assumes that Σ is normalized such that the largest eigenvalue of Σ can be upper bounded by a positive constant.

Assumption 2.2

There exists an absolute constant $c_\ell > 0$ such that:

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

Remark 2.2

Assumption 2.2 ensures the ℓ_1 norm for each row/column is lower bounded by a constant. This assumption is much milder than assuming the minimum eigenvalue of Σ is bounded away from 0. As an example, Assumption 2.2 is satisfied when every diagonal entry of Σ is bounded below by c_ℓ Note that Assumption 2.1 automatically holds for appropriately normalized features. However the assumption is nontrivial when considered jointly with Assumption 2.2, because normalization can potentially cause a violation of Assumption 2.2. We show that both Assumptions 2.1 and 2.2 hold in the examples considered in Section 2.2.

Assumption 2.3

The estimated covariance matrix $\hat{\Sigma}$ that is used to construct the matrix Γ satisfies

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

where c_ℓ is defined in Assumption 2.2.

Remark 2.3

Assumption 2.3 states that we need a sufficiently accurate estimator $\hat{\Sigma}$ for Σ. If Assumption 2.3 is satisfied then we can use $\hat{\Sigma}$ to construct Γ for our optimization problem stated in (1.5). We estimate Σ using side information that is not necessarily based on $\left(X^{(i)}\right)_{i=1}^{n}$. For instance, in the cytochrome P450 enzyme setting described in Section 1.1, we can leverage the recombination information of each chimeric protein to help estimate Σ. We elaborate on this in Section 4.1. In an MRI context, one can leverage prior knowledge of correlations between MRI features Caoa et al. (2018). In climate forecasting settings, physics-based simulations can be used to generate accurate covariance estimates.

In some settings, our source of side information may not directly yield an estimate of Σ, but rather a collection of m i.i.d. unlabeled feature vectors $\left(\breve{X}^{(i)}\right)_{i=1}^{m}$ that are potentially independent of the design features $\left(X^{(i)}\right)_{i=1}^{n}$ with $\breve{X}^{(i)} \sim \mathcal{N}\left(\boldsymbol{0}, \Sigma_{p \times p}\right)$. In this case, we need to estimate Σ based on $\left(\breve{X}^{(i)}\right)_{i=1}^{m}$, and there is a large literature on high-dimensional covariance estimation in high dimensions under different structural assumptions (see Bickel and Levina (2008b,a); Cai and Liu (2011); Cai et al. (2016); Donoho et al. (2013); Baik and Silverstein (2006)). As an example, we consider estimators based on thresholding the sample *covariance matrix under block structural assumptions developed by Bickel and Levina (2008a). We show that when the covariance matrix is block structured with K blocks, and m* = $\Omega(K^2 \log p)$, *Assumption 2.3 is satisfied. See* Appendix A *for more details.*

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

Definition 2.1

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

$$k_{\varnothing} := 1,\; k_T := \inf_{\beta \in \mathbb{R}^p} \frac{\sqrt{|T|}\|\beta\|_2}{\|(\widetilde{\Gamma}\beta)_T\|_1} \; for\; T \neq \varnothing\,.$$

This compatibility factor k_T reflects the degree of compatibility of the ℓ_1-regularizer $\|(\widetilde{\Gamma}\beta)_T\|_1$ and the ℓ_2-error norm $\|\beta\|_2$ for a set T. This compatibility factor appears explicitly in the bounds of our main theorem.

Definition 2.2

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

$$\rho := \|\widetilde{\Gamma}^{\dagger}\|_{1,2} = \max_{j=1,2,\ldots,m+p} \|s_j\|_2\,.$$

Remark 2.4

Definitions 2.1 and 2.2 are first proposed in Hütter and Rigollet (2016), though the definition of ρ *is based on* $\widetilde{\Gamma}$ *rather than* Γ. *Later we will see that* ρ *and* k_T *are crucial for our main results. The quantity* $\frac{\rho}{k_T}$ *is similar in flavour to the condition number of the matrix* $\widetilde{\Gamma}$.

Finally, we define the *estimated graph Laplacian* $L := \Gamma^{\top}\Gamma$. Recall that Γ, and therefore L, are constructed using the estimated covariance matrix $\widehat{\Sigma}$ rather than Σ. Spectral properties of L will play a crucial role in the mean-squared error bounds we derive.

Theorem 1

Suppose $\lambda_1 > 0$ *and Assumptions 2.1 to 2.3 are satisfied and suppose the estimated covariance matrix* $\widehat{\Sigma}$ *is constructed independently from the sample* $\left(X^{(i)}\right)_{i=1}^{n}$. *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*

$$\|\widehat{\beta} - \beta^*\|_2^2 \le C_u \min_T \max\left\{\frac{\lambda_1^2\left|T\right|}{k_T^2 \lambda_{\min}^2(\Sigma + \lambda_S L)}, \frac{\lambda_1\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1 + \lambda_1^2\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1^2}{\lambda_{\min}(\Sigma + \lambda_S L)}\right\},$$

provided $\frac{\lambda_1^2|T|}{k_T^2 \lambda_{\min}^2(\Sigma + \lambda_S L)} \to 0$ *(*i.e., *that the estimator is consistent).*

Remark 2.5

The assumption that the estimated covariance matrix $\widehat{\Sigma}$ *is constructed independently from the sample* $\left(X^{(i)}\right)_{i=1}^n$ *fits the settings we are motivated by where it is possible to exploit side information to estimate* $\widehat{\Sigma}$*. If instead the covariance matrix is estimated using* $\left(X^{(i)}\right)_{i=1}^n$ *one achieves a similar bound to Theorem 1 but without a factor of* ρ*. In some cases, such as the examples considered in Section 2.2* $\rho \ll 1$ *so assuming* $\widehat{\Sigma}$ *and* $\left(X^{(i)}\right)_{i=1}^n$ *are independent leads to sharper bounds.*

Remark 2.6

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* $D_{jj} = \sum_{k=1}^p \left|\widehat{\Sigma}_{j,k}\right|, 1 \le j \le p$*, then*

$$\Sigma + L := \Sigma - \widehat{\Sigma} + D .$$

Hence if Σ *and* $\widehat{\Sigma}$ *are "close" as is specified by Assumption 2.3, 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 2.2 and 2.3 are satisfied and $0 \le \lambda_S \le 1$*. Then*

$$\lambda_{\min}(\Sigma + \lambda_S L) \ge (1 - \lambda_S)\lambda_{\min}(\Sigma) + \lambda_S \frac{c_\ell}{2} .$$

Thus even if $\lambda_{\min}(\Sigma) = 0$*, choosing* λ_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 2.7

The consistency statement above is needed due to a condition in the statement of Lemma 8, and it must hold for any subset T for which we want to apply the theorem, but it need not hold for all possible subsets. In our examples, frequently choose $T = Supp\left(\widetilde{\Gamma}\beta^*\right)$.

Remark 2.8

$\|L\beta^*\|_\infty$ *can be seen as a measure of the* misalignment *of the signal* β^* *and the graph represented by the matrix* Γ. *Note that we require* $\lambda_1 \geq 8\lambda_S\|L\beta^*\|_\infty$. *Hence there is a clear trade-off in the choice of* λ_S. *Choosing* λ_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., β^* *is well-aligned with L) in order to obtain consistent mean-squared error bounds. Note that analysis of OWL (Figueiredo and Nowak, 2016) assumes* $L\beta^* = \mathbf{0}$ *(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* β^*, *just like the classical LASSO.*

Remark 2.9

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:*

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{\left(\frac{\log p}{n} + \lambda_S^2\|\beta^*\|_\infty^2\right)\|\beta^*\|_0}{\lambda_{\min}^2\left(\Sigma + \lambda_S I_{p \times p}\right)}.$$

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* β^* *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 ρ 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*

$$k_T^{-1} \le \frac{\lambda_{\mathrm{TV}}\sqrt{2\|\widehat{\Sigma}\|_{1,1}|T_1|} + \sqrt{|T_2|}}{\sqrt{|T_1| + |T_2|}}.$$

The proof for this lemma can be found in Appendix F.

Remark 2.10

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

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

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

Lemma 3

Let G denote the graph associated with $\widehat{\Sigma}$. Then

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

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

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

Theorem 2

Suppose $\lambda_1 > 0$ and Assumptions 2.1 to 2.3 are satisfied and suppose the estimated covariance matrix $\widehat{\Sigma}$ is constructed independently from the sample $\left(X^{(i)}\right)_{i=1}^n$. If we choose

$$\lambda_1 \ge 48\sqrt{\frac{\sigma^2 c_u \log p}{n}\max_{1 \le k \le K}(\frac{1}{|B_k|} + \frac{2}{1 + \mu_k\lambda_{\mathrm{TV}}^2})} + 8\lambda_S\|L\beta*\|_\infty$$

Then there exist absolute positive constants C_1 and C such that

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

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

The proof of Theorem 2 is provided in Section B. 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 λ_{TV} and its impact on the mean-squared error (MSE) bounds. *As* λ_{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 λ_1 decreases with λ_{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 $\lambda_1^2 \lambda_{TV}^2 \| \widehat{\Sigma} \|_{1,1} \| \Gamma \beta^* \|_0$ or $\lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_0$ and the lower bound on $\lambda_1 \lambda_{TV}$ increases as λ_{TV} increases. Determining optimal error rates is in general a challenging problem. However, in the special cases of the block and lattice graphs considered in Section 2.2 our bounds are consistent with known optimal rates. It is straightforward to extend the proofs of Theorems 1 and 2 in order to derive prediction error bounds on $\| X \widehat{\beta} - X \beta^* \|_2^2$. This is discussed in more detail in Appendix D.

2.1. 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 $\lambda_S = \lambda_{\mathrm{TV}} = 0$. From Theorem 2 we can see that

$$\| \widehat{\beta} - \beta^* \|_2^2 \leq \frac{\sigma^2 c_u \log p}{n} \| \beta^* \|_0, \tag{2.1}$$

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

$$\| \widehat{\beta} - \beta^* \|_2^2 \leq \lambda_1^2 \| \beta^* \|_0 + \min\left(\lambda_1^2 \lambda_{TV}^2 \| \widehat{\Sigma} \|_{1,1} \| \Gamma \beta^* \|_0, \lambda_1 \lambda_{TV} \| \Gamma \beta^* \|_1 \right)$$

where $\lambda_1^2 = O(\max_{1 \leq k \leq K} \frac{\sigma^2 c_u \log p}{n |B_k|})$ and $\lambda_1^2 \lambda_{\mathrm{TV}}^2 = O(\max_{1 \leq k \leq K} \frac{|B_k|}{\mu_k} \max_{1 \leq k \leq K} \frac{\sigma^2 c_u \log p}{n |B_k|})$. The upper bound may be well below the classical LASSO bound in (2.1) when $\min_k |B_k| \gg 1$ and $\Gamma \beta^* \approx \mathbf{0}$.

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 β^*; 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 β^* 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 λ_S—The smoothing penalty plays the role of a pre-conditioner where the trade-off is the addition of another term $\lambda_S\|L\beta^*\|_\infty$. This can also be seen in the optimization problem (1.5) where X is transformed to $\widetilde{X}$, so even if the restricted eigenvalue condition is not satisfied for X, it is satisfied for $\widetilde{X}$. What distinguishes our results from previous work using pre-conditioners for the LASSO (Jia et al., 2015; Wauthier et al., 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 2.9.

Role of λ_{TV}—As mentioned earlier, the total variation penalty promotes estimates well-aligned with the graph. As λ_{TV} increases, the sparsity parameter λ_1 decreases while $\lambda_1\lambda_{TV}$ increases. By increasing λ_{TV} we can also adapt to settings where β^* 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 μ_k (the smallest non-zero eigenvalue of a block) and $|B_k|$ (the block size). Clearly the larger μ_k and $|B_k|$, the lower the bound on λ_1 which potentially suggests lower mean-squared error. On the other hand, as we illustrate with specific examples later, larger μ_k typically indicates higher correlation between more covariates and larger $|B_k|$ corresponds to nodes being correlated, which means $\lambda_{\min}(\Sigma)$ is smaller.

2.2. Specific covariance graph structures

In this section we explore three specific graph structures and discuss suitable choices of λ_S,λ_1 and λ_{TV}. For each graph structure we assume

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

we refer to r as the correlation coefficient. Note that here a is a normalization parameter that we set to ensure such that Assumptions 2.1 and 2.2 are satisfied. We will talk about the specific choices of a for each graph structure below. Our general results allow us to quantify the impact of misspecification of Σ, but for interpretability and simplicity of exposition, we will assume in this section that $\widehat{\Sigma} = \Sigma$ – that is, that we have perfect side information about the correlation graph.

2.2.1. 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 Σ (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

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

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

We set $a = \frac{K}{p}$ to ensure that Assumptions 2.1 and 2.2 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 $\max_{1 \le k \le K} \frac{1}{|B_k|}, \mu_k, \rho, \lambda_{\min}(\Sigma + \lambda_S L)$:

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

$$\begin{gathered}\max_{1 \le k \le K} \frac{1}{|B_k|} = \frac{K}{p}, \\ \mu_k = r, \quad for\ all\ \ k \\ \rho \le \sqrt{\frac{K}{p} + \frac{2}{1 + r\lambda_{\mathrm{TV}}^2}}, \\ \lambda_{\min}(\Sigma + \lambda_S L) \ge (1 - \lambda_S)(1 - r)\frac{K}{p} + \lambda_S r\,.\end{gathered}$$

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

Corollary 1: *For a block complete graph with details described above, suppose that* $\hat{\Sigma} = \Sigma$. *If*

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

and $\lambda_1 \lambda_{TV} \|\Gamma\beta^*\|_1 \le 1$. *Then with probability at least* $1 - \frac{C_1}{p}$

$$\|\hat{\beta} - \beta^*\|_2^2 \le \frac{C\left(\lambda_1^2 \|\beta^*\|_0 + \min\left\{\lambda_1^2 \lambda_{TV}^2 \|\Gamma\beta^*\|_0, \lambda_1 \lambda_{TV} \|\Gamma\beta^*\|_1\right\}\right)}{\min\left\{\left[(1 - \lambda_S)(1 - r)\frac{K}{p} + \lambda_S r\right], \left[(1 - \lambda_S)(1 - r)\frac{K}{p} + \lambda_S r\right]^2\right\}}$$

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

Consider a setting where $r \approx 1$ and $\Gamma\beta^* \approx \mathbf{0}$ (near-perfect alignment which corresponds to the parameters in each block having the same values). Let $K_1 \le K$ denote the number of blocks which have features that are active in β^*. If we choose $\lambda_S \asymp 1, \lambda_{TV}^2 \asymp \frac{p}{K}$, and $\lambda_1^2 \asymp \frac{K \log p}{pn}$, then

$$\|\hat{\beta} - \beta^*\|_2^2 \precsim \frac{K_1 \log p}{n};$$

that is, the MSE is not determined by the number of nonzeros in β^*, but rather by K_1, the number of clusters of active nodes. In the case of perfect correlation between the blocks this matches the minimax optimal rate up to log factors (Raskutti et al. (2011)). A similar scaling was derived in Figueiredo and Nowak (2016) also under the assumption that $\Gamma\beta^* \approx \mathbf{0}$.

2.2.2. Chain covariance graph—

The covariance matrix correspnding 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 2.1 and 2.2 are clearly satisfied and requiring $r \in \left(0, \frac{1}{2}\right)$ ensures Σ 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 ρ and $\lambda_{\min}(\Sigma + \lambda_S L)$ for the chain covariance graph:

Lemma 5: *For a chain graph with details described above, suppose that* $\widehat{\Sigma} = \Sigma$. *Then*

$$\rho \le \sqrt{\frac{1}{p} + \frac{2\pi}{r\lambda_{TV} + 1}},$$
$$\lambda_{\min}(\Sigma + \lambda_S L) \ge (1 - \lambda_S)(1 - 2r) + \lambda_S .$$

Using Lemma 5 we have the following corollary for the chain graph:

Corollary 2: *For a chain graph with details described above, suppose that* $\widehat{\Sigma} = \Sigma$. *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 $\lambda_1\lambda_{TV}\|\Gamma\beta^*\|_1 \le 1$, *then with probability at least* $1 - \frac{C_1}{p}$ *we have*

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

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 β^* is strong but imperfect. Suppose that within β^* there are $O(1)$ blocks which are active, and within each active block all the coefficients 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 Corollary 2 says

$$\mathrm{MSE}_{\mathrm{GTV}} \preceq \frac{\sqrt{\|\beta^*\|_0}\log p}{n}$$

which is stronger than the LASSO guarantee of

$$\mathrm{MSE}_{\mathrm{LASSO}} \preceq \frac{\sqrt{\|\beta^*\|_0}\log p}{n}.$$

2.2.3. 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

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

We require $r \in \left(0, \frac{1}{4}\right)$ so that Σ is positive semi-definite. Clearly Assumptions 2.1 and 2.2 are satisfied for any $r \in \left(0, \frac{1}{4}\right)$, 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 ρ and $\lambda_{\min}(\Sigma + \lambda_S L)$:

Lemma 6: *For a lattice graph with details described above, suppose that* $\widehat{\Sigma} = \Sigma$. *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: *For a lattice graph with details described above, suppose that* $\widehat{\Sigma} = \Sigma$. *If we choose*

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

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

$$\|\widehat{\beta} - \beta^*\|_2^2 \leq \frac{C\left(\lambda_1^2\|\beta^*\|_0 + \min\left\{\lambda_1^2\lambda_{TV}^2\|\Gamma\beta^*\|_0, \lambda_1\lambda_{TV}\|\Gamma\beta^*\|_1\right\}\right)}{\min\left\{[(1 - \lambda_S)(1 - 4r) + \lambda_S], [(1 - \lambda_S)(1 - 4r) + \lambda_S]^2\right\}}$$

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 β^* 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 $n \preceq p$ and we set $\lambda_{TV} \approx \sqrt{n}$, $\lambda_S \approx 0$ and $\lambda_1 \approx \frac{\log p}{n}$. Corollary 3 says

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

which is stronger than the LASSO guarantee of

$$\mathrm{MSE}_{\mathrm{LASSO}} \preceq \frac{\|\beta^*\|_0 \log p}{n}.$$

Note that the $\mathrm{MSE}_{\mathrm{GTV}}$ bound from this example is identical to the $\mathrm{MSE}_{\mathrm{GTV}}$ bound from the example considered in the chain graph section. On one hand, our bound on ρ 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 β^* than between the lattice graph and β^*. 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 $\|\Gamma\beta^*\|_0 \approx \sqrt{\|\beta^*\|_0}$. Our overall rate matches the optimal rates derived in the lattice graph denoising setting considered in Hütter and Rigollet (2016).

3. 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. Specific details on how the covariance matrix Σ is constructed for each graph structure is discussed in Section O in the Appendix. The data is generated according to $y = X\beta^* + \epsilon$ with $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$. Each row of X is independently generated from $\mathcal{N}(\mathbf{0}, \Sigma_{p \times p})$ and ϵ is generated from $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n \times n})$ with $\sigma = 0.01$. Additionally, we generate $X_{\mathrm{ind}} \in \mathbb{R}^{1000 \times p}$ with each row of X_{ind} independently generated from $\mathcal{N}(\mathbf{0}, \Sigma_{p \times p})$. This X_{ind} provides side information that can be used to improve estimates of Σ. This X_{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 $(\mathrm{MSE} = \|\hat{\beta} - \beta^*\|_2^2)$. For all methods, tuning parameters are chosen based on five-fold cross-validation (in the case of GTV, we perform a three-dimensional search to find λ_1, λ_{TV} and λ_S). 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, $\widehat{\Sigma}$, 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 Γ based on $\widehat{\Sigma}$ and then estimate $\widehat{\beta}$ using (1.5).

GTV-Indep (Our method)

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

CRL-Esti

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 Bühlmann 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_{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_{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 and the number of covariates p. The results are summarized in Figure 1. 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. 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 the number of active nodes increases, which is consistent with our theoretical results.

We next test how the error scales with $\|\Gamma\beta^*\|_0$ and $\|\Gamma\beta^*\|_1$. In Figure 2a we take a chain graph with $p = 280$ nodes and let the first $s = 80$ nodes be active. For the active nodes we set $\beta_j^* \sim \mathcal{N}\left(1, \sigma^2\right)$ for varying values of σ. 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 2b 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 $\mathcal{N}\left(1, .01^2\right)$. 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 β^*.

4. Biochemistry application: Cytochrome P450 enzymes

In this section we describe an application of the proposed GTV methodology to protein thermostability data. As described in Section 1.1, the thermostability data we use was provided by the Romero Lab at UW-Madison. The data contains thermostability measurements for 242 proteins in the P450 protein family. For each protein, 50 structure features were simulated via RosettaCommons (Alford et al., 2017) and the goal is to understand the relationship between the 50 structural features and thermostability. Hence the design matrix $X \in \mathbb{R}^{242 \times 50}$ consists of the structural features. The response variable $y \in \mathbb{R}^{242}$ contains the thermostability measurements. Additionally, we have side information in the form of the amino acid sequences that make each of the 242 proteins; this is used to estimate the covariance matrix amongst the structural features.

4.1. Estimation of the covariance matrix with side information

One advantage of our GTV method is that side information can be incorporated to estimate the strength of correlations among features. It is a well known fact that the structure of the protein is a function of its amino acid sequence. We exploit this sequence and structure relationship and model the structural features as linear functions of sequence features. Then we use this model to obtain a better approximation of the covariance of structural features.

The proteins were created by the recombination of 3 other proteins. Each protein's amino acid sequence can be thought of as having 8 pieces/blocks where each piece came from one of 3 parent proteins (Figure 3). So the amino acid sequence can be represented as 8 categorical features, each with 3 categories. Each feature represents one piece of the sequence and indicates which parent that piece came from. We can use the one-hot encoding of these 8 categorical features to obtain 24 binary features that represent an amino acid sequence for a protein. Because each piece comes from one of three parents, the sum of the 3 binary features for each piece of the sequence must be 1. So only 2 parameters are needed for each piece of the sequence. Hence a model of the amino acid sequence has $K = 16$ parameters.

Hence we model $p = 50$ structural features as linear functions of $K = 16$ binary sequence features via a multivariate linear regression model. More concretely, we assume a linear model

$$X^{(i)} = A^T S^{(i)} + \delta^{(i)} \tag{4.1}$$

where $X^{(i)} \in \mathbb{R}^p$ is a vector of the ith structural feature and $S^{(i)} \in (0, 1)^K$ is the binary sequence features of the ith enzyme in the dataset. The matrix $A \in \mathbb{R}^{K \times p}$ is an unknown parameter matrix which determines the relationship between $X^{(i)}$ and $S^{(i)}$, and we assume Gaussian noise $\delta^{(i)} \sim \mathcal{N}\left(0, \sigma_\delta^2\right)$ independent from $S^{(i)}$ and $\epsilon^{(i)}$.

We note that the model assumption (4.1) amounts to assuming that the thermostability y can be modeled by the sequence matrix S which is of rank K. Although modeling y directly via S is possible, the results will not provide an understanding of how structural features contribute to the thermostability of a protein, which is the goal of our analysis.

Exploiting the structure of X in (4.1), we estimate the covariance matrix of X given sequence S as

$$\hat{\Sigma}_{\text{ind}} := \widehat{\text{Var}}\left(\mathbb{E}\left[X^{(i)} \middle| S^{(i)}\right]\right) = \hat{A}^T \widehat{\text{Var}}\left(S^{(i)}\right)\hat{A} = \hat{A}^T \hat{\Sigma}_s \hat{A},$$

where $\hat{\Sigma}_s$ is an empirical covariance matrix of $\left(S^{(i)}\right)_{i=1}^n$ and $\hat{A}$ is the LSE of A, i.e.

$$\hat{A} = \arg\min_{A \in \mathbb{R}^{K \times p}} \|\mathbf{X} - \mathbf{S}A\|_F^2 .$$

We note that the dimensions of A and Σ_s are K by p and K by K, respectively. Thus we reduce the estimation problem of a p by p matrix to a smaller problem, with $K = 16$ being much less than $p = 50$.

4.2. Results

We compare our GTV method (with and without side information) with Ordered Weighted LASSO (OWL), Cluster Representative LASSO (CRL), and the standard LASSO (LASSO), and the Elastic Net (EN) method. For all models, the tuning parameters were selected via five-fold cross validation on the training set. For OWL, the weights were set corresponding to the OSCAR regularizer.

To compare the performance of the five methods on the real P450 data, we considered two performance criteria: prediction accuracy and stability of estimated coefficients. To measure stability between estimated coefficients, we considered following two criteria:

1. $\mathrm{Cor}(\hat{\beta}_i, \hat{\beta}_j)$ where $\hat{\beta}_i$ and $\hat{\beta}_j$ are estimates from two different fittings for the same model.

2. Tanimoto Distance (Kalousis et al., 2007):

$$D(i, j) := 1 - \frac{\left|\mathrm{supp}\left(\hat{\beta}_i\right)\right| + \left|\mathrm{supp}\left(\hat{\beta}_j\right)\right| - 2\left|\mathrm{supp}\left(\hat{\beta}_i\right) \cap \mathrm{supp}\left(\hat{\beta}_j\right)\right|}{\left|\mathrm{supp}\left(\hat{\beta}_i\right)\right| + \left|\mathrm{supp}\left(\hat{\beta}_j\right)\right| - \left|\mathrm{supp}\left(\hat{\beta}_i\right) \cap \mathrm{supp}\left(\hat{\beta}_j\right)\right|}$$

   where supp refers to the support set.

For prediction accuracy, we use 10-fold cross validation. We trained the six models on each training set and evaluated the prediction performances on the test set. On the other hand, stability measures were calculated by spliting the entire P450 dataset into ten non-overlapping subsamples and fitting the six models using each of the subsamples.

Table 1 summarizes prediction accuracy. The result for EN is excluded since the tuning parameter for the l_2 penalty λ_S was chosen to be 0 in all cross-validation folds, and the result for EN is the same as LASSO. From the Table 1, we see that GTV Esti has the highest accuracy. GTV Ind (GTV with side information) is the next most accurate. CRL Ind and CRL Esti show very bad prediction performance. CRL is expected to perform badly in the case where variables are not grouped into tight clusters or coefficients within a group have opposite signs and their sum is close to zero Bühlmann et al. (2013). In our application, in most cross-validation folds Algorithm 1 in Bühlmann et al. (2013) resulted in one huge cluster in the case of CRL Esti, whose member features do not have similar effects on the response variable. We observed similar phenomenon in the case of CRL Ind, although to a lesser extent than the CRL Esti, where we observed one cluster with nine features with opposite effects and the remaining clusters are of size 1. As a result, both CRL methods demonstrated very poor prediction results.

Figure 4 demonstrates the correlation and variable selection stability. GTV Ind and GTV Esti show the most stable performances overall. In terms of correlations, all five methods generated highly correlated coefficients across different fits, except OWL which had a few outliers. For variable selection stability, both GTV methods and OWL produced the same support sets in all fits. On the contrary, the support sets from LASSO and both CRL methods greatly varied across fits. Only about 30% of the support sets overlap between any pair of fits. It appears that relatively strong correlation in the design but the lack of tightly grouped clusters contributed to the instability of clustering and support recovery in LASSO and CRL methods.

5. Conclusion

This paper describes a new graph-based regularization method for high-dimensional regression with highly-correlated designs and alignment between the covariance and regression coefficients. 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). 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 β^* and graph. We also demonstrate in both simulations and a biochemistry application superior performance of our method compared to 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 and more extensive evaluations on real-world data.

Acknowledgement

The authors would like to thank Ian Kinsella for helpful discussions, suggestions, and preliminary experiments. The authors would also like to thank Philip Romero and Jerry Duan for creating and providing access to the biochemistry example.

Funding: Raskutti and Li were supported by NSF DMS 1407028. Song was supported by NIH R01 GM131381-01. Mark was supported by NSF Awards 0353079 and 1447449. Willett was supported by NIH Award 1 U54 AI117924-01 and NSF Awards 0353079, 1447449, 1740707, and 1839338.

Appendix A. Covariance estimation

Assume we observe a collection of m i.i.d. unlabeled feature vectors $\left(\breve{X}^{(i)}\right)_{i=1}^{m}$ that may be independent of the design features $\left(X^{(i)}\right)_{i=1}^{n}$ with $\breve{X}^{(i)} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_{p \times p}\right)$. In this case, we need to estimate Σ based on $\left(\breve{X}^{(i)}\right)_{i=1}^{m}$, and there is a large literature on high-dimensional covariance estimation in high dimensions under different structural assumptions (see Bickel and Levina (2008b,a); Cai and Liu (2011); Cai et al. (2016); Donoho et al. (2013); Baik and Silverstein (2006)). As an example, we consider estimators based on thresholding the sample covariance matrix under block and sparsity assumptions developed by Bickel and Levina (2008a).

A.1. Sparse covariance matrix

To be specific, suppose the true covariance matrix Σ belongs to the following class:

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

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 $t = \Omega\left(\sqrt{\frac{\log p}{m}}\right)$, then

$$\|\hat{\Sigma} - \Sigma\|_{1,1} = \Omega_P\left(c_0(p)M\left(\frac{\log p}{m}\right)^{\frac{1-q}{2}}\right).$$

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,

$$\|\hat{\Sigma} - \Sigma\|_{1,1} = \Omega_P\left(s\sqrt{\frac{\log p}{m}}\right),$$

meaning if $m = \Omega(s^2 \log p)$, Assumption 2.3 is satisfied.

A.2. Block covariance matrix

On the other hand, if the covariance matrix Σ is not sparse but rather block-structured, we can use an alternative bound developed in Bickel and Levina (2008a). If Σ has K identical blocks where each block has p/K elements, we can ensure Assumptions 2.1 and 2.2 are satisfied if $\Sigma_{j,k} = \Omega\left(\frac{K}{p}\right)$ 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

$$\max_{j,k}\left|\frac{\hat{\Sigma}_{j,k} - \Sigma_{j,k}}{K/p}\right| = \Omega_P\left(\sqrt{\frac{\log p}{m}}\right) \tag{A.1}$$

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

$$\max_{j,k}\left|\hat{\Sigma}_{j,k} - \Sigma_{j,k}\right| = \Omega_P\left(\frac{K}{p}\sqrt{\frac{\log p}{m}}\right) \tag{A.2}$$

and

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

by (A.2), so that when $m = \Omega(K^2 \log p)$ Assumption 2.3 is satisfied.

Appendix B. Proof of Theorem 1

Much of our analysis follows standard steps for analysis of regularized M-estimators (see Bickel et al. (2009); Negahban et al. (2012); van de Geer (2000)), but we face two additional challenges not present in these works. First, since the regularization penalty in Equation (1.5) is $\|\widetilde{\Gamma}\beta\|_1$ rather than $\|\beta\|_1$ we need to deal with error terms involving $\widetilde{X}\widetilde{\Gamma}^{\dagger}$ instead of $\widetilde{X}$. 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 $\widetilde{X}$ rather than X. We incorporate techniques from Raskutti et al. (2010) in order to accomplish this.

Based on the optimization problem (1.5), by the definition of $\hat{\beta}$ and the basic inequality,

$$\frac{1}{n}\|\tilde{y} - \widetilde{X}\hat{\beta}\|_2^2 + \lambda_1\|\widetilde{\Gamma}\hat{\beta}\|_1 \le \frac{1}{n}\|\tilde{y} - \widetilde{X}\beta^*\|_2^2 + \lambda_1\|\widetilde{\Gamma}\beta^*\|_1 .$$

By simple re-arrangement,

$$\frac{1}{n}\|\widetilde{X}(\hat{\beta} - \beta^*)\|_2^2 \le \frac{2}{n}(\tilde{y} - \widetilde{X}\beta^*)^\top \widetilde{X}(\hat{\beta} - \beta^*) + \lambda_1(\|\widetilde{\Gamma}\beta^*\|_1 - \|\widetilde{\Gamma}\hat{\beta}\|_1).$$

For the remainder of the proof let $\Delta: = \hat{\beta} - \beta^*$. Then

$$\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 \le \frac{2}{n}(\tilde{y} - \widetilde{X}\beta^*)^\top \widetilde{X}\Delta + \lambda_1(\|\widetilde{\Gamma}\beta^*\|_1 - \|\widetilde{\Gamma}\hat{\beta}\|_1).$$

First we control the term $(\tilde{y} - \widetilde{X}\beta^*)^\top \widetilde{X}\Delta$. Using basic algebra,

$$(\tilde{y} - \widetilde{X}\beta^*)^\top \widetilde{X}\Delta = \epsilon^\top X\Delta - n\lambda_S\beta^{*\top}\Gamma^\top\Gamma\Delta .$$

First, for the second term, we bound $n\lambda_S\beta^{*\top}\Gamma^\top\Gamma\Delta$ by

$$\begin{aligned} n\lambda_S\beta^{*\ \top}\Gamma^\top\Gamma\Delta &\le n\lambda_S\|\Gamma^\top\Gamma\beta^*\|_\infty\|\Delta\|_1 \\ &\le n\lambda_S\|L\beta^*\|_\infty\|\widetilde{\Gamma}\Delta\|_1 \\ &\le n\frac{\lambda_1}{8}\|\widetilde{\Gamma}\Delta\|_1, \end{aligned}$$

where the last inequality follows from the constraint that $\lambda_1 \ge 8\lambda_S\|L\beta^*\|_\infty$.

For the first term, we use the following bound which we provide the proof in Appendix L. This Lemma is presented in more generality than our statements of Theorems 1 and 2. In those Theorems we focus specifically on the case where Γ is constructed independently of X which covers our main setting of interest where side information is used to construct the estimated covariance matrix $\widehat{\Sigma}$. However, Lemma 7 also allows for possible dependence with a modification to ρ.

Lemma 7

Suppose we have $\lambda_1 \geq 48\tilde{\rho}\sigma\sqrt{\frac{c_u \log p}{n}}$ *and* $n \geq C_1 \log p$, *where* $\tilde{\rho} = \min\{\rho, 1\}$ *when* Γ *is constructed independently of* X *or* $\tilde{\rho} = 1$ *otherwise, and* $n \geq C_1 \log p$. *Then with probability at least* $1 - \frac{C_2}{p}$,

$$\epsilon^\top X\Delta \leq \frac{n\lambda_1}{8}\|\widetilde{\Gamma}\Delta\|_1$$

for absolute constants $C_1, C_2 > 0$.

Combining the constraints for λ_1 with the inequalities above,

$$\frac{2}{n}\left(\tilde{y} - \widetilde{X}\beta*\right)^\top \widetilde{X}\Delta \leq \frac{\lambda_1}{4}\|\widetilde{\Gamma}\Delta\|_1 + \frac{\lambda_1}{4}\|\widetilde{\Gamma}\Delta\|_1 = \frac{\lambda_1}{2}\|\widetilde{\Gamma}\Delta\|_1 .$$

Putting these pieces together we have

$$\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 \leq \frac{\lambda_1}{2}\left(\|\widetilde{\Gamma}\Delta\|_1 + 2\|\widetilde{\Gamma}\beta * \|_1 - 2\|\widetilde{\Gamma}\hat{\beta}\|_1\right). \tag{B.1}$$

Furthermore by the triangle inequality and the fact that $\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 \geq 0$ we have

$$0 \leq \|\widetilde{\Gamma}\left(\hat{\beta} - \beta*\right)\|_1 + 2\|\widetilde{\Gamma}\beta*\|_1 - 2\|\widetilde{\Gamma}\hat{\beta}\|_1 \leq 3\|(\widetilde{\Gamma}\Delta)_T\|_1 - \|(\widetilde{\Gamma}\Delta)_{T^c}\|_1 + 4\|\left(\widetilde{\Gamma}\beta*\right)_{T^c}\|_1 .$$

Therefore Δ lies in the translated cone

$$\mathscr{C}: = \left\{v: \|(\widetilde{\Gamma}v)_{T^c}\|_1 \leq 3\|(\widetilde{\Gamma}v)_T\|_1 + 4\|\left(\widetilde{\Gamma}\beta*\right)_{T^c}\|_1\right\}. \tag{B.2}$$

Moreover by the definition of k_T we have

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

from (B.1) we have

$$\frac{1}{2n}\|\widetilde{X}\Delta\|_2^2 \leq \lambda_1\|\left(\widetilde{\Gamma}\beta*\right)_{T^c}\|_1 + \frac{3\lambda_1}{4k_T}\sqrt{|T|}\|\Delta\|_2 . \tag{B.3}$$

B.1. Restricted Eigenvalue Condition

From (B.1) and (B.2) we need to lower bound

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

for all Δ belonging to the cone $\mathscr{C}$ defined in (B.2). The result is stated in the following lemma:

Lemma 8

For all Δ *belonging to the cone defined in (B.2) if we have*

$$\lambda_1 \le c_2\sqrt{\frac{\lambda_{\min}(\Sigma + \lambda_S L)}{|T|}}k_T, \tag{B.4}$$

then

$$\Delta^\top\left(\frac{X^\top X}{n} + \lambda_S L\right)\Delta \ge c_1\lambda_{\min}(\Sigma + \lambda_S L)\|\Delta\|_2^2 - c_3\lambda_1^2\|(\widetilde{\Gamma}\beta^*)_{T^c}\|_1^2 \tag{B.5}$$

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

The proof for this lemma is largely based on a technique used in Raskutti et al. (2010), and is provided in Appendix M. The condition in Equation (B.5) is the source of the consistency condition in Theorem 1.

B.2. Final Part for Proof

From (B.3) and (B.5),

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

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

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

with

$$\begin{aligned}
a &= 1,\\
b &= \frac{3\lambda_1\sqrt{|T|}}{2c_1k_T\lambda_{\min}(\Sigma + \lambda_S L)},\\
c &= \frac{1}{c_1\lambda_{\min}(\Sigma + \lambda_S L)}\left(2\lambda_1\|(\widetilde{\Gamma}\beta^*)_{T^c}\|_1 + c_3\lambda_1^2\|(\widetilde{\Gamma}\beta^*)_{T^c}\|_1^2\right).
\end{aligned}$$

By solving this quadratic inequality,

$$\|\Delta\|_2^2 \le 4\max\left\{b^2, \left|c\right|\right\}.$$

Therefore these exists a positive constant C_u such that

$$\|\hat{\beta} - \beta^*\|_2^2 \le C_u \max\left\{\frac{\lambda_1^2\left|T\right|}{k_T^2\lambda_{\min}^2(\Sigma + \lambda_S L)}, \frac{\lambda_1\|\left(\widetilde{\Gamma}\beta^*\right)T^c\|_1 + \lambda_1^2\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1^2}{\lambda_{\min}(\Sigma + \lambda_S L)}\right\}.$$

Note that the above inequality is true for all T, thus

$$\|\hat{\beta} - \beta^*\|_2^2 \le C_u \min_T \max\left\{\frac{\lambda_1^2\left|T\right|}{k_T^2\lambda_{\min}^2(\Sigma + \lambda_S L)}, \frac{\lambda_1\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1 + \lambda_1^2\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1^2}{\lambda_{\min}(\Sigma + \lambda_S L)}\right\}.$$

This completes the proof.

Appendix C. 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 $T = \mathrm{supp}\left(\widetilde{\Gamma}\beta^*\right)$ then $\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1 = 0$ and by Lemma 2,

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

Then by Theorem 1 we have

$$\|\hat{\beta} - \beta^*\|_2^2 \le \frac{2C_u}{\lambda_{\min}^2(\Sigma + \lambda_S L)}\left(\lambda_1^2\|\beta^*\|_0 + 2\lambda_1^2\lambda_{\mathrm{TV}}^2\|\hat{\Sigma}\|_{1,1}\|\Gamma\beta^*\|_0\right). \qquad \text{(C.1)}$$

On the other hand if we choose $T = \mathrm{supp}(\beta^*)$, $\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\|_1 = \lambda_{\mathrm{TV}}\|\Gamma\beta^*\|_1$ and by Lemma 2, $k_T^{-1} \le 1$. Thus if $\lambda_1\lambda_{\mathrm{TV}}\|\Gamma\beta^*\|_1 \le 1$ by Theorem 1

$$\|\hat{\beta} - \beta^*\|_2^2 \le C_u\left(\frac{\lambda_1^2\|\beta^*\|_0}{\lambda_{\min}^2(\Sigma + \lambda_S L)} + \frac{2\lambda_1\lambda_{\mathrm{TV}}\|\Gamma\beta^*\|_1}{\lambda_{\min}(\Sigma + \lambda_S L)}\right). \qquad \text{(C.2)}$$

Theorem 2 follows by combining (C.1) and (C.2) and taking the minimum over these two choices of T.

Appendix D. Prediction Error Bounds

In this section we observe that the proofs of Theorems 1 and 2 also give rise to bounds on the prediction error of our estimator. Starting from Equation (B.3) in the proof of Theorem 1,

$$\frac{1}{2n}\|\widetilde{X}\Delta\|_2^2 \leq \lambda_1\left\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\right\|_1 + \frac{3\lambda_1}{4k_T}\sqrt{|T|}\|\Delta\|_2 . \tag{D.1}$$

The Restricted Eigenvalue condition in Lemma 8 gives that

$$\|\Delta\|_2 \leq c\sqrt{\lambda_{\min}(\Sigma + \lambda_S L)}\frac{1}{\sqrt{n}}\|X\Delta\|_2 .$$

Thus

$$\frac{1}{2n}\|\widetilde{X}\Delta\|_2^2 - \frac{3c\lambda_1}{4k_T\sqrt{n}}\sqrt{|T\,|\,\lambda_{\min}(\Sigma + \lambda_S L)}\|X\Delta\|_2 - \lambda_1\left\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\right\|_1 \leq 0. \tag{D.2}$$

As in the proof of Theorem 1, we can solve this quadratic inequality to conclude

Theorem 3

(Theorem 1 for Prediction Error). *Suppose the conditions of Theorem 1 hold. Then with probability at least* $1 - \frac{C_1}{p}$ *we have*

$$\frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2 \leq \min_T \max\left(\frac{\lambda_1^2\left|T\right|}{k_T^2\lambda_{\min}(\Sigma + \lambda_S L)}, \lambda_1\left\|\left(\widetilde{\Gamma}\beta^*\right)_{T^c}\right\|_1\right)$$

This result holds for all choices of T. Choosing T to be $\operatorname{supp}\left(\widetilde{\Gamma}\beta^*\right)$ and supp(β^*) as in the proof of Theorem 2 gives an analogous result for prediction error.

Theorem 4

(Theorem 2 for Prediction Error). *Suppose the conditions of Theorem 2 hold. Then with probability at least* $1 - \frac{C_1}{p}$ *we have*

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

Appendix E. Proof of Lemma 1

First note that

$$\begin{aligned}\lambda_{\min}(\Sigma + \lambda_S L) &= \lambda_{\min}((1 - \lambda_S)\Sigma + \lambda_S(\Sigma + L)) \\ &\geq (1 - \lambda_S)\lambda_{\min}(\Sigma) + \lambda_S\lambda_{\min}(\Sigma + L)\end{aligned}$$

where the second inequality follows from Weyl's inequality. For the remainder of the proof, we bound $\lambda_{\min}(\Sigma + L)$. Recall that

$$\Sigma + L = \Sigma - \widehat{\Sigma} + D \tag{E.1}$$

where $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix with

$$D_{jj} = \sum_{k=1}^{p} \left|\widehat{\Sigma}_{j,k}\right|, \ 1 \le j \le p .$$

Then

$$\lambda_{\min}(\Sigma + L) = \lambda_{\min}(\Sigma - \widehat{\Sigma} + D) \ge \lambda_{\min}(\Sigma - \widehat{\Sigma}) + \lambda_{\min}(D)$$

by Weyl's inequality. Since

$$\lambda_{\min}(\Sigma - \widehat{\Sigma}) = -\lambda_{\max}(\widehat{\Sigma} - \Sigma) \ge -\|\Sigma - \widehat{\Sigma}\|_{op} \ge -\|\Sigma - \widehat{\Sigma}\|_{1,1} .$$

Hence

$$\begin{aligned}
\lambda_{\min}(\Sigma + L) &\ge \lambda_{\min}(D) - \|\Sigma - \widehat{\Sigma}\|_{1,1} \\
&\ge \min_j \sum_{k=1}^{p} \left|\widehat{\Sigma}_{j,k}\right| - \frac{c_\ell}{4} (\text{ by Assumption 2.3}) \\
&\ge \min_j \left[\sum_{k=1}^{p} \left|\Sigma_{j,k}\right| - \sum_{k=1}^{p} \left|\Sigma_{j,k} - \widehat{\Sigma}_{j,k}\right| \right] - \frac{c_\ell}{4} \\
&\ge \min_j \sum_{k=1}^{p} \left|\Sigma_{j,k}\right| - \max_j \sum_{k=1}^{p} \left|\Sigma_{j,k} - \widehat{\Sigma}_{j,k}\right| - \frac{c_\ell}{4} \\
&\ge c_\ell - \frac{c_\ell}{4} - \frac{c_\ell}{4} = \frac{c_\ell}{2} (\text{ by Assumptions 2.2 and 2.3}) .
\end{aligned}$$

Appendix F. Proof of Lemma 2

By the definition of k_T we have

$$\begin{aligned}
\sqrt{|T|} k_T^{-1} &= \sup_\beta \frac{\|(\widetilde{\Gamma}\beta)_T\|_1}{\|\beta\|_2} \\
&= \sup_{\beta: \|\beta\|_2 = 1} \|(\widetilde{\Gamma}\beta)_T\|_1 \\
&= \sup_{\beta: \|\beta\|_2 = 1} \lambda_{\mathrm{TV}} \|(\Gamma\beta)_{T_1}\|_1 + \|\beta_{T_2}\|_1 \\
&\le \sup_{\beta: \|\beta\|_2 = 1} \lambda_{\mathrm{TV}} \|(\Gamma\beta)_{T_1}\|_1 + \sqrt{|T_2|} \|\beta\|_2 \\
&\le \sup_{\beta: \|\beta\|_2 = 1} \lambda_{\mathrm{TV}} \|(\Gamma\beta)_{T_1}\|_1 + \sqrt{|T_2|} .
\end{aligned}$$

Next we will bound the term $\|(\Gamma\beta)_{T1}\|_1$. First note that

$$
\begin{aligned}
\|(\Gamma\beta)_{T_1}\|_1 \leq & \sqrt{|T_1|}\|(\Gamma\beta)_{T_1}\|_2 \\
\leq & \sqrt{|T_1| \sum_{(j,k) \in E \cap T_1} \left|\widehat{\Sigma}_{j,k}\right|\left|\beta_j - \operatorname{sign}\left(\widehat{\Sigma}_{j,k}\right)\beta_k\right|^2} \\
\leq & \sqrt{|T_1| \sum_{(j,k) \in E \cap T_1} \left|\widehat{\Sigma}_{j,k}\right|\left(2\left|\beta_j\right|^2 + 2\left|\beta_k\right|^2\right)} \\
\leq & \sqrt{|T_1| \sum_{j=1}^{p} \left(\sum_{k:(j,k) \in E \cap T_1} 2\left|\widehat{\Sigma}_{j,k}\right|\right)\left|\beta_j\right|^2} \\
\leq & \sqrt{|T_1|}\sqrt{\max_{1 \leq j \leq p}\left[\left(\sum_{k:(j,k) \in E \cap T_1} 2\left|\widehat{\Sigma}_{j,k}\right|\right)\right]}\sqrt{\sum_{j=1}^{p}\left|\beta_j\right|^2} \\
\leq & \sqrt{|T_1|}\sqrt{\max_{1 \leq j \leq p}\left[\left(\sum_{k:(j,k) \in E \cap T_1} 2\left|\widehat{\Sigma}_{j,k}\right|\right)\right]} \\
\leq & \sqrt{|T_1|}\sqrt{2\|\widehat{\Sigma}\|_{1,1}}\,.
\end{aligned}
$$

Thus

$$
k_T^{-1} \leq \frac{\lambda_{\mathrm{TV}}\sqrt{2\|\widehat{\Sigma}\|_{1,1}|T_1|} + \sqrt{|T_2|}}{\sqrt{|T_1| + |T_2|}},
$$

which completes the proof.

Appendix G. Proof of Lemma 3

Note that Γ is the edge incidence matrix and $L = \Gamma^\top\Gamma$ is the weighted graph Laplacian matrix. Let the singular value decomposition for Γ to be $\Gamma = U_{m \times p}D_{p \times p}V_{p \times p}^\top$. Next recall that $\widetilde{\Gamma} = \begin{bmatrix} \lambda_{\mathrm{TV}}\Gamma \\ I \end{bmatrix}$, then we have

$$
\begin{aligned}
\widetilde{\Gamma}^\dagger &= \left(\lambda_{\mathrm{TV}}^2\Gamma^\top\Gamma + I\right)^{-1}\left[\lambda_{\mathrm{TV}}\Gamma^\top \quad I\right] \\
&= \left(\lambda_{\mathrm{TV}}^2 V D^2 V^\top + I\right)^{-1}\left[\lambda_{\mathrm{TV}} V D U^\top \quad I\right] \\
&= V\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-1} V^\top\left[\lambda_{\mathrm{TV}} V D U^\top \quad I\right] \\
&= \left[\underbrace{V\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-1}\lambda_{\mathrm{TV}} D U^\top}_{=:A} \quad \underbrace{V\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-1} V^\top}_{=:B}\right].
\end{aligned}
$$

From the definition of ρ we can see that the maximum diagonal entry of $\left(\widetilde{\Gamma}^\dagger\right)^\top\widetilde{\Gamma}^\dagger$ will just be ρ^2. Since

$$
\left(\widetilde{\Gamma}^\dagger\right)^\top\widetilde{\Gamma}^\dagger = \begin{bmatrix} A^\top A & A^\top B \\ B^\top A & B^\top B \end{bmatrix},
$$

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_ks. Let μ_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 be the i^{th} column of V, and $\tilde{v}_i^\top$ be the i^{th} row of V, i.e. $V := \left[v_1, \ldots, v_p\right] = \left[\tilde{v}_1^\top, \ldots, \tilde{v}_p^\top\right]^\top$. Similarly, we define u_i be the i^{th} column of U and $\tilde{u}_i^\top$ be the i^{th} row of U. Note that $\tilde{v}_i$ 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:

$$B^\top B = V\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-2} V^\top,$$

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

$$\begin{aligned}
\operatorname{maxdiag}\left(\boldsymbol{B}^\top \boldsymbol{B}\right) = & \max_{i \in \{1, \ldots, p\}} \tilde{v}_i^\top \left(\lambda_{\mathrm{TV}}^2 \boldsymbol{D}^2 + \boldsymbol{I}\right)^{-2} \tilde{v}_i \\
= & \max_{i \in \{1, \ldots, p\}} \sum_{j=1}^{p} \frac{v_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 \boldsymbol{D}_{jj}^2 + 1\right)^2} \\
= & \max_{i \in \{1, \ldots, p\}} \sum_{j \in B_{k(i)}} \frac{v_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 \boldsymbol{D}_{jj}^2 + 1\right)^2} \\
\leq & \max_{i \in \{1, \ldots, p\}} \left\{ \frac{1}{\left|\boldsymbol{B}_{k(i)}\right|} + \sum_{\substack{j \in B_{k(i)}: \\ D_{jj}^2 > 0}} \frac{v_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 \boldsymbol{D}_{jj}^2 + 1\right)^2} \right\} \\
\leq & \max_{i \in \{1, \ldots, p\}} \left\{ \frac{1}{\left|\boldsymbol{B}_{k(i)}\right|} + \frac{1}{\left(\lambda_{\mathrm{TV}}^2 \mu_{k(i)} + 1\right)^2} \sum_{\substack{j \in B_{k(i)}: \\ D_{jj}^2 > 0}} v_{j,i}^2 \right\} \\
\leq & \max_{i \in \{1, \ldots, p\}} \left\{ \frac{1}{\left|\boldsymbol{B}_{k(i)}\right|} + \frac{1}{\left(\lambda_{\mathrm{TV}}^2 \mu_{k(i)} + 1\right)^2} \right\} \\
\leq & \max_{k \in \{1, \ldots, K\}} \left\{ \frac{1}{\left|\boldsymbol{B}_k\right|} + \frac{1}{\left(\lambda_{\mathrm{TV}}^2 \mu_k + 1\right)^2} \right\}.
\end{aligned} \tag{G.1}$$

On the other hand we note that

$$A^{\top}A = U\lambda_{\mathrm{TV}}^2 D^2\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-2} U^{\top},$$

similarly the maximum diagonal element for $A^{\top}A$ can be upper bounded as:

$$\begin{aligned}
\max \operatorname{diag}\left(A^{\top}A\right) = & \max_{i \in \{1,\ldots,m\}} \tilde{u}_i^{\top} \lambda_{\mathrm{TV}}^2 D^2\left(\lambda_{\mathrm{TV}}^2 D^2 + I\right)^{-2} \tilde{u}_i \\
= & \max_{i \in \{1,\ldots,m\}} \sum_{j=1}^{p} \frac{\lambda_{\mathrm{TV}}^2 D_{jj}^2 u_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1\right)^2} \\
= & \max_{i \in \{1,\ldots,m\}} \sum_{j=1}^{p} \frac{\left(\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1 - 1\right) u_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1\right)^2} \\
= & \max_{i \in \{1,\ldots,m\}} \sum_{j=1}^{p} \left\{ \frac{u_{j,i}^2}{\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1} - \frac{u_{j,i}^2}{\left(\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1\right)^2} \right\} \\
\le & \max_{i \in \{1,\ldots,m\}} \sum_{j \in \left\{1,\ldots,p\right\}: D_{jj}^2 > 0} \left\{ \frac{u_{j,i}^2}{\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1} \right\} \\
\le & \max_{i \in \{1,\ldots,m\}} \max_{j \in \{1,\ldots,p\}: D_{jj}^2 > 0} \frac{1}{\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1} \sum_{j=1}^{p} u_{j,i}^2 \\
\le & \max_{i \in \{1,\ldots,m\}} \max_{j \in \{1,\ldots,p\}: D_{jj}^2 > 0} \frac{1}{\lambda_{\mathrm{TV}}^2 D_{jj}^2 + 1} \\
\le & \max_{k \in \{1,\ldots,K\}} \frac{1}{\lambda_{\mathrm{TV}}^2 \mu_k + 1}.
\end{aligned} \tag{G.2}$$

Then by combining the results above we have

$$\begin{aligned}
\rho^2 \le & \max_{1 \le k \le K} \left\{ \frac{1}{|B_k|} + \frac{1}{\left(\lambda_{\mathrm{TV}}^2 \mu_k + 1\right)^2} + \frac{1}{\lambda_{\mathrm{TV}}^2 \mu_k + 1} \right\} \\
\le & \max_{1 \le k \le K} \left\{ \frac{1}{|B_k|} + \frac{2}{\lambda_{\mathrm{TV}}^2 \mu_k + 1} \right\}.
\end{aligned}$$

This completes the proof of Lemma 3.

Appendix H. Proof of Lemma 4

By the definition of the block complete graph in Section 2.2.1 we can see that $|B_k| = \frac{p}{K}$ for $1 \le k \le K$ thus we have $\max_{1 \le k \le K} \frac{1}{|B_k|} = \frac{K}{p}$. Note that μ_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 $1 \le k \le K$. Also note that $\lambda_{\min}(\Sigma) = a(1 - r)$, so we have

$$
\begin{aligned}
\lambda_{\min}(\Sigma + \lambda_S L) &= \lambda_{\min}\left[(1-\lambda_S)\Sigma + \lambda_S(\Sigma + L)\right] \\
&\geq (1-\lambda_S)\lambda_{\min}(\Sigma) + \lambda_S \lambda_{\min}(\Sigma + L) \\
&= (1-\lambda_S)a(1-r) + \lambda_S\left[a + ar\left(\frac{p}{K} - 1\right)\right] (\text{by } ((\text{E.1})) \text{ and } \widehat{\Sigma} = \Sigma) \\
&\geq (1-\lambda_S)(1-r)\frac{K}{p} + \lambda_S r (\text{by using } a = \frac{K}{p}) .
\end{aligned}
$$

This completes the proof of Lemma 4.

Appendix I. Proof of Lemma 5

Let Γ be the edge incidence matrix for the chain graph and let $\Gamma = UDV^T$ denote the SVD of Γ. Note that the chain graph has one connected component, so in the language of Lemma 3 we have $|B_1| = p$. From Equations (G.2) and (G.3) in the proof of Lemma 3 it follows that

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

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

$$
\frac{1}{p} + \sum_{j: D_{j,j}^2 > 0} \frac{v_{j,i}^2}{\left(\lambda_{TV}^2 D_{j,j}^2 + 1\right)^2} . \tag{I.1}
$$

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

$$
v_{j,i} = \begin{cases} \sqrt{\frac{1}{p}} & \text{for } j = 1, i \in \{1, \ldots, p\} \\ \sqrt{\frac{2}{p}} \cos\left(\frac{(i - 1/2)(j-1)\pi}{p}\right), & \text{for } j \geq 2, i \in \{1, \ldots, p\} \end{cases}
$$

so in particular, $v_{j,i}^2 \leq \frac{2}{p}$ for all *i, j*. Thus Equation (I.1) is

$$
\leq \frac{1}{p} + \frac{2}{p} \sum_{j: D_{j,j}^2 > 0} \frac{1}{\left(\lambda_{TV}^2 D_{j,j}^2 + 1\right)^2} .
$$

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 $\sigma_j = 2 - 2\cos\left(\frac{j\pi}{p}\right)$ for $j = 1, \dots, p - 1$. We have $2 - 2\cos\left(\frac{j\pi}{p}\right) \geq \frac{j^2}{p^2}$ for $1 \leq j \leq p - 1$ so this is

$$\begin{aligned}
&\leq \frac{1}{p} + \frac{2}{p}\sum_{j=1}^{p-1} \frac{1}{\left(\frac{r^2\lambda_{TV}^2 j^2}{p^2} + 1\right)^2} \\
&= \frac{1}{p} + 2p^3 \sum_{j=1}^{p-1} \frac{1}{\left(r^2\lambda_{TV}^2 j^2 + p^2\right)^2} \\
&= \frac{1}{p} + \frac{2p^3}{r^4\lambda_{TV}^4} \sum_{j=1}^{p-1} \frac{1}{\left(j^2 + \left(\frac{p}{r\lambda_{TV}}\right)^2\right)^2} .
\end{aligned}$$

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

$$\begin{aligned}
&\leq \frac{1}{p} + \frac{2p^3}{r^4\lambda_{TV}^4}\int_{x=0}^{\infty} \frac{1}{(x^2 + (\frac{p}{r\lambda_{TV}})^2)^2} dx \\
&= \frac{1}{p} + \frac{2p^3}{r^4\lambda_{TV}^4} \frac{\pi}{4\left(\frac{p}{r\lambda_{TV}}\right)^3} = \frac{1}{p} + \frac{\pi}{2r\lambda_{TV}} .
\end{aligned} \tag{I.2}$$

Left singular vectors

We next focus on bounding

$$\sum_{j:D_{j,j}^2 > 0} \frac{u_{j,i}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1} . \tag{I.3}$$

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

$$\Gamma\Gamma_{i,j}^T = \begin{cases} 2 \text{ if } i = j \\ -1 \text{ if } |i - j| = 1 \\ 0 \text{ otherwise} \end{cases}$$

Strang (2007, Section 1.5), gives $p - 1$ orthonormal eigenvectors u_j of $\Gamma\Gamma^T$ which are of the form $u_{j,i} = \sqrt{\frac{2}{p}}\sin\left(\frac{\pi ij}{p}\right)$. In particular $u_{j,i}^2 \leq \frac{2}{p}$ so Equation (I.3) is

$$\leq \frac{2}{p} \sum_{j:\, D_{j,j}^2 > 0} \frac{1}{\lambda_{TV}^2 D_{j,j}^2 + 1}\,. \tag{I.4}$$

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 $2 - 2\cos\left(\frac{\pi j}{p}\right)$ for $j = 1, \ldots\, p - 1$ and since $2 - 2\cos\left(\frac{\pi j}{p}\right) \geq \frac{j^2}{p^2}$ for $j = 1, \ldots, p - 1$ we get that this is

$$\begin{aligned} &\leq \frac{2}{p} \sum_{j=1}^{p-1} \frac{1}{\frac{r^2 \lambda_{TV}^2 j^2}{p^2} + 1} \\ &= 2p \sum_{j=1}^{p-1} \frac{1}{p^2 + r^2 \lambda_{TV}^2 j^2} \\ &= \frac{2p}{r^2 \lambda_{TV}^2} \sum_{j=1}^{p-1} \frac{1}{j^2 + \left(\frac{p}{r\lambda_{TV}}\right)^2}\,. \end{aligned}$$

Since $f(j) = \frac{1}{j^2 + \left(\frac{p}{r\lambda_{TV}}\right)^2}$ is montonically decreasing on $\mathbb{R}^+$ we have that this is

$$\begin{aligned} &\leq \frac{2p}{r^2 \lambda_{TV}^2} \int_{x=0}^{\infty} \frac{1}{x^2 + \left(\frac{p}{r\lambda_{TV}}\right)^2} dx \\ &= \frac{2p}{r^2 \lambda_{TV}^2} \frac{r\lambda_{TV}}{p} \arctan\left(\frac{r\lambda_{TV} x}{p}\right)\Bigg|_{x=0}^{x=\infty} = \frac{\pi}{r\lambda_{TV}}\,. \end{aligned} \tag{I.5}$$

Moreover, since $u_{j,i}^2$ and $v_{j,i}^2$ are bounded by $\frac{2}{p}$ we immediately have the bound

$$\rho^2 \leq \frac{1}{p} + 2.$$

Combining this with Equations (I.2) and (I.5) we conclude that

$$\rho^2 \leq \min\left(\frac{1}{p} + 2, \max(\frac{\pi}{r\lambda_{TV}}, \frac{1}{p} + \frac{\pi}{2r\lambda_{TV}})\right) \leq \frac{1}{p} + \frac{2\pi}{r\lambda_{TV} + 1}$$

as claimed.

For the final part of the proof, note that $\lambda_{\min}(\Sigma) = a\left[1 + 2r\cos\left(\frac{p}{p+1}\pi\right)\right]$ (see Noschese et al. (2013, Section 2)), so we have

$$\begin{aligned}\lambda_{\min}(\Sigma + \lambda_S L) &= \lambda_{\min}[(1-\lambda_S)\Sigma + \lambda_S(\Sigma + L)] \\ &\geq (1-\lambda_S)\lambda_{\min}(\Sigma) + \lambda_S\lambda_{\min}(\Sigma + L) \\ &= (1-\lambda_S)a\left[1 + 2r\cos\left(\frac{p}{p+1}\pi\right)\right] + \lambda_S a(1+r)(\text{by } (E.1) \text{ and } \widehat{\Sigma} = \Sigma) \\ &\geq (1-\lambda_S)\left[1 + 2r\cos\left(\frac{p}{p+1}\pi\right)\right] + \lambda_S(\text{ by using } a = 1) \\ &\geq (1-\lambda_S)(1-2r) + \lambda_S\,.\end{aligned}$$

This completes the proof of Lemma 5.

Appendix J. 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 (G.2) and (G.3) in the proof of Lemma 3 it follows that

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

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

$$\frac{1}{p} + \sum_{j: D_{j,j}^2 > 0} \frac{v_{j,i}^2}{\left(\lambda_{TV}^2 D_{j,j}^2 + 1\right)^2}\,. \tag{J.1}$$

The v_j correspond to the normalized eigenvectors of the unweighted Laplacian for the Lattice graph. We denote the Laplacian L_{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

$$L_{Lat} = L_{\sqrt{p}} \otimes I_{\sqrt{p}} + I_{\sqrt{p}} \otimes L_{\sqrt{p}}\,.$$

Let $\{w_k\}_{j = 1,\ldots,\sqrt{p}}$ denote the normalized eigenvectors of $L_{\sqrt{p}}$ and σ_{k-1} the corresponding eigenvalues. Then

$$\begin{aligned}L_{Lat}(w_k \otimes w_l) &= L_{\sqrt{p}} w_k \otimes I_{\sqrt{p}} w_l + I_{\sqrt{p}} w_k \otimes L_{\sqrt{p}} w_l \\ &= \sigma_{k-1} w_k \otimes w_l + w_k \otimes \sigma_{l-1} w_l = (\sigma_{k-1} + \sigma_{l-1})(w_k \otimes w_l)\,.\end{aligned}$$

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

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

We have $\frac{D_{j,j}^2}{r^2} = \lambda_j$ where λ_j denotes the *j*th eigenvalue of L_{Lat}. We concluded above that the eigenvalues of L_{Lat} are of the form $\sigma_k + \sigma_l$ where $\{\sigma_k\}_{k = 0,\ldots,\sqrt{p}-1}$ 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

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

Algebraic rearrangement gives that this is

$$\begin{aligned} &= \frac{1}{p} + 4p\sum_{k=1}^{\sqrt{p}}\sum_{l=1}^{\sqrt{p}} \frac{1}{\left(r^2\lambda_{TV}^2\left(k^2+l^2\right)+p\right)^2} + 8p\sum_{k=1}^{\sqrt{p}} \frac{1}{\left(r^2\lambda_{TV}^2k^2+p\right)^2} \\ &= \frac{1}{p} + \frac{4p}{r^4\lambda_{TV}^4}\sum_{k=1}^{\sqrt{p}}\sum_{l=1}^{\sqrt{p}} \frac{1}{\left(k^2+l^2+\frac{p}{r^2\lambda_{TV}^2}\right)^2} + \frac{8p}{r^4\lambda_{TV}^4}\sum_{k=1}^{\sqrt{p}} \frac{1}{\left(k^2+\frac{p}{r^2\lambda_{TV}^2}\right)^2}. \end{aligned}$$

The above functions are monotonically decreasing in *k* and *l* for *k,l* ≥ 0 and so we can say this is

$$\begin{aligned} &\leq \frac{1}{p} + \frac{4p}{r^4\lambda_{TV}^4}\int_{x=0}^{\infty}\int_{y=0}^{\infty} \frac{1}{\left(x^2+y^2+\frac{p}{r^2\lambda_{TV}^2}\right)^2}dydx + \frac{8p}{r^4\lambda_{TV}^4}\int_{x=0}^{\infty} \frac{1}{\left(x^2+\frac{p}{r^2\lambda_{TV}^2}\right)^2}dx \\ &= \frac{1}{p} + \frac{4p}{r^4\lambda_{TV}^4}\frac{\pi r^2\lambda_{TV}^2}{4p} + \frac{8p}{r^4\lambda_{TV}^4}\frac{\pi r^3\lambda_{TV}^3}{4p^{3/2}} = \frac{1}{p} + \frac{4\pi}{r^2\lambda_{TV}^2} + \frac{8\pi}{r\lambda_{TV}\sqrt{p}} \\ &. \end{aligned} \tag{J.2}$$

Left singular vectors

We next focus on bounding

$$\sum_{j:\, D_{j,j}^2 > 0} \frac{u_{j,i}^2}{\lambda_{TV}^2 D_{j,j}^2 + 1}. \tag{J.3}$$

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 $u_{j,i}^2 \le \frac{4}{p}$ for all i, j and so Equation (J.3) is

$$\le \frac{4}{p} \sum_{j:\, D_{j,j}^2 > 0} \frac{1}{\lambda_{TV}^2 D_{j,j}^2 + 1}.$$

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

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

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

$$\frac{4}{p} \sum_{j:\, D_{j,j}^2 > 0} \frac{1}{\lambda_{TV}^2 D_{j,j}^2 + 1} \le \frac{4}{p} \sum_{k=0}^{\sqrt{p}} \sum_{l=0}^{\sqrt{p}} \frac{\mathbb{1}(k,l) \ne (0,0)}{r^2 \lambda_{TV}^2 \frac{k^2 + l^2}{p} + 1}.$$

Algebraic manipulation gives that this is

$$\begin{aligned} &= 4 \sum_{k=1}^{\sqrt{p}} \sum_{l=1}^{\sqrt{p}} \frac{1}{r^2 \lambda_{TV}^2 \left(k^2 + l^2\right) + p} + 8 \sum_{k=1}^{\sqrt{p}} \frac{1}{r^2 \lambda_{TV}^2 k^2 + p} \\ &= \frac{4}{r^2 \lambda_{TV}^2} \sum_{k=1}^{\sqrt{p}} \sum_{l=1}^{\sqrt{p}} \frac{1}{k^2 + l^2 + \frac{p}{r^2 \lambda_{TV}^2}} + \frac{8}{r^2 \lambda_{TV}^2} \sum_{k=1}^{\sqrt{p}} \frac{1}{k^2 + \frac{p}{r^2 \lambda_{TV}^2}}. \end{aligned}$$

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

$$\begin{aligned} &\le \frac{4}{r^2 \lambda_{TV}^2} \int_{x=0}^{\sqrt{p}} \int_{y=0}^{\infty} \frac{1}{x^2 + y^2 + \frac{p}{r^2 \lambda_{TV}^2}} dy dx + \frac{8}{r^2 \lambda_{TV}^2} \int_{x=0}^{\infty} \frac{1}{x^2 + \frac{p}{r^2 \lambda_{TV}^2}} dx \\ &= \frac{2\pi}{r^2 \lambda_{TV}^2} \int_{x=0}^{\sqrt{p}} \frac{1}{\sqrt{x^2 + \frac{p}{r^2 \lambda_{TV}^2}}} dx + \frac{8}{r^2 \lambda_{TV}^2} \frac{\pi r \lambda_{TV}}{2\sqrt{p}}. \end{aligned} \tag{J.4}$$

We compute this last integral explicitly as

$$\begin{aligned}\int_{x=0}^{\sqrt{p}} \frac{1}{\sqrt{x^2 + \frac{p}{r^2\lambda_{TV}^2}}} dx &= \frac{\pi}{2}\log(\sqrt{\frac{p}{r^2\lambda_{TV}^2} + x^2} + x)\Big|_{x=0}^{x=\sqrt{p}} \\ &= \frac{\pi}{2}\log\left(\frac{\sqrt{\frac{p}{r^2\lambda_{TV}^2} + p} + \sqrt{p}}{\sqrt{\frac{p}{r^2\lambda_{TV}^2}}}\right).\end{aligned}$$

Some additional algebra, along with the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a,b \ge 0$ gives that this is

$$\le \frac{\pi}{2}\log(2 + r\lambda_{TV}).$$

Overall we've concluded that Equation (J.4) is

$$\le \frac{\pi^2\log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2} + \frac{8\pi}{r\lambda_{TV}\sqrt{p}}. \qquad \text{(J.5)}$$

Moreover, since $u_{i,j}^2$ and $v_{i,j}^2$ are bounded by $\frac{4}{p}$ we immediately have the bound

$$\rho^2 \le 5.$$

Combining this with Equations (J.2) and (J.5) we conclude that

$$\rho^2 \le \min\left(5, \frac{1}{p} + \frac{4\pi\log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2} + \frac{8\pi}{r\lambda_{TV}\sqrt{p}}\right) \le \frac{1}{p} + \frac{5\pi\log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2 + 1} + \frac{10\pi}{r\lambda_{TV}\sqrt{p} + 1}$$

as claimed.

For the final part of the proof recall that $r \in \left(0, \frac{1}{4}\right)$. Thus Σ is diagonally dominant with $\Sigma_{i,i} - \sum_{j \ne i}\Sigma_{i,j} \ge 1 - 4r > 0$ for all i and therefore $\lambda_{\min}(\Sigma) \ge 1 - 4r$. This implies that

$$\begin{aligned}\lambda_{\min}(\Sigma + \lambda_S L) &= \lambda_{\min}[(1 - \lambda_S)\Sigma + \lambda_S(\Sigma + L)] \\ &\ge (1 - \lambda_S)\lambda_{\min}(\Sigma) + \lambda_S\lambda_{\min}(\Sigma + L) \\ &= (1 - \lambda_S)(1 - 4r) + \lambda_S(1 + 2r)(\text{ by } (E.1) \text{ and } \widehat{\Sigma} = \Sigma) \\ &\ge (1 - \lambda_S)(1 - 4r) + \lambda_S.\end{aligned}$$

This completes the proof of Lemma 6.

Appendix K. Extension to Logistic Regression

In the main body of the paper we consider only a linear model in the interest of simplicity. However, it is straightforward to extend the theory in this paper to generalized linear models.

In this section we need to assume that $\|\beta^*\|_1 \le u$ for a universal constant u. We will informally sketch an extension to logistic regression. Consider a logistic model where

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\lambda_i) \\ \lambda_i &= \frac{1}{1 + \exp(-\langle \beta^*, X_i \rangle)}\,. \end{aligned}$$

Instead of using squared loss, we want to use the logistic loss function

$$L(\beta; X, y) = \sum_{i=1}^{n} \log(1 + \exp(\langle \beta^*, X_i \rangle)) - y_i \langle \beta^*, X_i \rangle\,.$$

The GTV estimator for the logistic model takes the form

$$\begin{aligned} \hat{\beta} = \underset{\beta : \|\beta\|_1 \le u}{\arg\min}\ & \frac{1}{n} L(\beta) + \lambda_S \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right| \left(\beta_j - \hat{s}_{j,k} \beta_k\right)^2 \\ & + \lambda_1 \left(\lambda_{\mathrm{TV}} \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right|^{1/2} \left|\beta_j - \hat{s}_{j,k} \beta_k\right| + \|\beta\|_1 \right). \end{aligned} \tag{K.1}$$

For convenience define

$$R(\beta) := \lambda_S \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right| \left(\beta_j - \hat{s}_{j,k} \beta_k\right)^2 + \lambda_1 \left(\lambda_{\mathrm{TV}} \sum_{j,k} \left|\hat{\Sigma}_{j,k}\right|^{1/2} \left|\beta_j - \hat{s}_{j,k} \beta_k\right| + \|\beta\|_1 \right).$$

To derive similar theoretical bounds in this setting, first note that by definition

$$\frac{1}{n} L(\hat{\beta}) \le \frac{1}{n} L(\beta^*) + \left(R(\beta^*) - R(\hat{\beta}) \right).$$

We now use standard steps for the analysis of generalized models in order to reduce our problem to the linear setting in the proof of Theorem 1. For the remainder of the section, we use the shorthand $f(x) = \log(1 + \exp(x))$. Using the definition of $L(\beta)$ and rearranging terms yields

$$\sum_{i=1}^{n} \frac{1}{n} f(\langle \hat{\beta}, X_i \rangle) - f(\langle \beta^*, X_i \rangle) - y_i \langle \Delta, X_i \rangle \le R(\beta^*) - R(\hat{\beta})\,.$$

Define $\epsilon_i := y_i - \mathbb{E}[y_i \mid X_i] = y_i - f'(\langle \beta^*, X_i \rangle)$ and then

$$\begin{aligned} & \frac{1}{n} f(\langle \hat{\beta}, X_i \rangle) - f(\langle \beta^*, X_i \rangle) - f'(\langle \beta^*, X_i \rangle) \langle \Delta, X_i \rangle \le \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle \Delta, X_i \rangle + R(\beta^*) \\ & - R(\hat{\beta})\,. \end{aligned} \tag{K.2}$$

For x,y contained in an interval $[-d, d]$, f is a strongly convex function so that

$$f(x) - f(y) - f'(y)(x - y) \geq \psi \|x - y\|_2^2$$

for a strong convexity parameter ψ which depends on d. Applying this to Equation (K.2),

$$\frac{\psi}{n}\langle \Delta, X_i \rangle^2 \leq \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \langle \Delta, X_i \rangle + R(\beta^*) - R(\widehat{\beta}) . \tag{K.3}$$

Assuming $\|\beta^*\|_1, \|\widehat{\beta}\|_1 \leq u$ for a universal constant u, the convexity parameter ψ is also bounded by a universal constant. Rearranging terms and ignoring the factor of ψ, the inequality in Equation (K.3) is exactly the inequality at the beginning of the proof of Theorem 1. Thus all the bounds derived in the linear setting also apply in the logistic regression setting up to a factor of the convexity parameter ψ.

Appendix L. Proof of Lemma 7

We have,

$$\begin{aligned} \epsilon^\top X\Delta &= \epsilon^\top X\widetilde{\Gamma}^\dagger \widetilde{\Gamma}\Delta \\ &\leq \left\| \left(X\widetilde{\Gamma}^\dagger \right)^\top \epsilon \right\|_\infty \|\widetilde{\Gamma}\Delta\|_1, \end{aligned}$$

where $\widetilde{\Gamma}^\dagger$ is the pseudo-inverse of $\widetilde{\Gamma}$, since $\widetilde{\Gamma}^\dagger \widetilde{\Gamma} = I_{p \times p}$. Also,

$$\begin{aligned} \epsilon^\top X\Delta &\leq \|X^\top \epsilon\|_\infty \|\Delta\|_1 \\ &\leq \|X^\top \epsilon\|_\infty \|\widetilde{\Gamma}\Delta\|_1, \end{aligned}$$

since $\|\widetilde{\Gamma}\Delta\|_1 = \lambda_{TV}\|\Gamma\Delta\|_1 + \|\Delta\|_1 \geq \|\Delta\|_1$. Therefore, in the following, we will bound $\|(XD)^\top \epsilon\|_\infty$ when $D = \widetilde{\Gamma}^\dagger$ or $I_{p \times p}$. In particular, provided that D is *independent* of X, we will establish

$$\frac{1}{n}\|(XD)^\top \epsilon\|_\infty \leq 6\sigma\rho_D\sqrt{\frac{c_u \log p}{n}}$$

with probability at least $1-C/p$, for an absolute constant $C > 0$, where we define $\rho_D := \max_{1 \leq j \leq n_D}\|D_j\|_2$ and n_D is the number of columns of D (i.e. $n_D = m+p$ when $D = \widetilde{\Gamma}^\dagger$, and $n_D = p$ when $D = I_{p \times p}$). We note that both choices of D are allowed when Γ is independent of X, but we force $D = I_{p \times p}$ when Γ is dependent of X.

First, we define the event,

$$\mathscr{E} := \left\{ \max_{1 \leq j \leq n_D} \|XD_j\|_2^2 \leq 2\rho_D^2 c_u n \right\} .$$

We have,

$$\begin{aligned}P\left(\max_{1 \le j \le n_D}\left|\frac{\left(XD_j\right)^\top \epsilon}{n}\right| \ge t\right) &\le \mathbb{E}\left(\mathbb{1}\left\{\max_{1 \le j \le n_D}\left|\frac{\left(XD_j\right)^\top \epsilon}{n}\right| \ge t\right\} \cdot \mathbb{1}_{\mathscr{E}}\right) + P\left(\mathscr{E}^c\right) \\ &\le \mathbb{E}\left(P\left(\max_{1 \le j \le n_D}\left|\frac{\left(XD_j\right)^\top \epsilon}{n}\right| \ge t \mid \left(X^{(i)}\right)_{i=1}^n\right) \cdot \mathbb{1}_{\mathscr{E}}\right) \\ &+ P\left(\mathscr{E}^c\right)\end{aligned} \tag{L.1}$$

Given $\left(X^{(i)}\right)_{i=1}^n$, note that $\forall j$,

$$\frac{\left(XD_j\right)^\top \epsilon}{n} \sim \mathcal{N}(0, \frac{\sigma^2}{n^2}\|XD_j\|_2^2).$$

Then, by a gaussian tail bound and a union bound, for any $t > 0$,

$$P\left(\max_{1 \le j \le n_D}\left|\frac{\left(XD_j\right)^\top \epsilon}{n}\right| \ge t \mid \left(X^{(i)}\right)_{i=1}^n\right) \le \sum_{j=1}^{n_D} 2\exp\left(-\frac{n^2t^2}{2\sigma^2\|XD_j\|_2^2}\right).$$

It follows,

$$\begin{aligned}&\mathbb{E}\left(P\left(\max_{1 \le j \le n_D}\left|\frac{\left(XD_j\right)^\top \epsilon}{n}\right| \ge t \mid \left(X^{(i)}\right)_{i=1}^n\right) \cdot \mathbb{1}_{\mathscr{E}}\right) \\ &\le 2\mathbb{E}\left(\sum_{j=1}^{n_D} \exp\left(-\frac{n^2t^2}{2\sigma^2\|XD_j\|_2^2}\right) \cdot \mathbb{1}_{\mathscr{E}}\right) \\ &\le 2\exp\left(-\frac{nt^2}{4\sigma^2\rho_D^2c_u} + \log n_D\right) \\ &\le 2/p,\end{aligned} \tag{L.2}$$

noting $n_D \le p^2 + p$ for both choices of D, with the choice of $t^2 = 36\sigma^2\rho_D^2c_u\log p/n$.

Now we calculate $P\left(\mathscr{E}^c\right)$. For any fixed unit vector v, $v^\top X^{(i)}$ is sub-gaussian with parameter c_u. Let $v_j := D_j/\|D_j\|_2$. By Bernstein's inequality (e.g. Theorem 2.8.1 in Vershynin (2018)) and the fact that $\|Xv\|_2^2 = \sum_{i=1}^n \left(v^\top X^{(i)}\right)^2$,

$$P\left(\max_{1 \le j \le n_D}\left|\|Xv_j\|_2^2 - \mathbb{E}\left[\|Xv_j\|_2^2\right| \ge c_u n\delta\right) \le 2\exp\left(-cn\left(\delta^2 \wedge \delta\right) + \log n_D\right) \le 2/p$$

taking $\delta^2 = 2\log n_D/cn$, provided $n \ge (2/c)\log n_D$, since $n_D \ge p$. Then w.p. $1 - 2/p$, for all j

$$\begin{aligned}\|Xv_j\|_2^2 &\le \mathbb{E}\left[\|Xv_j\|_2^2\right] + c_u n\delta \\ &\le nc_u(1+\delta).\end{aligned}$$

Then, with the same probability,

$$\max_{1 \le j \le n_D} \|XD_j\|_2^2 \le \rho_D^2 nc_u(1 + \sqrt{2\log n_D/cn}) \le 2\rho_D^2 nc_u .$$

Therefore

$$P(\mathscr{E}^c) = P(\max_{1 \le j \le n_D} \|XD_j\|_2^2 \ge 2\rho_D^2 c_u n) \le 2/p . \tag{L.3}$$

Combining (L.1), (L.2), and (L.3), we conclude

$$P\left(\max_{1 \le j \le n_D}\left|\frac{(XD_j)^\top \epsilon}{n}\right| \ge t\right) \le 4/p \tag{L.4}$$

for the choice of $t = 6\sigma\rho_D\sqrt{c_u \log p/n}$.

Appendix M. Proof of Lemma 8

The proof of Lemma 8 involves two parts.

Part 1: We first show that the following inequality

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{1}{4}\|\Sigma^{1/2}\Delta\|_2 - 9\frac{\lambda_1}{\sigma}\|\widetilde{\Gamma}\Delta\|_1 \tag{M.1}$$

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

1. (1) Since we want to lower bound $\frac{\|X\Delta\|_2}{\sqrt{n}}$ in terms of $\|\Sigma^{1/2}\Delta\|_2$ and $\|\widetilde{\Gamma}\Delta\|_1$, we define the set $V(r) := \left\{\Delta \in \mathbb{R}^p \mid \|\Sigma^{1/2}\Delta\|_2 = 1, \|\widetilde{\Gamma}\Delta\|_1 \le r\right\}$ 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 E[$M(r,X)$], where $M(r,X)$ is defined as:

$$M(r, X) := 1 - \inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}} = \sup_{\Delta \in V(r)}\left\{1 - \frac{\|X\Delta\|_2}{\sqrt{n}}\right\}.$$

1. (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 $\mathbb{E}[M(r, X)]$.

2. (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, so that we can show that (M.1) holds with high probability.

In the following proof we only provide details for proving step (1), and sketch the proof for step (2) and (3) since our proof for step (2) and (3) are almost identical to those in Raskutti et al. (2010).

For step (1) we prove the following lemma:

Lemma 9

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 $S^{n-1} = \left\{u \in \mathbb{R}^n \mid \|u\|_2 = 1\right\}$. 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'e inequality, if

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

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

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

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

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

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

$$\begin{aligned} \mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u, \Delta}\right] &\leq \mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Z_{u, \Delta}\right] \\ &= \mathbb{E}\left[\inf_{u \in S^{n-1}} g^\top u\right] + \mathbb{E}\left[\sup_{\Delta \in V(r)} h^\top \Sigma^{1/2} \Delta\right] \\ &= -\mathbb{E}[\|g\|_2] + \mathbb{E}\left[\sup_{\Delta \in V(r)} h^\top \Sigma^{1/2} \Delta\right]. \end{aligned}$$

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

Lemma 10

Suppose we have $\lambda_1 \geq 48 \rho \sigma \sqrt{\frac{c_u \log p}{n}}$, *then we have that*

$$\lambda_1 \geq 8 \frac{\sigma}{\sqrt{n}} \mathbb{E}\left[\left\|\left(\Sigma^{1/2} \tilde{\Gamma}^\dagger\right)^\top h\right\|_\infty\right].$$

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

The proof for this lemma will be provided shortly. Thus

$$\begin{aligned} \mathbb{E}\left[\sup_{\Delta \in V(r)} \left|h^\top \Sigma^{1/2} \Delta\right|\right] &= \mathbb{E}\left[\sup_{\Delta \in V(r)} \left|h^\top \Sigma^{1/2} \tilde{\Gamma}^\dagger \tilde{\Gamma} \Delta\right|\right] \\ &\leq \mathbb{E}\left[\sup_{\Delta \in V(r)} \|h^\top \Sigma^{1/2} \tilde{\Gamma}^\dagger\|_\infty \|\tilde{\Gamma} \Delta\|_1\right] \\ &\leq \mathbb{E}\left[\left\|\left(\Sigma^{1/2} \tilde{\Gamma}^\dagger\right)^\top h\right\|_\infty\right] r \\ &\leq 3r \frac{\lambda_1}{\sigma} \sqrt{n}, \end{aligned}$$

where the last inequality holds with high probability from Lemma 10. Also by standard χ^2 tail bounds (Ledoux and Talagrand, 1991) when $n \geq 10$ we have $\mathbb{E}[\|g\|_2] \geq \frac{3}{4} \sqrt{n}$. By combining hese pieces together

$$\mathbb{E}\left[-\inf_{\Delta \in V(r)} \|X\Delta\|_2\right] \leq -\frac{3}{4} \sqrt{n} + 3r \frac{\lambda_1}{\sigma} \sqrt{n}.$$

Thus by dividing by $\sqrt{n}$ and adding 1 to both sides we have

$$\mathbb{E}[M(r,X)] = \mathbb{E}[1 - \inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}}] \le \frac{1}{4} + 3r\frac{\lambda_1}{\sigma}.$$

In step (2), similarly as in Raskutti et al. (2010), we can show that $M(r,X) = h(W) := \sup_{v \in V(r)}(1 - \|W\Sigma^{1/2}v\|_w/\sqrt{n})$ is $1/\sqrt{n}$ – Lipschitz in W, where $W \in \mathbb{R}^{n \times p}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then by the Gaussian concentration inequality,

$$P(\left|M(r,X) - \mathbb{E}[M(r,X)] \ge t(r)/2)\right|) \le 2\exp\left(-t^2(r)/8\right),$$

for $t(r) := 1/4 + 3r\lambda_1/\sigma$. Combining with step (1), we have

$$P\left(M(r,X) \ge \frac{3t(r)}{2}\right) \le 2\exp\left(-nt^2(r)/8\right). \tag{M.2}$$

Finally, in step (3), we use a standard peeling argument to extend (M.2) for all r. More concretely, we define the event

$$\mathcal{E} := \left\{\exists \Delta \in \mathbb{R}^p \text{ s.t. } \|\Sigma^{1/2}\Delta\|_2 = 1, \left(1 - \|X\Delta\|_2/\sqrt{n}\right) \ge 3t\left(\|\tilde{\Gamma}\Delta\|_1\right)\right\}.$$

By a peeling argument (e.g. Lemma 3 in Raskutti et al. (2010)), it can be established that $P\left[\mathcal{E}^c\right] \ge 1c_4\exp(c_5 n)$ for some numerical constants c_4 and c_5. On $\mathcal{E}^c$, we have,

$$1 - \frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{3}{4} + 9\frac{\lambda_1}{\sigma}\|\tilde{\Gamma}\Delta\|_1,$$

for all Δ such that $\|\Sigma^{1/2}\Delta\|_2 = 1$. Therefore, we conclude that with probability at least $1 - c_4 \exp(-c_5 n)$,

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{1}{4}\|\Sigma^{1/2}\Delta\|_2 - 9\frac{\lambda_1}{\sigma}\|\tilde{\Gamma}\Delta\|_1.$$

Part 2: Next we can go to the second part of the proof. From (B.1) and (B.2) we know that

$$\begin{aligned}\|\tilde{\Gamma}\Delta\|_1 &\le 4\|(\tilde{\Gamma}\Delta)_T\|_1 + 4\|\left(\tilde{\Gamma}\beta^*\right)_{T^c}\|_1 \\ &\le \frac{4\sqrt{|T|}\|\Delta\|_2}{k_T} + 4\|\left(\tilde{\Gamma}\beta^*\right)_{T^c}\|_1.\end{aligned}$$

Then

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{1}{4}\|\Sigma^{1/2}\Delta\|_2 - 9\frac{\lambda_1}{\sigma}\left(4\|\left(\tilde{\Gamma}\beta^*\right)_{T^c}\|_1 + \frac{4\sqrt{|T|}\|\Delta\|_2}{k_T}\right).$$

Thus there exist constants $c', c'' > 0$ such that

$$\Delta^\top\left(\frac{X^\top X}{n} + \lambda_S L\right)\Delta \geq c'\Delta^\top(\Sigma + \lambda_S L)\Delta - c''\lambda_1^2\left(\|(\widetilde{\Gamma}\beta *)_{T^c}\|_1^2 + \frac{\left|T\right| \|\Delta\|_2^2}{k_T^2}\right).$$

Since

$$\Delta^\top(\Sigma + \lambda_S L)\Delta \geq \lambda_{\min}(\Sigma + \lambda_S L)\|\Delta\|_2^2,$$

then when λ_1 satisfies (B.4) for some constant $c_2 > 0$,

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

for absolute constants $c_1, c_3 > 0$.

Appendix N. Proof of Lemma 10

We will use a classical Lemma for Gaussian processes by Anderson (1955) to prove our result.

Lemma 11

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

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

First note that $\Sigma \leq c_u I_{p \times p}$ and by using Lemma 11 we have for any $y > 0$ the following inequality

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

Since $h \sim \mathcal{N}\left(0, I_{p \times p}\right)$ then also by known results on Gaussian maxima (Boucheron et al. (2013, Theorem 2.5)) we have $\sup\left\langle\widetilde{\Gamma}^\dagger, h\right\rangle \leq 3\rho\sqrt{\log(m+p)}$ 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

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

Thus with high probability $\|\left(\Sigma^{1/2}\widetilde{\Gamma}^{\dagger}\right)^{\top}h\|_{\infty} \le 3\rho\sqrt{c_u \log(m+p)}$, then using the fact that $m \le \frac{p(p-1)}{2}$, $\|\left(\Sigma^{1/2}\widetilde{\Gamma}^{\dagger}\right)^{\top}h\|_{\infty} \le 6\rho\sqrt{c_u \log p}$ holds with probability at least $1 - \frac{c}{p}$. This completes the proof of Lemma 10.

Appendix O. Simulation Details

The graphs and corresponding covariance structures are constructed as follows:

Block Complete Graph

Σ is block diagonal with K blocks, each of size $\frac{p}{K} \times \frac{p}{K}$. Following the discussion in Section 2.2.1, all the diagonal elements are set to $\frac{K}{p}$ and all the off-diagonal elements in each block are set to $\frac{Kr}{p}$ with $r \in (0, 1)$. Here r is the correlation coefficient and will be set to different values in the experiments. Specifically, let

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

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

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

Chain Graph

Following the discussion in Section 2.2.2, we set elements in the main diagonal of Σ to be one, the first off-diagonal elements to be r with $r \in \left(0, \frac{1}{2}\right)$, 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 β^* is set to have $\beta_j^* \sim \mathcal{N}\left(1, 0.01^2\right)$ for $1 \le j \le s$ 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

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

Lattice Graph

Following the discussion in Section 2.2.3, we construct Σ as follows.

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

The corresponding true coefficient vector β^* with s active elements is set to $\beta_j^* \sim \mathcal{N}\left(1, 0.01^2\right)$ if $j \le \sqrt{s}$ mod $\sqrt{p}$ and $j \le \sqrt{ps}$ 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.

Appendix P. Biochemistry Table.

Table 2:

Description of structural features used in the biochemistry data analysis.

Structure Feature	Description
buried_np_AFILMVWY_per_res	Buried nonpolar surface area on nonpolar amino acids/count buried and core residues.
buried_np_per_res	Buried nonpolar surface area of the protein divided by count buried non polar residues.
buried_over_exposed	buried_np_per_res divided by solvent available surface area (sasa) of hydrophobic residues.
buried_over_exposed_AFILMVWY	buried_np_AFILMVWY_per_res divided by sasa of hydrophobic residues.
cbeta	A solvation term intended to correct for the excluded volume effect introduced by the simulation and favor compact structures. It is based on the ratio of probabilities of a residue having a given number of neighbors in a compact structure vs. random coil and summed over all residues.
cenpack	A centroid energy term.
contact_all_per_res	Count sidechain carbon-carbon contacts among all residues under the given distance cutoff divided by count residues of the sequence modeled.
contact_buried_core_boundary_per_res	Count sidechain carbon-carbon contacts among the buried and boundary residues under the given distance cutoff divided by count buried and boundary residues.
contact_buried_core_per_res	Count sidechain carbon-carbon contacts among the buried residues under the given distance cutoff divided by count buried residues.
degree_core_boundary_per_res	Count number of residues within a set distance of buried and boundary residues divided by count buried and boundary residues.
degree_core_per_res	Count number of residues within a set distance of buried residues divided by count buried residues.
degree_per_res	Count number of residues within a set distance of other residues divided by count residues of the sequence modeled.
env	A context-dependent one-body energy term that describes the solvation of a particular residue (based on the hydrophobic effect). It is based on the probability of a residue having the specified type given its number of neighboring residues.
exposed_hydrophobics_per_res	Sasa of hydrophobic residues divided by count residues of the sequence modeled.

Structure Feature	Description
exposed_polars_per_res	Sasa of polar residues divided by count residues of the sequence modeled.
exposed_total_per_res	Sasa of whole protein divided by count residues of the sequence modeled.
fa_atr	Lennard-Jones attractive.
fa_dun	Internal energy of sidechain rotamers as derived from Dunbrack's statistics.
fa_elec	Coulombic electrostatic potential with a distance-dependent dielectric. Supports canonical and noncanonical residue types.
fa_intra_rep	Lennard-Jones repulsive between atoms in the same residue.
fa_intra_sol_xover4	Intra-residue LK solvation, counted for the atom-pairs beyond torsion-relationship. Supports arbitrary residues types.
fa_rep	Lennard-Jones repulsive.
fa_sol	Lazaridis-Karplus solvation energy.
hbond_bb_sc	Sidechain-backbone hydrogen bond energy.
hbond_lr_bb	Backbone-backbone hbonds distant in primary sequence.
hbond_sc	Sidechain-sidechain and sidechain-backbone hydrogen bond energy.
hbond_sr_bb	Backbone-backbone hbonds close in primary sequence.
hs_pair	Describes packing between strands and helices. It is based on the probability that two pairs of residues (1 pair in the sheet and 1 pair in the helix) will have their current dihedral angles given the separation (in sequence and physical distance) between the helix and the strand.
lk_ball_wtd	Weighted sum of lk_ball & lk_ball_iso (w1*lk_ball + w2*lk_ball_iso); w2 is negative so that anisotropic contribution(lk_ball) replaces some portion of isotropic contribution (fa_sol=lk_ball_iso). Supports arbitrary residue types.
n_charged	Number of charged residues.
netcharge	The total charge.
omega	Omega angles.
one_core_each	The fraction of secondary structure elements (helices and strands) with one large hydrophobic residue (FILMVYW) at a position in the core layer of the protein.
p_aa_pp	Probability of observing an amino acid, given its phi/psi energy method declaration.
pack	Packing statistics. Calculated on whole protein.
pair	A two-body energy term for residue pair interactions (electrostatics and disulfide bonds). For each pair of residues, it is based on the probability that both of these two residues will have their specified types given their sequence separation and the physical distance between them, normalized by the product of the probabilities that each residue will have its specified type given the same information.
polar_over_hydrophobic	exposed_polars_per_res divided by exposed_hydrophobics_per_res.
pro_close	Proline ring closure energy.
rama_prepro	Backbone torsion preference term that takes into account of whether preceding amono acid is Proline or not. Currently supports the 20 canonical alpha-amino acids, their mirror-image D-amino acids, oligoureas, and N-methyl amino acids. Arbitrary new building-blocks can also be supported provided that an N-dimensional mainchain potential can be generated somehow.
ref	Reference energy for each amino acid.
rg	Favors compact structures and is calculated as the root mean square distance between residue centroids.
rsigma	Scores strand pairs based on the distance between them and the register of the two strands.

Structure Feature	Description
sheet	Favors the arrangement of individual beta strands into sheets. It is derived from the probability that a structure with a given number of beta strands will have the current number of beta sheets and lone beta strands.
ss_contributes_core	The fraction of secondary structure elements (helices and strands) with one large hydrophobic residue (FILMVYW) at a position in either the core or interface layer of the protein.
ss_mis	Generate secondary structure predictions from sequence. Calculated on whole protein.
ss_pair	Describes hydrogen bonding between beta strands.
total_score_per_res	The sum of all features, averaged by residue number.
two_core_each	The fraction of secondary structure elements (helices and strands) with two large hydrophobic residues (FILMVYW) at positions in the core layer of the protein.
vdw	Represents only steric repulsion and not attractive van der Waals forces (those are modeled in terms rewarding compact structures, such as the rg term; local interactions are implicitly included from fragments). It is calculated over pairs of atoms only in cases where: 1. the interatomic distance is less than the sum of the atoms' van der Waals radii, and 2. the interatomic distance does not depend on the torsion angles of a single residue.
yhh_planarity	Helps control the alcohol hydrogen in tyrosine.

References

Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL Jr, Das R, Baker D, Kuhlman B, Kortemme T, and Gray JJ (2017, 6). The rosetta All-Atom energy function for macromolecular modeling and design. >J. Chem. Theory Comput 13(6), 3031–3048. [PubMed: 28430426]

Anderson TW (1955). The integral of a symmetric convex set and some probability inequalities. Proc. of American Mathematical Society 6, 170–176.

Baik J and Silverstein JW (2006, 7). Eigenvalues of large sample covariance matrices of spiked populations models. Journal of Multivariate Analysis 97(6), 1382–1408.

Barnston AG and Smith TM (1996). Specification and prediction of global surface temperature and precipitation from global sst using cca. Journal of Climate 9(11), 2660–2697.

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

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

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

Bogdan M, van den Berg E, Su W, and Candes E (2013). Statistical estimation and testing via the ordered ℓ_1 norm. Technical Report arXiv:1310.1969.

Bondell HD and Reich BJ (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64(1), 115–123. [PubMed: 17608783]

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

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

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

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

Candes E and Tao T (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35(6), 2313–2351.

Caoa P, Liua X, Liuc H, Yanga J, Zhaoa D, Huang M, and Zaiane O (2018). Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in alzheimers disease. Computer Methods and Programs in Biomedicine 162, 19–45. [PubMed: 29903486]

Davidson KR and Szarek SJ (2001). Local operator theory, random matrices, and Banach spaces In Handbook of Banach Spaces, Volume 1, pp. 317–336. Amsterdan, NL: Elsevier.

Daye Z and Jeng X (2009). Shrinkage and model selection with correlated variables via weighted fusion. Computational Statistics & Data Analysis 53(4), 1284–1298.

DelSole T and Banerjee A (2017). Statistical seasonal prediction based on regularized regression. Journal of Climate 30(4), 1345–1361.

Donoho DL, Gavish M, and Johnstone IM (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. arXiv preprint arXiv:1311.0851.

El Anbari M and Mkhadri A (2014). Penalized regression combining the ℓ_1 norm and a correlation based penalty. Sankhya 76(1), 82102.

Figueiredo M and Nowak R (2016). Ordered weighted $\ell1$ regularized regression with strongly correlated covariates: Theoretical aspects. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pp. 930–938.

Geisler JE, Blackmon ML, Bates GT, and Munoz S (1985). Sensitivity of january climate response to the magnitude and position of equatorial pacific sea surface temperature anomalies. Journal of the atmospheric sciences 42(10), 1037–1049.

Guengerich FP (2002, 5). Cytochrome P450 enzymes in the generation of commercial products. Nat. Rev. Drug Discov. 1(5), 359–366. [PubMed: 12120411]

Hallac D, Leskovec J, and Boyd S (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, Tibshirani R, and Wainwright M (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.

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

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

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

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

Kalousis A, Prados J, and Hilario M (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and information systems 12(1), 95–116.

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

Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, and Arnold FH (2007, 9). A diverse family of thermostable cytochrome p450s created by recombination of stabilizing fragments. Nat. Biotechnol 25(9), 1051–1056. [PubMed: 17721510]

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

Mamalakis A, Yu J-Y, Randerson JT, AghaKouchak A, and Foufoula-Georgiou E (2018). A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern us. Nature communications 9(1), 2332.

Marial J and Yu B (2013). Supervised feature selection in graphs with path coding penalties and network flows. Journal of Machine Learning Research 14, 2449–2485.

Needell D and Ward R (2013a). Near-optimal compressed sensing guarantees for total variation minimization. IEEE Transactions on Image Processing 22(10), 3941–3949. [PubMed: 23708808]

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

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

Niehaus F, Bertoldo C, Kähler M, and Antranikian G (1999, 6). Extremophiles as a source of novel enzymes for industrial application. Appl. Microbiol. Biotechnol 51(6), 711–729. [PubMed: 10422220]

Noschese S, Pasquini L, and Reichel L (2013). Tridiagonal toeplitz matrices: properties and novel applications. Numerical linear algebra with applications 20(2), 302–326.

Zhao P, G. R. and Yu B (2009). The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics 37(6A), 3468–3497.

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

Raskutti G, Wainwright MJ, and Yu B (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Transactions on Information Theory 57(10), 6976–6994.

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

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

Sharma DB, Bondell HD, and Zhang HH (2013). Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics 22(2), 319–340. [PubMed: 23772171]

Sharpnack J, Singh A, and Rinaldo A (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 Huang H-C (2010). Grouping pursuit through a regularization solution surface. Journal of the American Statistical Association 105(490), 727–739. [PubMed: 20689721]

Shuman DI, Narang SK, Frossard P, Ortega A, and Vandergheynst P (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.

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, Saunders M, Rosset S, Zhu J, and Knight K (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 Taylor J (2011, 06). The solution path of the generalized lasso. The Annals of Statistics 39(3), 1335–1371.

Tutz G and Ulbricht J (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 Buhlmann P (2009). On the conditions used to prove oracle results for the lasso. Electronic Journal of Statistics 3, 1360–1392.

Vershynin R (2018, 9). High-Dimensional Probability by Roman Vershynin. Cambridge University Press.

Viallon V and Lambert-Lacroix S, Hoefling H, and Picard F (2016). On the robustness of the generalized fused lasso to prior specifications. Statistics and Computing 26(1), 285–301.

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

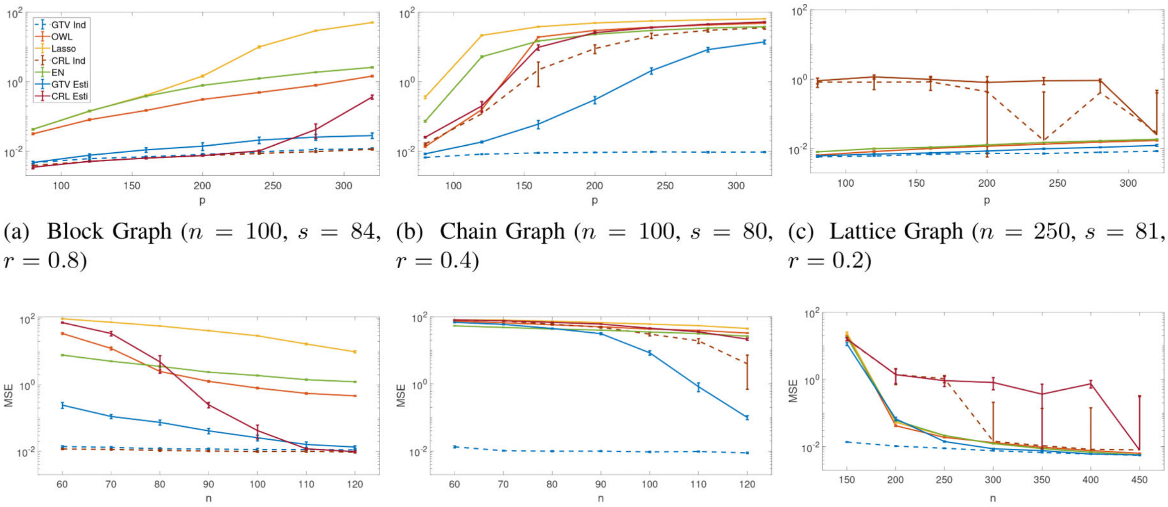
Wauthier FL, Jojic N, and Jordan MI (2013). A comparative framework for preconditioned lasso algorithms In Advances in Neural Information Processing Systems 26, pp. 1061–1069. Curran Associates, Inc.

Witten D, Shojaie A, and Zhang F (2014, 2). The cluster elastic net for high-dimensional regression with unknown variable grouping. Technometrics 56(1), 112–122. [PubMed: 24817772]

Wu TT, Chen YF, Hastie T, Sobel E, and Lange K (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6), 714–721. [PubMed: 19176549]

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

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



(a) Block Graph ($n = 100$, $s = 84$, $r = 0.8$)

(b) Chain Graph ($n = 100$, $s = 80$, $r = 0.4$)

(c) Lattice Graph ($n = 250$, $s = 81$, $r = 0.2$)

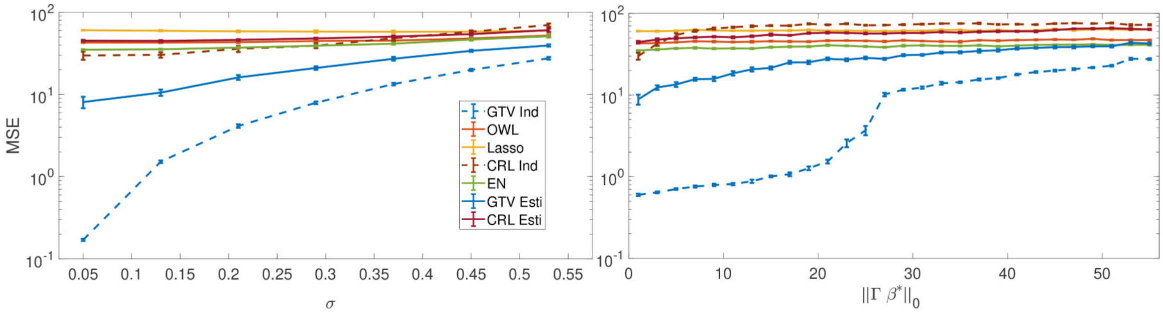
(d) Block Graph ($p = 280$, $s = 84$, $r = 0.8$)

(e) Chain Graph ($p = 280$, s = 80, $r = 0.4$)

(f) Lattice Graph ($p = 289$, s = 81, $r = 0.2$)

Figure 1.

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.



(a) Robustness to increases in σ. An increase in σ causes an increase in $\|\Gamma\beta^*\|_1$ while holding $\|\Gamma\beta^*\|_0$ constant.

(b) Robustness to increases in $\|\Gamma\beta^*\|_0$.

Figure 2.

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* β**. Medians 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.*

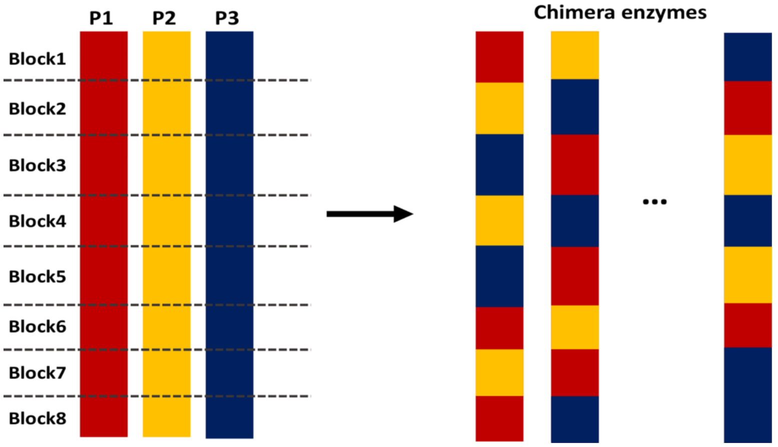
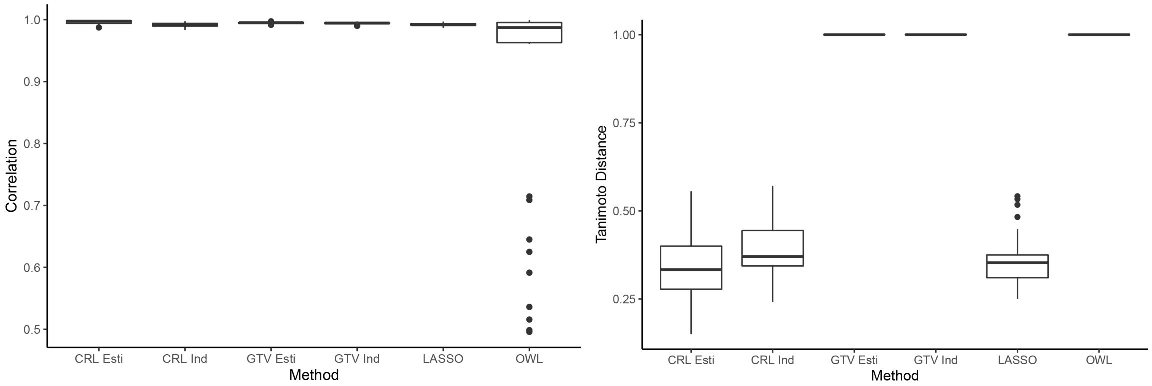


Figure 3.

A diagram of the process of creating Chimeras enzymes. P1, P2, and P3 are three parent proteins. They are each made up of an amino acid sequence (represented by red, yellow, or blue). There are 8 *pieces/blocks in each sequence. Chimera enzymes are made from recombining blocks from the* 3 *parents. The P450 dataset we use consist of Chimeras.*



(a) Correlation stability

(b) Tanimoto distance

Figure 4.

Box Plots of the two stability measures of each model on the P450 dataset. Correlation and Tanimoto distance were calculated between 10 different fittings for each model, leading to 45 measurements per model for each kind of the stability measrue.

Table 1

The average prediction error for each model on the P450 dataset.

	GTV Ind	GTV Esti	LASSO	CRL Ind	CRL Esti	OWL
Prediction Error	5.10	5.08	5.11	13.78	31.21	5.35