SIAM J. MATH. DATA SCI.
Vol. 2, No. 2, pp. 480–504

Graph-Based Regularization for Regression Problems with Alignment and Highly Correlated Designs*

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

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.

Key words. linear models, graphs, high-dimensional statistics

AMS subject classifications. 62-09, 62J99, 62P10

DOI. 10.1137/19M1287365

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

*Received by the editors September 17, 2019; accepted for publication (in revised form) March 2, 2020; published electronically June 16, 2020.

https://doi.org/10.1137/19M1287365

Funding: The first and third authors were supported by NSF DMS 1407028. The second author was supported by NSF Awards 0353079 and 1447449. The fourth author was supported by NIH Award 1U54 AI117924-01 and NSF Awards 0353079, 1447449, 1740707, and 1839338. The fifth author was supported by NIH R01 GM131381-01.

[†]These authors contributed equally to the paper.

[‡]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706 (yuanli@stat.wisc.edu, raskutti@stat.wisc.edu, hb.song@wisc.edu, dneiman5@gmail.com).

[§]Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (bmark2@wisc.edu).

[¶]Departments of Statistics and Computer Science, University of Chicago, Chicago, IL 60637 (willett@uchicago.edu).

and other related conditions (see Candes and Tao (2007); Bickel et al. (2009); van de Geer and Buhlmann (2009)). In this paper we develop an approach for parameter estimation in high-dimensional linear regression with highly correlated designs.

More specifically, we consider observations of the form

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

where $y \in \mathbb{R}^n$ is the response variable, $X \in \mathbb{R}^{n\times p}$ is the observation or *design* matrix, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ 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 $\Sigma := \frac{1}{n}\mathbb{E}[X^\top X]$ 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 in the supplementary materials 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 a large 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, β^* 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 et al. (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 3D 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.[1] 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 whose 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 3D 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 P.1 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 our 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 $(X^{(i)}, y^{(i)})_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$ are drawn i.i.d. according to

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

Let $y = (y^{(1)}, y^{(2)}, \dots, y^{(n)})^\top \in \mathbb{R}^n$, $X = [X^{(1)}, X^{(2)}, \dots, X^{(n)}]^\top \in \mathbb{R}^{n\times p}$, and $\epsilon = (\epsilon^{(1)}, \epsilon^{(2)}, \dots, \epsilon^{(n)})^\top \in \mathbb{R}^n$. 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 $\hat{s}_{j,k} := \text{sign}(\hat{\Sigma}_{j,k})$. Based on the estimated

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

covariance matrix $\hat{\Sigma}$, we consider the following estimator for β^*:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda_S \sum_{j,k} |\hat{\Sigma}_{j,k}|(\beta_j - \hat{s}_{j,k}\beta_k)^2 + \lambda_1 \left(\lambda_{\mathrm{TV}} \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k}\beta_k| + \|\beta\|_1 \right), \tag{1.2}$$

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 $w_{j,k}$ $(1 \le j \ne k \le p)$ 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 $E := \{(j, k) \ : \ w_{j,k} \ne 0, \ j \ne k\}$ and the entries of the weight matrix $W \in \mathbb{R}^{p\times p}$ are $W_{j,k} = w_{j,k}$. Given this graph, the regularization term $\sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2}|\beta_j - \hat{s}_{j,k}\beta_k|$ is a measure of the *graph total variation* (GTV) of the signal β with respect to the graph G, and $\sum_{j,k} |\hat{\Sigma}_{j,k}|(\beta_j - \hat{s}_{j,k}\beta_k)^2$ 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

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

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

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

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

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

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

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

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

The three regularizers play the following roles:

- We refer to $\|\Gamma\beta\|_2^2 = \sum_{j,k} |\hat{\Sigma}_{j,k}|(\beta_j - \hat{s}_{j,k}\beta_k)^2$ as the *Laplacian smoothing penalty*; Hebiri and van de Geer (2011) studied a variant of this regularizer with $\hat{\Sigma}_{j,k}$ replaced with arbitrary nonnegative 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 $\|\Gamma\beta\|_1 = \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2}|\beta_j - \hat{s}_{j,k}\beta_k|$ 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 $\hat{\Sigma}$, as we do). This regularizer promotes estimates $\hat{\beta}$ that are *well aligned* with the graph structure; for instance, a group of nodes with large edge weights connecting them (i.e., a group of columns of X that are highly correlated) is 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 amounts 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 with 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 (MSE) (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 nonzero eigenvalue.
- New MSE 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 outperforms 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 MSE in a variety of settings.
- A validation of our method on real biochemistry data that demonstrates the advantages 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 MSE; 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 (GTV) estimation (Shuman et al., 2013). Wang et al. (2016) consider a generalization of GTV 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) considers 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 Alzheimer's 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

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

In contrast to the fused LASSO, the clustered LASSO considers *all* pairwise differences of elements of 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) considers 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

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

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 nonoverlapping; (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 nonconvex 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,

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

where $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $|\beta|_{[j]}$ is the jth 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 precomputed and will depend on the value of y.

An estimator called the *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

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

The first regularization term, unlike the total variation term in (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 semidefinite 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 exist 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*

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

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*

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

where c_ℓ is as 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 $(X^{(i)})_{i=1}^n$. 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 $(\check{X}^{(i)})_{i=1}^m$ that are potentially independent of the design features $(X^{(i)})_{i=1}^n$ with $\check{X}^{(i)} \sim \mathcal{N}(\mathbf{0}, \Sigma_{p\times p})$. In this case, we need to estimate Σ based on $(\check{X}^{(i)})_{i=1}^m$, and there is a large literature on high-dimensional covariance estimation in high dimensions under different structural assumptions (see Bickel and Levina (2008a,b); 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 (2008b). 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 $\tilde{\Gamma}$ appearing in our regularizer.

Definition 2.1 (Compatibility factor k_T, Hütter and Rigollet (2016)). *We define the compatibility factor k_T of matrix $\tilde{\Gamma}$ for a set $T \subset \{1, 2, \ldots, p, p+1, \ldots, p+m\}$ as*

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

This compatibility factor k_T reflects the degree of compatibility of the ℓ_1-regularizer $\|(\tilde{\Gamma}\beta)_T\|_1$ 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 $S := \tilde{\Gamma}^\dagger = [s_1, \dots, s_{m+p}]$, where $\tilde{\Gamma}^\dagger$ is the Moore–Penrose pseudoinverse of the matrix $\tilde{\Gamma}$, and define the inverse scaling factor as*

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

Remark 2.4. Definitions 2.1 and 2.2 were first proposed in Hütter and Rigollet (2016), though the definition of ρ is based on $\tilde{\Gamma}$ 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 flavor to the condition number of the matrix $\tilde{\Gamma}$.

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

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

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

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 $\hat{\Sigma}$ is constructed independently from the sample $(X^{(i)})_{i=1}^n$ fits the settings we are motivated by where it is possible to exploit side information to estimate $\hat{\Sigma}$. If instead the covariance matrix is estimated using $(X^{(i)})_{i=1}^n$, 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 $\hat{\Sigma}$ and $(X^{(i)})_{i=1}^n$ 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 $D_{jj} = \sum_{k=1}^p |\hat{\Sigma}_{j,k}|$, $1 \leq j \leq p$, then

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

Hence if Σ and $\hat{\Sigma}$ 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 Assumptions 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 1, 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, we frequently choose $T = \mathrm{Supp}(\tilde{\Gamma}\beta^*)$.

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 $\lambda_1 \ge 8\lambda_S\|L\beta^*\|_\infty$. 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 MSE 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 MSE 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 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 preconditioning is still achieved. However, the important difference and why we prefer the graph Laplacian penalty is that using our analysis the condition $\lambda_1 \ge 8\lambda_S\|L\beta^*\|_\infty$ would be replaced by $\lambda_1 \ge 8\lambda_S\|\beta^*\|_\infty$. Hence if we were in the strictly sparse case and $\lambda_{\mathrm{TV}} = 0$, we would recover the MSE bound:

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

Note that this exactly matches the MSE 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 cannot 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*

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

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 $T = \text{Supp}(\tilde{\Gamma}\beta)$ for some β; then $T_1 = \text{Supp}(\Gamma\beta)$ and $T_2 = \text{Supp}(\beta)$ and Lemma 2 can be reduced to

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

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

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

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

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 kth connected component.

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

Theorem 2. *Suppose $\lambda_1 > 0$ and Assumptions 2.1–2.3 are satisfied, and suppose the estimated covariance matrix $\hat{\Sigma}$ is constructed independently from the sample $(X^{(i)})_{i=1}^n$. If we choose*

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

then there exist absolute positive constants C_1 and C such that

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

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

The proof of Theorem 2 is provided in section B in the Appendix. 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 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 $\lambda_1^2\lambda_{TV}^2\|\hat{\Sigma}\|_{1,1}\|\Gamma\beta^*\|_0$ or $\lambda_1\lambda_{TV}\|\Gamma\beta^*\|_1$ and the lower bound on $\lambda_1\lambda_{TV}$ increases as 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 $||X\hat{\beta} - X\beta^*||_2^2$. 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 $\lambda_S = \lambda_{\text{TV}} = 0$. From Theorem 2 we can see that

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

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 $\lambda_{\text{TV}} = C \max_{1 \leq k \leq K} \sqrt{\frac{|B_k|}{\mu_k}}$; then we obtain the bound

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

where $\lambda_1^2 = O(\max_{1 \leq k \leq K} \frac{\sigma^2 c_u \log p}{n|B_k|})$ and $\lambda_1^2 \lambda_{\text{TV}}^2 = O(\max_{1 \leq k \leq K} \frac{|B_k|}{\mu_k} \max_{1 \leq k \leq K} \frac{\sigma^2 c_u \log p}{n|B_k|})$. 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^*; CRL and OWL 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 preconditioner 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 $\tilde{X}$, so even if the restricted eigenvalue condition is not satisfied for X, it is satisfied for $\tilde{X}$. What distinguishes our results from previous work using preconditioners for LASSO (Jia and Rohe, 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 that are 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 nonzero 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 MSE. 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

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

we refer to r as the correlation coefficient. Note that here a is a normalization parameter that we set to ensure 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 $\hat{\Sigma} = \Sigma$, 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*

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

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 MSE bound for the block complete graph.

Corollary 2.3. *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* $1 - \frac{C_1}{p}$

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

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

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 $K_1 \leq K$ 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

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

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 corresponding to the chain graph satisfies $\Sigma_{jj} = 1$ for all j and $\Sigma_{jk} = r$ for all $(j,k) \in E$ where $E = \{(1,2),(2,3),\ldots,(p-1,p)\}$. Assumptions 2.1 and 2.2 are clearly satisfied, and requiring $r \in (0, \frac{1}{2})$ ensures Σ is positive semidefinite. 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 $\hat{\Sigma} = \Sigma$. Then*

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

Using Lemma 5, we have the following corollary for the chain graph.

Corollary 2.4. *For a chain graph with details described above, suppose that $\hat{\Sigma} = \Sigma$. If we choose*

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

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

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

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 2.4 says

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

which is stronger than the LASSO guarantee of

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

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 $r \in (0, \frac{1}{4})$ so that Σ is positive semidefinite. Clearly Assumptions 2.1 and 2.2 are satisfied for any $r \in (0, \frac{1}{4})$, and we note that the lattice graph is fully connected, so $K = 1$ and $B_1 = \{1, 2, \dots, p\}$. 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 $\hat{\Sigma} = \Sigma$. Then*

$$\rho \leq \sqrt{\frac{1}{p} + \frac{5\pi \log(2 + r\lambda_{TV})}{r^2\lambda_{TV}^2 + 1} + \frac{10\pi}{r\lambda_{TV}\sqrt{p} + 1}},$$
$$\lambda_{\min}(\Sigma + \lambda_S L) \geq (1 - \lambda_S)(1 - 4r) + \lambda_S.$$

Using Lemma 6 we have the following corollary for the lattice graph.

Corollary 2.5. *For a lattice graph with details described above, suppose that $\hat{\Sigma} = \Sigma$. 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

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

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 2.5 says

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

which is stronger than the LASSO guarantee of

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

Note that the MSE_{GTV} bound from this example is identical to the MSE_{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 Appendix O. 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 $X_{\text{ind}} \in \mathbb{R}^{1000\times p}$ 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 MSE ($\text{MSE} = \|\hat{\beta} - \beta^*\|_2^2$). For all methods, tuning parameters are chosen based on fivefold 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, $\hat{\Sigma}$, using hard thresholding of the sample covariance matrix with a threshold chosen by cross-validation (see Bickel and Levina (2008b) for more details). We construct the edge incidence matrix Γ based on $\hat{\Sigma}$ and then estimate $\hat{\beta}$ using (1.5).

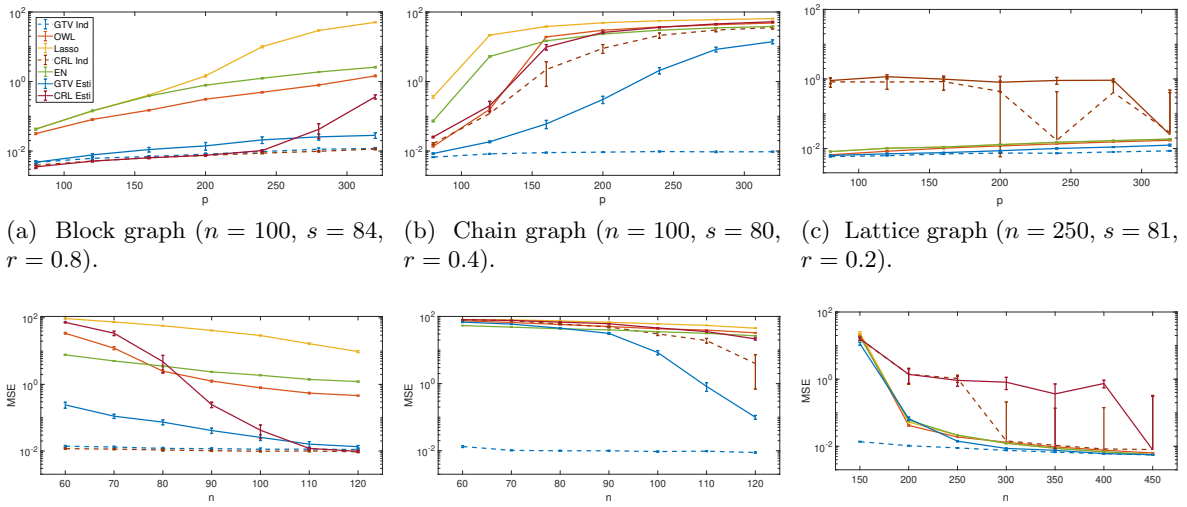
GTV-Indep (our method). This approach is equivalent to GTV-Esti (above), except that the side information X_{ind} is used to compute the estimated covariance matrix $\hat{\Sigma}$.

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.



(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 $= 81$, $r = 0.2$).

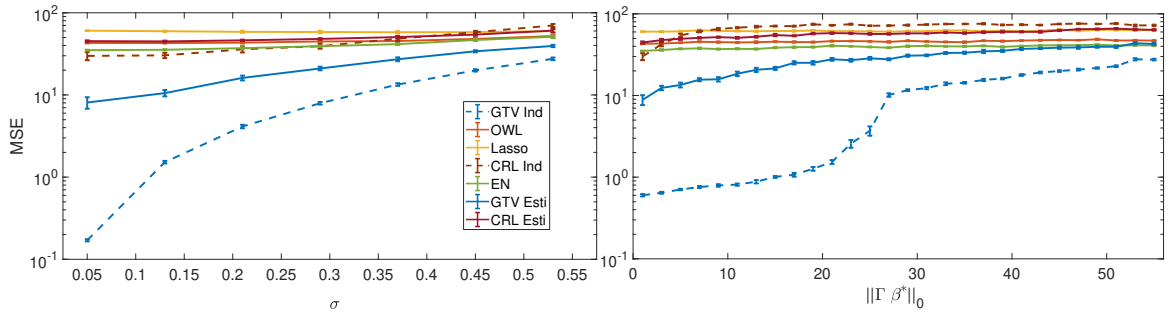
Figure 1. *MSE for varying covariance graph structures and values of n and p. Medians of* 100 *trials are shown, and error bars denote the standard deviation of the median estimated using the bootstrap method with* 500 *resamplings on the* 100 *MSEs. GTV-Esti yields lower MSEs than other methods for a broad range of n, p.*

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 MSE changes with the 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 $\beta_j^* \sim \mathcal{N}(1, \sigma^2)$ 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 $\mathcal{N}(1, .01^2)$. We measure MSE as 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



(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 blocks, 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 *MSEs.*

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

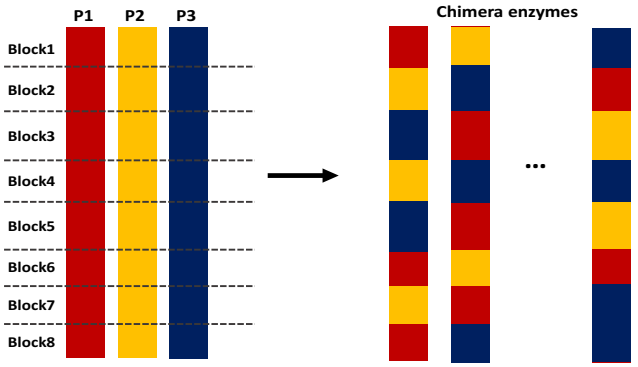


Figure 3. *A diagram of the process of creating Chimeras enzymes. P1, P2, and P3 are three parent proteins. They are each made up of an amino acid sequence (represented by red, yellow, or blue). There are 8 pieces/blocks in each sequence. Chimera enzymes are made from recombining blocks from the 3 parents.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

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 $\delta^{(i)} \sim \mathcal{N}(0, \sigma_\delta^2)$ 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

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

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

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

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), standard LASSO (LASSO), and the Elastic Net (EN) methods. For all models, the tuning parameters were selected via fivefold 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):

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

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 splitting the entire P450 dataset into ten nonoverlapping 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 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 a 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.

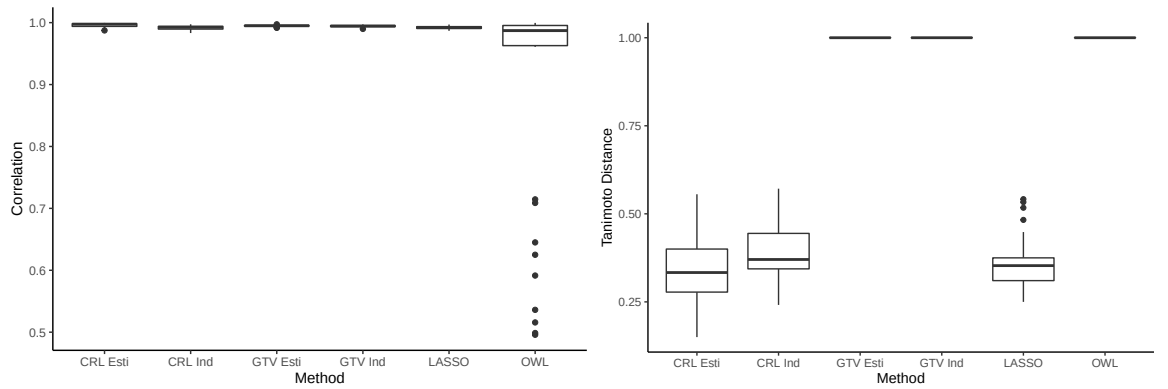
Table 1

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

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 the LASSO and CRL methods.



(a) Correlation stability. (b) Tanimoto distance.

Figure 4. *Box plots of the two stability measures of each model on the P*450 *dataset. Correlation and Tanimoto distance were calculated between* 10 *different fittings for each model, leading to* 45 *measurements per model for each kind of stability measure.*

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 Elastic Net (Zou and Hastie, 2005), Fused LASSO (Tibshirani et al., 2005), 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 precondition our design matrix and regularize regression weights to promote alignment with the covariance structure of the problem. We are able to provide MSE 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 clustering 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.

Acknowledgments. 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.

REFERENCES

R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray (2017), *The Rosetta All-Atom energy function for macromolecular modeling and design*, J. Chem. Theory Comput., 13, pp. 3031–3048, https://doi.org/10.1021/acs.jctc.7b00125.

T. W. Anderson (1955), *The integral of a symmetric convex set and some probability inequalities*, Proc. of Amer. Math. Soc., 6, pp. 170–176.

J. Baik and J. W. Silverstein (2006), *Eigenvalues of large sample covariance matrices of spiked populations models*, J. Multivariate Anal., 97, pp. 1382–1408.

A. G. Barnston and T. M. Smith (1996), *Specification and prediction of global surface temperature and precipitation from global SST using CCA*, J. Climate, 9, pp. 2660–2697.

P. Bickel and E. Levina (2008a), *Regularized estimation of large covariance matrices*, Ann. Statist., 36, pp. 199–227.

P. Bickel and E. Levina (2008b), *Covariance regularization by thresholding*, Ann. Statist., 36, pp. 2577–2604.

P. Bickel, Y. Ritov, and A. Tsybakov (2009), *Simultaneous analysis of lasso and Dantzig selector*, Ann. Statist., 37, pp. 1705–1732.

M. Bogdan, E. van den Berg, W. Su, and E. Candes (2013), *Statistical Estimation and Testing via the Ordered ℓ_1 Norm*, preprint, https://arxiv.org/abs/1310.1969v1.

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

S. Boucheron, G. Lugosi, and P. Massart (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, UK.

P. Bühlmann, P. Rütimann, S. van de Geer, and C. Zhang (2013), *Correlated variables in regression: Clustering and sparse estimation*, J. Statist. Planning Inference, 143, pp. 1835–1858.

T. T. Cai and W. Liu (2011), *Adaptive thresholding for sparse covariance matrix estimation*, J. Amer. Statist. Assoc., 106, pp. 672–684.

T. T. Cai, R. Zhao, and H. H. Zhou (2016), *Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation*, Electron. J. Statist., 10, pp. 1–59, https://doi.org/10.1214/15-EJS1081.

E. Candes and T. Tao (2007), *The Dantzig selector: Statistical estimation when p is much larger than n*, Ann. Statist., 35, pp. 2313–2351.

P. Caoa, X. Liua, H. Liuc, J. Yanga, D. Zhaoa, M. Huang, and O. Zaiane (2018), *Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimer's disease*, Comput. Methods Programs Biomed., 162, pp. 19–45.

K. R. Davidson and S. J. Szarek (2001), *Local operator theory, random matrices, and Banach spaces*, in Handbook of the Geometry of Banach Spaces, Vol. 1, North-Holland, Amsterdam, pp. 317–336.

Z. Daye and X. Jeng (2009), *Shrinkage and model selection with correlated variables via weighted fusion*, Comput. Statist. Data Anal., 53, pp. 1284–1298.

T. DelSole and A. Banerjee (2017), *Statistical seasonal prediction based on regularized regression*, J. Climate, 30, pp. 1345–1361.

D. L. Donoho, M. Gavish, and I. M. Johnstone (2013), *Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model*, preprint, https://arxiv.org/abs/1311.0851.

M. El Anbari and A. Mkhadri (2014), *Penalized regression combining the ℓ_1 norm and a correlation based penalty*, Sankhya, 76, pp. 82–102.

M. Figueiredo and R. Nowak (2016), *Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, pp. 930–938.

J. E. Geisler, M. L. Blackmon, G. T. Bates, and S. Munoz (1985), *Sensitivity of January climate response to the magnitude and position of equatorial Pacific sea surface temperature anomalies*, J. Atmospher. Sci., 42, pp. 1037–1049.

F. P. Guengerich (2002), *Cytochrome P450 enzymes in the generation of commercial products*, Nat. Rev. Drug Discov., 1, pp. 359–366, https://doi.org/10.1038/nrd792.

D. Hallac, J. Leskovec, and S. Boyd (2015), *Network lasso: Clustering and optimization in large graphs*, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, pp. 387–396.

T. Hastie, R. Tibshirani, and M. Wainwright (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Boca Raton, FL.

M. Hebiri and S. van de Geer (2011), *The smooth-lasso and other $\ell_1+\ell_2$-penalized methods*, Electron. J. Statist., 5, pp. 1184–1226.

J. Huang, S. Ma, H. Li, and C.-H. Zhang (2011), *The sparse Laplacian shrinkage estimator for high-dimensional regression*, Ann. Statist., 39, pp. 2021–2046, https://doi.org/10.1214/11-AOS897.

J. Hütter and P. Rigollet (2016), *Optimal Rates for Total Variation Denoising*, preprint, https://arxiv.org/abs/1603.09388.

J. Jia and K. Rohe (2015), *Preconditioning the lasso for sign consistency*, Electron. J. Statist., 9, pp. 1150–1172.

A. Kalousis, J. Prados, and M. Hilario (2007), *Stability of feature selection algorithms: A study on high-dimensional spaces*, Knowledge Inform. Syst., 12, pp. 95–116.

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

Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, and F. H. Arnold (2007), *A diverse family of thermostable cytochrome p450s created by recombination of stabilizing fragments*, Nat. Biotechnol., 25, pp. 1051–1056, https://doi.org/10.1038/nbt1333.

J. Liu, L. Yuan, and J. Ye (2013), *Dictionary LASSO: Guaranteed Sparse Recovery under Linear Transformation*, preprint, https://arxiv.org/abs/1305.0047.

A. Mamalakis, J.-Y. Yu, J. T. Randerson, A. A. Kouchak, and E. Foufoula-Georgiou (2018), *A new interhemispheric teleconnection increases predictability of winter precipitation in Southwestern US*, Nat. Commun., 9, 2332.

J. Marial and B. Yu (2013), *Supervised feature selection in graphs with path coding penalties and network flows*, J. Mach. Learn. Res., 14, pp. 2449–2485.

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

D. Needell and R. Ward (2013b), *Stable image reconstruction using total variation minimization*, SIAM J. Imaging Sci., 6, pp. 1035–1058, https://doi.org/10.1137/120868281.

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

F. Niehaus, C. Bertoldo, M. Kähler, and G. Antranikian (1999), *Extremophiles as a source of novel enzymes for industrial application*, Appl. Microbiol. Biotechnol., 51, pp. 711–729.

S. Noschese, L. Pasquini, and L. Reichel (2013), *Tridiagonal Toeplitz matrices: Properties and novel applications*, Numer. Linear Algebra Appl., 20, pp. 302–326.

G. Raskutti and M. Yuan (2015), *Convex Regularization for High-Dimensional Tensor Regression*, preprint, https://arxiv.org/abs/1512.01215v1.

G. Raskutti, M. J. Wainwright, and B. Yu (2010), *Restricted eigenvalue conditions for correlated Gaussian designs*, J. Mach. Learn. Res., 11, pp. 2241–2259.

G. Raskutti, M. J. Wainwright, and B. Yu (2011), *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Trans. Inform. Theory, 57, pp. 6976–6994.

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

D. B. Sharma, H. D. Bondell, and H. H. Zhang (2013), *Consistent group identification and variable selection in regression with correlated predictors*, J. Comput. Graphic. Statist., 22, pp. 319–340, https://doi.org/10.1080/15533174.2012.707849.

J. Sharpnack, A. Singh, and A. Rinaldo (2012), *Sparsistency of the edge lasso over graphs*, Art. Intell. Statist., pp. 1028–1036.
Y. She (2010), *Sparse regression with exact clustering*, Electron. J. Statist., 4, pp. 1055–1096.
X. Shen and H.-C. Huang (2010), *Grouping pursuit through a regularization solution surface*, J. Amer. Statist. Assoc., 105, pp. 727–739.
D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst (2013), *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*, IEEE Signal Process. Mag., 30, pp. 83–98.
G. Strang (2007), *Computational Science and Engineering*, Wellesley-Cambridge Press, Wellesley, MA.
R. Tibshirani (1996), *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58, pp. 267–288.
R. Tibshirani and J. Taylor (2011), *The solution path of the generalized lasso*, Ann. Statist., 39, pp. 1335–1371, https://doi.org/10.1214/11-AOS878.
R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005), *Sparsity and smoothness via the fused lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67, pp. 91–108.
G. Tutz and J. Ulbricht (2009), *Penalized regression with correlation-based penalty*, Stat. Comput., 19, pp. 239–253.
S. van de Geer (2000), *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, UK.
S. van de Geer and P. Buhlmann (2009), *On the conditions used to prove oracle results for the lasso*, Electron. J. Statist., 3, pp. 1360–1392.
R. Vershynin (2018), *High-Dimensional Probability*, Cambridge University Press, Cambridge, UK.
S. Viallon, V. Lambert-Lacroix, H. Hoefling, and F. Picard (2016), *On the robustness of the generalized fused lasso to prior specifications*, Stat. Comput., 26, pp. 285–301.
Y. Wang, J. Sharpnack, A. Smola, and R. Tibshirani (2016), *Trend filtering on graphs*, J. Mach. Learn. Res., 17, pp. 1–41.
F. L. Wauthier, N. Jojic, and M. I. Jordan (2013), *A comparative framework for preconditioned lasso algorithms*, in Advances in Neural Information Processing Systems, Vol. 26, NeurIPS, San Diego, CA, pp. 1061–1069, https://papers.nips.cc/paper/5104-a-comparative-framework-for-preconditioned-lasso-algorithms.pdf.
D. M. Witten, A. Shojaie, and F. Zhang (2014), *The cluster elastic net for high-dimensional regression with unknown variable grouping*, Technometrics, 56, pp. 112–122, https://doi.org/10.1080/00401706.2013.810174.
T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009), *Genome-wide association analysis by lasso penalized logistic regression*, Bioinformatics, 25, pp. 714–721.
P. Zhao, G. Rocha, and B. Yu (2009), *The composite absolute penalties family for grouped and hierarchical variable selection*, Ann. Statist., 37, pp. 3468–3497.
H. Zou (2006), *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc., 101, pp. 1418–1429.
H. Zou and T. Hastie (2005), *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67, pp. 301–320.