Elementary Estimators for High-Dimensional Linear Regression

Eunho Yang EUNHO@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Aurélie C. Lozano ACLOZANO@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Abstract

We consider the problem of structurally constrained high-dimensional linear regression. This has attracted considerable attention over the last decade, with state of the art statistical estimators based on solving regularized convex programs. While these typically non-smooth convex programs can be solved by the state of the art optimization methods in polynomial time, scaling them to very large-scale problems is an ongoing and rich area of research. In this paper, we attempt to address this scaling issue at the source, by asking whether one can build *simpler* possibly closed-form estimators, that yet come with statistical guarantees that are nonetheless comparable to regularized likelihood estimators. We answer this question in the affirmative, with variants of the classical ridge and OLS (ordinary least squares estimators) for linear regression. We analyze our estimators in the high-dimensional setting, and moreover provide empirical corroboration of its performance on simulated as well as real world microarray data.

1. Introduction

We consider the problem of high-dimensional linear regression, where the number of variables p could potentially be even larger than the number of observations n. Under such high-dimensional regimes, it is now well understood that consistent estimation is typically not possible unless one imposes low-dimensional structural constraints upon the regression parameter vector. Popular structural constraints include that of sparsity, where very few entries of the high-dimensional regression parameter are assumed to be non-zero, group-sparse constraints, and low-rank structure with matrix-structured parameters, among others.

The development of consistent estimators for such structurally constrained high-dimensional linear regression has attracted considerable recent attention. A key class of estimators are based on regularized maximum likelihood estimators; in the case of linear regression with Gaussian noise, these take the form of regularized least squares estimators. For the case of sparsity, a popular instance is constrained basis pursuit or LASSO (Tibshirani, 1996), which solves an ℓ_1 regularized (or equivalently ℓ_1-constrained) least squares problem, and has been shown to have strong statistical guarantees, including prediction error consistency (van de Geer & Buhlmann, 2009), consistency of the parameter estimates in ℓ_2 or some other norm (van de Geer & Buhlmann, 2009; Meinshausen & Yu, 2009; Candes & Tao, 2006), as well as variable selection consistency (Meinshausen & Bühlmann, 2006; Wainwright, 2009; Zhao & Yu, 2006). For the case of group-sparse structured linear regression, ℓ_1/ℓ_q regularized least squares (with $q \geq 2$) has been proposed (Tropp et al., 2006; Zhao et al., 2009; Yuan & Lin, 2006; Jacob et al., 2009), and shown to have strong statistical guarantees, including convergence rates in ℓ_2-norm (Lounici et al., 2009; Baraniuk et al., 2008)) as well as model selection consistency (Obozinski et al., 2008; Negahban & Wainwright, 2009). For the matrix-structured least squares problem, nuclear norm regularized estimators have been studied for instance in (Recht et al., 2010; Bach, 2008). For other structurally constrained least squares problems, see (Huang et al., 2011; Bach et al., 2012; Negahban et al., 2012) and references therein. All of these estimators solve convex programs, though with non-smooth components due to the respective regularization functions. The state of the art optimization methods for solving these programs are iterative, and can approach the optimal solution within any finite accuracy with computational complexity that scales polynomially with the number

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

of variables and number of samples, rendering these very expensive for very large-scale problems.

In another line of work, the Dantzig estimator (Candes & Tao, 2007) solves a linear program to estimate the sparse linear regression parameter with stronger statistical guarantees when compared to the LASSO (Bickel et al., 2009). But here again, the linear program while convex, is computationally expensive for very large-scale problems. Other class of estimators for structured linear regression include greedy methods. These include forward-backward selection (Zhang et al., 2008b), matching pursuit (Mallat & Zhang, 1993), and orthogonal matching pursuit (Zhang, 2010), among others. The caveat with such greedy methods however is that some of these require considerably more stringent conditions to hold when compared to regularized convex programs noted above, and either require the knowledge of the exact structural complexity such as sparsity level, or are otherwise unstable i.e. sensitive to tuning parameters such as their stopping thresholds. Thus overall, the class of regularized convex programs have proved more popular for large-scale structured linear regression, and accordingly there has been a strong line of recent research on large-scale optimization methods for such programs (see e.g. Friedman et al. (2007); Hsieh et al. (2011) and references therein), including parallel and distributed variants.

In this paper, we consider the following simple question:

"If we restrict ourselves to estimators with *closed-form* solutions; can we nonetheless obtain consistent estimators that have the sharp convergence rates of the regularized convex programs and other estimators noted above?"

A natural closed-form estimator for the high-dimensional linear regression problem is the ordinary least squares (OLS) estimator. This is the maximum likelihood solution under a Gaussian noise assumption, and is thus consistent with optimal convergence rates under *classical* statistical settings where the number of variables p is fixed as a function of the number of samples n. However under high-dimensional regimes, it is not only inconsistent but the least squares estimation problem does not even have a unique minimum. Another classical estimator is the ridge, or ℓ_2-regularized least squares estimator, which is also available in closed form; but unlike the OLS estimator, the ridge-regularized estimation problem has a unique solution. While simple, it is however not known to be for instance ℓ_2-norm consistent under high-dimensional settings. Another prominent closed-form estimator and an OLS variant is marginal regression: this regresses each covariate separately on the response to get the corresponding regression parameter. Correlation or sure screening (Fan & Lv, 2008) (also termed simple thresholding (Donoho, 2006)) is a closely related method for the high-dimensional sparse linear regression setting, where the regression parameters are set to soft-thresholded values of the correlation of the covariates with the response. However, as (Genovese et al., 2012) showed, as a flip side of the simplicity of marginal regression, this method requires very stringent conditions (loosely, extremely weak correlations between the covariates) in order to select a relevant subset of covariates.

In this paper, we are surprisingly able to answer our question in the affirmative, by building on the OLS and ridge estimators: we provide close variants that are not only available in closed form, but also come with strong statistical guarantees similar to those of the regularized convex program based estimators. The estimators are reminiscent of the Dantzig estimator, but in contrast are actually available in closed form, are thus much more scalable. As we show, for the sparse structure case the ridge-variant comes with slightly worse guarantees when compared to the LASSO, but the OLS variant actually has comparable guarantees to that of the LASSO. We also provide a unified statistical analysis for general structures *beyond sparsity*. We corroborate these results via simulations as well as on a real-world microarray dataset. Overall, the estimators and analyses in the paper thus motivate a new line of research on simpler and possibly closed form estimators for very high-dimensional statistical estimation.

2. Setup

We consider the linear regression model,

$$y_i = x_i^\top \theta^* + w_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\theta^* \in \mathbb{R}^p$ is the fixed unknown regression parameter of interest, $y_i \in \mathbb{R}$ is a real-valued response, $x_i \in \mathbb{R}^p$ is a known observation vector, and $w_i \in \mathbb{R}$ is an unknown noise term. For technical simplicity, we assume that these noise terms are independent zero-mean Gaussian random variables, $w_i \in \mathcal{N}(0, \sigma^2)$, for some $\sigma > 0$. Suppose we collate the n observations from the linear regression model in (1) in vector and matrix form. Let $y \in \mathbb{R}^n$ denote the vector of n responses, $X \in \mathbb{R}^{n \times p}$ denote the design matrix consisting of the linear regression observation vectors, and $w \in \mathbb{R}^n$ the vector of n noise terms. The linear regression model thus entails: $y = X\theta^* + w$.

We are interested in the high-dimensional setting, where the number of variables p may be of the same order as, or even substantially larger than the sample size n, so that $p \gg n$. Under such high-dimensional settings, it is now well understood that it is typically necessary to impose structural constraints on the model parameters θ^*.

2.1. Unified Framework of Negahban et al. (2012)

We follow the unified statistical framework of Negahban et al. (2012) to formalize the notion of *structural con-*

straints. There, they use subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$ where $\mathcal{M}$ is the *model subspace* in which the model parameter θ^* and similarly structured parameters lie, and which is typically *low-dimensional*, while $\overline{\mathcal{M}}^{\perp}$ is the *perturbation subspace* of parameters that represents perturbations away from the model subspace.

They also define the property of *decomposability* of a regularization function, which captures the suitablity of a regularization function $\mathcal{R}$ to particular structure. Specifically, a regularization function $\mathcal{R}$ is said to be *decomposable* with respect to a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, if $\mathcal{R}(u+v) = \mathcal{R}(u) + \mathcal{R}(v)$, for all $u \in \mathcal{M},\ v \in \overline{\mathcal{M}}^{\perp}$. Note that when $\mathcal{R}(\cdot)$ is a norm, by the triangle inequality, the LHS is always less than or equal to the RHS, so that the equality indicates the largest possible value for the LHS. In other words, the decomposable regularization function $\mathcal{R}(\cdot)$ heavily penalizes perturbations v from structured parameters u.

For any structure such as sparsity, low-rank, etc., we can define the corresponding low-dimensional model subspaces, as well as regularization functions that are decomposable with respect to the corresponding subspace pairs.

Example 1. *Given any subset $S \subseteq \{1, \ldots, p\}$ of the coordinates, let $\mathcal{M}(S)$ be the* subspace *of vectors in $\mathbb{R}^p$ that have support contained in S. It can be seen that any parameter $\theta \in \mathcal{M}(S)$ would be atmost $|S|$-sparse. For this case, we use $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$, so that $\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S)$. Negahban et al. (2012) show that the ℓ_1 norm $\mathcal{R}(\theta) = \|\theta\|_1$, commonly used as a sparsity-encouraging regularization function, is decomposable with respect to subspace pairs $(\mathcal{M}(S), \overline{\mathcal{M}}^{\perp}(S))$.*

Also of note is the *subspace compatibility constant* that captures the relationship between the regularization function $\mathcal{R}(\cdot)$ and the error norm $\|\cdot\|$, over vectors in subspace $\mathcal{M}$: $\Psi(\mathcal{M}, \|\cdot\|) := \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(u)}{\|u\|}$. We also define the projection operator $\Pi_{\bar{\mathcal{M}}}(u) := \operatorname{argmin}_{v \in \bar{\mathcal{M}}} \|u - v\|_2$.

2.2. Regularized Convex Program Estimators

We now briefly review key regularized convex program based estimators for high-dimensional linear regression focused on the sparse structure case, where the underlying true parameter θ^* is sparse, so that denoting the non-zero indices by $S(\theta^*) := \{i \in \{1, \ldots, p\} \mid \theta^*_i \neq 0\}$, the sparsity assumption constrains the cardinality $k = |S(\theta^*)|$ as a function of the problem size p.

The LASSO estimator (Tibshirani, 1996) solves the following ℓ_1 regularized least squares problem:

$$\underset{\theta}{\text{minimize}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$$

The Dantzig estimator (Candes & Tao, 2007) solves the following linear program:

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \ \|\theta\|_1 \\ &\text{s.t. } \frac{1}{n} \|X^\top (X\theta - y)\|_\infty \leq \lambda_n. \end{aligned} \tag{2}$$

This seeks a parameter $\widehat{\theta}$ with minimum ℓ_1 norm, and that yet satisfies a key component of the stationary conditions of the LASSO optimization problem.

2.3. Classical Closed-Form Estimators

When $n > p$, and $(X^\top X)$ is full-rank (and hence invertible), the OLS estimator is then given as follows: $\widehat{\theta} = \left(X^\top X\right)^{-1} \left(X^\top y\right)$. However, since the $p \times p$ matrix $(X^\top X)$ can have rank atmost n, the requirement that it is full-rank cannot be satisfied when $p > n$, hence the OLS estimator is no longer well-defined.

Ridge regularized least squares estimator solves the following problem:

$$\widehat{\theta} = \arg\min_{\theta} \left\{ \|y - X\theta\|_2^2 + \epsilon \|\theta\|_2^2 \right\}. \tag{3}$$

It can be seen that this has a unique minimum and is well-defined even in high-dimensional regimes when $p > n$. The unique solution moreover is available in closed-form as: $\widehat{\theta} = (X^\top X + \epsilon I)^{-1} X^\top y$.

2.4. Outline

In the next two sections, we derive variants of the ridge and OLS estimators for *general structurally constrained* high-dimensional linear regression models. We also provide a unified statistical analysis of these two estimators for general structural constraints, deriving corollaries for specific structures. Finally in the last section, we experimentally corroborate the performance of our estimators.

3. The Elem-Ridge Estimator

We will now propose a variant of the standard ridge-regularized least squares estimator (3), where we incorporate the ridge estimator within a structural constraint. Our estimator is specified by any regularization function $\mathcal{R} : \mathbb{R}^p \mapsto \mathbb{R}$. We assume that for the regularization function $\mathcal{R}(\cdot)$, the following "dual" function $\mathcal{R}^* : \mathbb{R}^p \mapsto \mathbb{R}$ is well-defined: $\mathcal{R}^*(u) = \sup_{\theta : \mathcal{R}(\theta) \neq 0} \frac{u^\top \theta}{\mathcal{R}(\theta)}$. When $\mathcal{R}(\cdot)$ is a norm for instance, it can be seen that $\mathcal{R}^*(\cdot)$ would be the corresponding dual norm. Armed with this notation, we consider the following general class of what we call *Elem-Ridge* estimators:

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \ \mathcal{R}(\theta) \\ &\text{s.t. } \mathcal{R}^* \left(\theta - (X^\top X + \epsilon I)^{-1} X^\top y\right) \leq \lambda_n. \end{aligned} \tag{4}$$

The estimator bears similarities to the Dantzig estimator (2) since both of these minimize some structural complexity of the parameter subject to certain constraints. However, unlike the Dantzig estimator, the estimator above is available in closed form for typical settings of the regularization function $\mathcal{R}(\cdot)$. For instance, when $\mathcal{R}(\cdot)$ is set to the ℓ_1 norm, the estimator (4) is given by

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \ \|\theta\|_1 \\ &\quad \text{s.t.} \ \left\|\theta - (X^\top X + \epsilon I)^{-1} X^\top y\right\|_\infty \leq \lambda_n. \end{aligned} \tag{5}$$

This can be seen to have a unique solution available in closed form as: $\widehat{\theta} = S_{\lambda_n}\Big((X^\top X + \epsilon I)^{-1} X^\top y\Big)$, where $[S_\lambda(u)]_i = \text{sign}(u_i)\max(|u_i| - \lambda, 0)$ is the soft-thresholding function.

Another interesting instantiation of $\mathcal{R}(\cdot)$ would be the group structured ℓ_1/ℓ_α norm defined as $\|\theta\|_{\mathcal{G},\alpha} := \sum_{g=1}^{L} \|\theta_{G_g}\|_\alpha$, where $\mathcal{G} := \{G_1, G_2, \ldots, G_L\}$ is a set of disjoint subsets/groups of the index-set $\{1, \ldots, p\}$ and α is a constant between 2 and ∞. With respect to this group norm, the estimator (4) will have the form of

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \ \|\theta\|_{\mathcal{G},\alpha} \\ &\quad \text{s.t.} \ \left\|\theta - (X^\top X + \epsilon I)^{-1} X^\top y\right\|^*_{\mathcal{G},\alpha} \leq \lambda_n \end{aligned}$$

where $\|\theta\|^*_{\mathcal{G},\alpha} := \max_g \|\theta_{G_g}\|_{\alpha^*}$ for a constant α^* satisfying $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$. At the time same time, the soft-thresholding operator for the group, $S_{\mathcal{G},\lambda}$ can be extended as follows: for any group g in $\mathcal{G}$, $[S_{\mathcal{G},\lambda}(u)]_g = \max(\|u_g\|_\alpha - \lambda, 0)\frac{u_g}{\|u_g\|_\alpha}$, and hence the optimal solution will have a closed-form as previous ℓ_1 case: $\widehat{\theta} = S_{\mathcal{G},\lambda_n}\Big((X^\top X + \epsilon I)^{-1} X^\top y\Big)$.

3.1. Error Bounds

We now provide a unified statistical analysis of the class of estimators in (4), for general structures, and general regularization functions $\mathcal{R}(\cdot)$. We follow the structural constraint notation of Negahban et al. (2012) detailed in the background section and assume the following:

(C1) The norm in the objective $\mathcal{R}(\cdot)$ is decomposable with respect to the subspace-pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$.

(C2) There exists a structured subspace-pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ such that the regression parameter satisfies $\Pi_{\mathcal{M}^\perp}(\theta^*) = \mathbf{0}$.

In (C2), we consider the case where θ^* is *exactly* sparse with respect to the subspace pair for technical simplicity.

Theorem 1. *Consider the linear regression model (1) where the conditions (C1) and (C2) hold. Suppose we solve the estimation problem (4) setting the constraint bound λ_n such that $\lambda_n \geq \mathcal{R}^*(\theta^* - (X^\top X + \epsilon I)^{-1} X^\top y)$. Then the optimal solution $\widehat{\theta}$ satisfies the following error bounds:*

$$\begin{aligned} &\mathcal{R}^*\big(\widehat{\theta} - \theta^*\big) \leq 2\lambda_n \ , \\ &\|\widehat{\theta} - \theta^*\|_2 \leq 4\Psi(\mathcal{M})\lambda_n \ , \\ &\mathcal{R}\big(\widehat{\theta} - \theta^*\big) \leq 8[\Psi(\mathcal{M})]^2\lambda_n \ . \end{aligned}$$

We note that Theorem 1 is a non-probabilistic result, and holds deterministically for any selection of λ_n or any distributional setting of the covariates X. It is also worthwhile to note that the conditions of the theorem entail that the "initial estimator" consisting of the standard ridge-regularized least squares estimator is in turn consistent with respect to the $\mathcal{R}^*(\cdot)$ norm. However, embedding this initial estimator within a structural constraint as in our Elem-Ridge estimator (4) allows us to guarantee additional error bounds in terms of $\mathcal{R}(\cdot)$ and ℓ_2 norms, which **do not hold** for the initial ridge-regularized estimator. While the statement of the theorem is a bit abstract, we derive its consequences under specific structural settings as corollaries.

3.2. Sparse Linear Models

We now derive a corollary of Theorem 1 for the specific case where θ^* is sparse with k non-zero entries. The condition described in (C2) can be written in this case as:

(C3) The regression parameter θ^* is exactly sparse with k non-zero entries. As discussed in Example 1, it is natural to select $\mathcal{M}(S)$ equal to the support set of θ^*.

We analyze the variant (5) which sets the regularization function $\mathcal{R}(\cdot)$ in (4) to the ℓ_1 norm. Note that the condition (C1) is automatically satisfied with this selection of regularization function since ℓ_1 norm is decomposable with respect to $\mathcal{M}(S)$ and its orthogonal complement, as discussed in Example 1. The only remaining issue to appeal to Theorem 1, is to set λ_n satisfying the condition in the statement: $\lambda_n \geq \|\theta^* - (X^\top X + \epsilon I)^{-1} X^\top y\|_\infty$. To do so, we leverage the analysis of the classical ridge-regression estimator from Zhang et al. (2008a), where they impose the following assumption:

(C-Ridge) Let $e_1, \ldots, e_q, e_{q+1}, \ldots, e_p$ be the singular vectors of $\frac{1}{n}X^\top X$ corresponding to the singular values $d_1 \geq \ldots \geq d_q > d_{q+1} = \ldots = d_p = 0$ where q is the rank of $\frac{1}{n}X^\top X$. Let $\theta^* = \sum_{i=1}^{p} \theta_i e_i$. Then, $\|\sum_{i=q+1}^{p} \theta_i e_i\|_\infty = O(\xi)$ with some sequence $\xi \to 0$.

Note that this assumption is trivially satisfied if $n \geq p$ and $X^\top X$ has full rank. When $p \gg n$, however, this assumption plays a role as an identifiable condition for ℓ_∞ consistency so that the penalty term favors true parameter over any others (see Zhang et al. (2008a) for details).

Appealing to Theorem 3.1 in Zhang et al. (2008a), under Condition (C-Ridge), the classical ridge-estimator satisfies the following error bound:

$$\|\theta^* - (X^\top X + \epsilon I)^{-1} X^\top y\|_\infty \leq O\left(\left(\frac{\sqrt{k}\log p}{n d_q}\right)^{1/3}\right)$$

where k is the sparsity level of θ^* in (C3). Note that this does not entail that the classical ridge-estimator would also be ℓ_2 or ℓ_1 norm consistent. But we can provide such bounds for the estimator in (4) by deriving the following corollary of Theorem 1.

Corollary 1. *Consider the optimization problem in (5), and suppose that all the conditions in (C3) and (C-Ridge) are satisfied. Furthermore, suppose also that we select*

$$\lambda_n := O\left(\left(\frac{\sqrt{k}\log p}{n d_q}\right)^{1/3}\right).$$

Then, there are universal positive constants (c_1, c_2) such that any optimal solution $\widehat{\theta}$ of (5) satisfies

$$\|\widehat{\theta} - \theta^*\|_\infty \leq O\left(\left(\frac{\sqrt{k}\log p}{n d_q}\right)^{1/3}\right),$$
$$\|\widehat{\theta} - \theta^*\|_2 \leq O\left(\left(\frac{k^2\log p}{n d_q}\right)^{1/3}\right),$$
$$\|\widehat{\theta} - \theta^*\|_1 \leq O\left(\left(\frac{k^3\sqrt{k}\log p}{n d_q}\right)^{1/3}\right)$$

with probability at least $1 - c_1\exp(-c_2 p)$.

Even though our estimator in (5) is consistent in ℓ_2 and ℓ_1 norms, its convergences rate can be seen to be inferior when compared with those of the LASSO; if we compare ℓ_2 error bounds for instance, the LASSO estimator satisfies $\|\widehat{\theta}_{LASSO} - \theta^*\|_2 \leq O\left(\sqrt{\frac{k\log p}{n}}\right)$ under some standard conditions such as restricted eigenvalue condition (Negahban et al., 2012). For reasons of space, we thus defer deriving corollaries of Theorem 1 for other structures such as group-sparsity, and focus on a variant of the OLS estimator with faster convergence rates in the next section.

4. The Elem-OLS Estimator

As noted in our discussion of the classical OLS estimator in Section 2.3, in the high-dimensional regime $p > n$, the matrix $X^\top X$ is rank-deficient, and the classical OLS estimator $(X^\top X)^{-1}X^\top y$ is no longer well-defined since $X^\top X$ is not invertible. In this section, we thus consider the following simple variant of the OLS estimator.

For any matrix A, we first define the following element-wise operator T_ν:

$$\left[T_\nu(A)\right]_{ij} = \begin{cases} A_{ii} + \nu & \text{if } i = j \\ \text{sign}(A_{ij})(|A_{ij}| - \nu) & \text{otherwise, } i \neq j \end{cases}$$

Suppose we apply this element-wise operator T_ν to the sample covariance matrix $\frac{X^\top X}{n}$, to obtain $T_\nu\left(\frac{X^\top X}{n}\right)$. We will now show that $T_\nu\left(\frac{X^\top X}{n}\right)$ will be invertible with high probability, even under high-dimensional settings, provided the following conditions hold:

(C-OLS1) (Σ-*Gaussian ensemble*) Each row of the design matrix $X \in \mathbb{R}^{n\times p}$ is i.i.d. sampled from $N(0, \Sigma)$.
(C-OLS2) The design matrix X is column normalized.

Lastly, we impose the following condition on the population matrix Σ:

(C-OLS3) The covariance Σ of Σ-Gaussian ensemble is strictly diagonally dominant: for all row i, $\delta_i := \Sigma_{ii} - \sum_{j\neq i}|\Sigma_{ij}| \geq \delta_{\min} > 0$ where $\delta_{\min}$ is a large enough constant so that $\|\Sigma\|_\infty \leq \frac{1}{\delta_{\min}}$.

Condition (C-OLS3) that Σ be strictly diagonally dominant is different from (and possibly stronger) than the conventional restricted eigenvalue assumption for the LASSO. As we will show in the sequel, this assumption guarantees that (a) the matrix $T_\nu[(X^T X)/n]$ is invertible, and (b) its induced ℓ_∞ norm is well bounded. We note that there could be more general cases under which $T_\nu[(X^T X)/n]$ satisfies these two conditions, and defer relaxing the assumption to future work.

We then have the following proposition.

Proposition 1. *Suppose conditions (C-OLS1) and (C-OLS3) hold. Then for any $\nu \geq 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau\log p'}{n}}$, the matrix $T_\nu\left(\frac{X^\top X}{n}\right)$ is invertible with probability at least $1 - 4/p'^{\tau-2}$ for $p' := \max\{n, p\}$ and any constant $\tau > 2$.*

Our key idea is to then use this *invertible* matrix $T_\nu\left(\frac{X^\top X}{n}\right)$ to modify the OLS estimator, and embed this within a structural constraint, so as to obtain the following class of what we call "Elem-OLS" estimators:

$$\begin{aligned} &\underset{\theta}{\text{minimize}}\ \mathcal{R}(\theta) \\ &\text{s. t. } \mathcal{R}^*\left(\theta - \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\frac{X^\top y}{n}\right) \leq \lambda_n. \end{aligned} \tag{6}$$

As in Elem-Ridge estimators discussed in the beginning of Section (4), the estimators from (6) are also available in closed form for typical settings of the regularization function $\mathcal{R}(\cdot)$.

4.1. Error Bounds

As for our earlier Elem-Ridge estimators, here too we provide a unified statistical analysis of Elem-OLS estimators in (6), for general structures, and general regularization functions $\mathcal{R}(\cdot)$. Specifically, we obtain the following counterpart of Theorem 1:

Theorem 2. *Consider the linear regression model (1) where the conditions (C1) and (C2) hold. Suppose we solve the estimation problem (6) setting the constraint bound λ_n such that $\lambda_n \geq \mathcal{R}^*\Big(\theta^* - \big[T_\nu\big(\frac{X^\top X}{n}\big)\big]^{-1}\frac{X^\top y}{n}\Big)$. Then the optimal solution $\widehat{\theta}$ satisfies the following error bounds:*

$$\mathcal{R}^*\big(\widehat{\theta} - \theta^*\big) \leq 2\lambda_n \ ,$$
$$\|\widehat{\theta} - \theta^*\|_2 \leq 4\Psi(\mathcal{M})\lambda_n \ ,$$
$$\mathcal{R}\big(\widehat{\theta} - \theta^*\big) \leq 8[\Psi(\mathcal{M})]^2\lambda_n \ .$$

As in the class of Elem-Ridge estimators, here we have an "initial estimator" consisting of a (novel) OLS-esque estimator $\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\frac{X^\top y}{n}$. The conditions of our theorem guarantee that even this initial estimator is consistent with respect to $\mathcal{R}^*(\cdot)$ norm. However, embedding this initial estimator within a structural constraint as in our Elem-OLS estimator (6) allows us to guarantee additional error bounds in terms of $\mathcal{R}(\cdot)$ and ℓ_2 norms, which do not hold for our initial OLS-esque estimator.

We now derive the consequences of our abstract theorem under specific structural settings.

4.2. Sparse Linear Models

Consider the case when the true parameter θ^* is sparse with k non-zero elements. We then consider the variant of (6) with the regularization function $\mathcal{R}(\cdot)$ set to the ℓ_1 norm:

$$\underset{\theta}{\text{minimize}} \ \|\theta\|_1 \quad \text{s.t.} \ \left\|\theta - \Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\frac{X^\top y}{n}\right\|_\infty \leq \lambda_n. \tag{7}$$

We can then derive the convergence rate of this estimator (7) as a corollary of Theorem 2:

Corollary 2. *Under the condition (C3), consider the optimization problem (7), setting $\nu := 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau\log p'}{n}} := a\sqrt{\frac{\log p'}{n}}$, and $p' := \max\{n, p\}$. Suppose that all the conditions in (C-OLS1)-(C-OLS3) are satisfied. Furthermore, suppose also that we select*

$$\lambda_n := \frac{1}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right).$$

Then, there are universal positive constants (c_1, c_2) such that any optimal solution $\widehat{\theta}$ of (7) satisfies

$$\|\widehat{\theta} - \theta^*\|_\infty \leq \frac{2}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right),$$
$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4}{\delta_{\min}}\left(2\sigma\sqrt{\frac{k\log p'}{n}} + a\sqrt{\frac{k\log p'}{n}}\|\theta^*\|_1\right),$$
$$\|\widehat{\theta} - \theta^*\|_1 \leq \frac{8}{\delta_{\min}}\left(2\sigma k\sqrt{\frac{\log p'}{n}} + ak\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right)$$

with probability at least $1 - c_1\exp(-c_2 p')$.

The rates in Corollary 2 are the almost same as those by standard LASSO; for instance, $\|\widehat{\theta}_{LASSO} - \theta^*\|_2 \leq O\Big(\sqrt{\frac{k\log p}{n}}\Big)$ analyzed in Negahban et al. (2012) even though the rates here have an additional $\|\theta^*\|_1$ term, which are negligible if k and $\|\theta^*\|_\infty$ are bounded by constants.

4.3. Group-sparse Linear Models

Another interesting structural constraint arises when θ^* is *group-sparse*. Consider a set of disjoint subsets/groups $\mathcal{G} := \{G_1, G_2, \ldots, G_L\}$ of the index-set $\{1, \ldots, p\}$, each of size at most $|G_j| \leq m$. For any vector $u \in \mathbb{R}^p$, let u_{G_j} denote the subvector with indices restricted to the set G_j. We assume that the model parameter θ^* is group-sparse with respect to these groups $\mathcal{G}$ so that the condition described in (C2) can be written in this case as:

(C4) The linear regression parameter θ^* has group-support of size atmost k, so that $|\{j \in \{1, \ldots, L\} : \theta^*_{G_j} \neq 0\}| \leq k$.

A natural regularization function for such a setting is the following group-structured ℓ_1/ℓ_α norm defined as $\|\theta\|_{\mathcal{G},\alpha} := \sum_{g=1}^L \|\theta_{G_g}\|_\alpha$, where α is a constant between 2 and ∞.

In this case, the assumption of column normalization in (C-OLS2) can be naturally generalized (Negahban et al., 2012):

(C-OLS4) Define the operator norm $\|\!|X_G|\!\|_{\alpha\to 2} := \max_{\|w\|_\alpha = 1}\|X_G w\|_2$. Then, $\frac{\|\!|X_{G_j}|\!\|_{\alpha\to 2}}{\sqrt{n}} \leq 1$ for all $j = 1, \ldots, L$.

We then consider the following variant of (6), with the regularization function $\mathcal{R}(\cdot)$ set to the above group-structured norm:

$$\underset{\theta}{\text{minimize}} \ \|\theta\|_{\mathcal{G},\alpha} \quad \text{s.t.} \ \left\|\theta - \Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\frac{X^\top y}{n}\right\|^*_{\mathcal{G},\alpha} \leq \lambda_n. \tag{8}$$

where $\|\theta\|^*_{\mathcal{G},\alpha} := \max_g \|\theta_{G_g}\|_{\alpha^*}$ for a constant α^* satisfying $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$.

Now we can derive the convergence rates of this estimator (8) as a corollary of Theorem 2:

Corollary 3. *Under the condition (C4), consider the optimization problem (8), setting $\nu := 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau\log p'}{n}} := a\sqrt{\frac{\log p'}{n}}$, and $p' := \max\{n, p\}$. Suppose that all the conditions in (C-OLS1), (C-OLS3) and (C-OLS4) are satisfied. Furthermore, suppose also that we*

Table 1. Average performance measure and standard deviation in parenthesis for ℓ_1-penalized comparison methods on simulated data for sparse linear models.

	Method	TP	FP	ℓ_2	ℓ_∞
n=1000,p=1000	Elem-OLS	**100.00 (0.00)**	**2.05 (1.15)**	**0.551 (0.071)**	**0.255 (0.041)**
	Elem-Ridge	**100.00 (0.00)**	2.44 (2.12)	0.741 (0.411)	0.435 (0.064)
	LASSO	**100.00 (0.00)**	9.84 (2.45)	0.563 (0.067)	0.270 (0.039)
	Thr-LASSO	**100.00 (0.00)**	8.33 (1.14)	0.560 (0.066)	0.274 (0.071)
	OMP	98.24 (0.64)	3.20 (1.38)	0.559 (0.113)	0.282 (0.055)
n=1000,p=2000	Elem-OLS	**100.00 (0.00)**	**2.22 (2.02)**	**0.656 (0.111)**	**0.314(0.071)**
	Elem-Ridge	**100.00 (0.00)**	11.94 (4.48)	3.8834 (0.411)	1.678 (0.349)
	LASSO	**100.00 (0.00)**	18.88 (6.93)	0.657(0.110)	0.316(0.075)
	Thr-LASSO	99.59(0.36)	14.35(2.66)	**0.656 (0.099)**	0.315(0.052)
	OMP	96.36(1.00)	10.25 (4.24)	0.735(0.222)	0.536(0.136)

select

$$\lambda_n := \frac{m^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right).$$

Then, there are universal positive constants (c_1, c_2) *such that any optimal solution* $\widehat{\theta}$ *of* (8) *satisfies*

$$\|\widehat{\theta} - \theta^*\|^*_{\mathcal{G},\alpha} \leq \frac{2m^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right),$$

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4\sqrt{k}m^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right),$$

$$\|\widehat{\theta} - \theta^*\|_{\mathcal{G},\alpha} \leq \frac{8km^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right)$$

with probability at least $1 - c_1\exp(-c_2 p')$.

5. Experiments

We demonstrate the performance of our elementary estimators on simulated as well as real-world datasets.

5.1. Simulation study

We first provide simulation studies that corroborate our theoretical results, and furthermore compares the finite sample performance of various methods.

Sparse linear models For our first set of simulations, we generate data according to the linear model $y = X\theta^* + w$. We construct the $n \times p$ design matrices X by sampling the rows independently from a multivariate Gaussian distribution $N(0, \Sigma)$ where $\Sigma_{i,j} = 0.5^{|i-j|}$. We then set the error term as $w \sim N(0, 1)$. For each simulation, the entries of the true model coefficient vector θ^* are set to be 0 everywhere, except for a randomly chosen subset of 10 coefficients, which are chosen independently and uniformly in the interval $(1, 3)$. We set the number of samples to $n = 1000$, and the number of covariates among $p \in \{1000, 2000\}$. We compare the performance of *Elem-OLS*: our OLS-variant elementary estimator in (7), with the operator T_ν and the regularization function $\mathcal{R}(\cdot)$ set to the ℓ_1 norm; *Elem-Ridge*: our ridge-variant elementary estimator in (5) with the regularization function $\mathcal{R}(\cdot)$ set to the ℓ_1 norm, *LASSO*: the standard LASSO estimator, *Thr-LASSO*: the LASSO estimator followed by post hard-thresholding and OLS refitting (van de Geer et al., 2011), and *OMP*: the Orthogonal Matching Pursuit estimator (Zhang, 2010). As performance measures, we used the True Positive Rate (TP), False Positive Rate (FP), ℓ_∞ and ℓ_2 error between the estimated and true parameter vectors. While our theorem specified an optimal setting of the regularization parameter λ, this optimal setting depended on unknown model parameters. Thus, as is standard with high-dimensional regularized convex programs, we set the tuning parameters in a holdout-validated fashion, as those that minimize the average squared error on an independent validation set of sample size n. For each setting, we present the average of the performance measures based on 100 simulations in Table 1. As can be seen from the table, the performance of Elem-OLS in term of ℓ_2 and ℓ_∞ errors is competitive with that of LASSO, which corroborates our statistical analyses of Section 4.2. In addition, the simulation results confirm the suboptimal performance of Elem-Ridge – especially as p becomes larger than n. Elem-OLS even achieves *superior* variable selection accuracy when compared to the LASSO. This is further illustrated in Figure 1, which depicts ROC curves for the various methods, i.e., how TP and FP evolve as the amount of regularization (stopping point for OMP) is varied. We find this remarkable, given that Elem-OLS is available in *closed-form*, whereas LASSO, OMP and variants require an iterative optimization procedure.

Group-sparse linear models The data is generated following the lines of the third model of Yuan & Lin (2006). We compare *Group-LASSO*, *Thr-Group-LASSO* (Group-LASSO followed by hard-thresholding), *Group-OMP* (Lozano et al., 2009), with our Elem-OLS estimator in (8) for group sparsity, which we refer to as *Group-Elem-OLS*. We present the average of the performance measures based on 100 simulations in Table 2. As can be seen from

Table 2. Average performance measure and standard deviation in parenthesis for ℓ_1-penalized comparison methods on simulated data for group-sparse linear models.

Method	TP	FP	ℓ_2	ℓ_∞
Group-Elem-OLS	**100 (0.00)**	**0.9 (0.10)**	1.300 (0.045)	**0.613 (0.027)**
Group-LASSO	**100 (0.00)**	1.8 (0.18)	**1.269 (0.095)**	0.642 (0.031)
Thr-Group-LASSO	99 (0.14)	1.2 (0.16)	1.296 (0.075)	0.628 (0.029)
Group-OMP	99 (0.16)	1.9 (0.15)	1.984(0.080)	0.642(0.030)

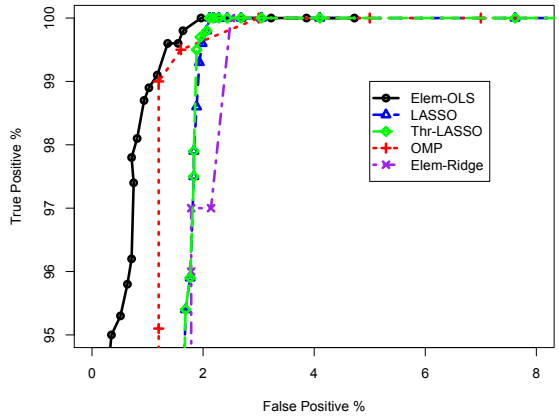

Figure 1. ROC curves for the Elem-OLS, Elem-Ridge, LASSO, Thresholded-LASSO and OMP estimators under sparse linear model.

the table, the performance of Group-Elem-OLS is comparable to the other methods in term of ℓ_2 and ℓ_∞ errors are comparable. Group-Elem-OLS achieves superior variable selection accuracy.

5.2. Real Data Analysis

We employed our estimators toward the analysis of gene expression data. We used microarray data pertaining to isoprenoid biosynthesis in Arabidopsis thaliana (A. thaliana) provided by Wille et al. (2004). A. thaliana is a plant widely used as model organism in genetics. Isoprenoids play key roles in major plant processes including photosynthesis, respiration and defense against pathogens. Here we focus on identifying the genes that exhibit significant association with the specific isoprenoid gene GGPPS11 (AGI code AT4G36810), which is known as a precursor to chlorophils, carotenoids, tocopherols and abscisic acids. Thus, the response is the expression level of the isoprenoid gene GGPPS11, while as covariates, we have the expression levels from 833 genes, coming from 58 different metabolic pathways. There are 131 samples. All variables are log transformed. Due to space limitations, we only report results for our OLS-variant elementary estimator with ℓ_1 regularization (Elem-OLS) and for the LASSO estimator. We evaluate the predictive accuracy of the methods by randomly partitioning the data into training and test sets, using 90 observations for training and the remainder for testing. The tuning parameters were selected using 5 fold cross-validation. We computed the prediction MSE for the testing set. The results for 20 random train/test partitions are presented in Table 3. Overall, our elementary estimator achieves superior prediction accuracy. We now look into the genes selected by the comparison methods on the full dataset. Out of 833 candidate genes, Elem-OLS and LASSO selected selected 79 and 73 genes respectively. For each method we ranked the selected genes according to the amplitude of their regression coefficients. The top 20 genes for each method are listed in Table 4 along with their associated pathways (the pathway names are abbreviated). From table 4 we can see Elem-OLS and LASSO have 7 genes in common among their respective top 20 genes. Interestingly, the gene AT1G17050 (a.k.a. PPDS1), which is known to belong to the same pathway as target gene GGPPS11 (isoprenoid non-mevalonate), is included in the top-20 genes of Elem-OLS, but not of LASSO.

Table 3. Average test MSE on microarray dataset. (Smaller values indicate higher predictive accuracy).

Elem-OLS	LASSO
$\mathbf{10.52 \pm 0.39}$	11.59 ± 0.39

Table 4. Top 20 selected genes for the microarray study, along with their associated pathways.

Elem-OLS		LASSO	
Pathway	*Gene (AGI)*	*Pathway*	*Gene (AGI)*
Carote	AT1G57770	Carote	AT1G57770
Cytoki	AT3G63110	Citrate	AT1G54340
Flavon	AT2G38240	Ethyl	AT1G01480
Flavon	AT5G08640	Ethyl	AT4G11280
Folate	AT1G78670	Flavon	AT5G08640
Glycer	AT2G44810	Folate	AT1G78670
Glycol	AT5G50850	Inosit	AT3G56960
Inosit	AT2G31830	Inosit	AT2G41210
Inosit	AT3G56960	Phenyl	AT2G27820
Non-Mev	AT1G17050	Porphy	AT3G51820
Phenyl	AT2G27820	Pyrimi	AT5G23300
Phytos	AT1G58440	Ribofl	AT4G13700
Porphy	AT1G09940	SGC	AT5G28030
Ribofl	AT4G13700	Starch	AT2G45880
Starch	AT2G45880	Starch	AT2G25540
Starch	AT5G03650	Starch	AT5G29958
Threo	AT1G72810	Toco	AT2G18950
Toco	AT2G18950	Trypto	AT5G17980
Trypto	AT5G48220	Trypto	AT3G03680

Acknowledgments

E.Y and P.R. acknowledge the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, DMS-1264033.

References

Bach, F. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Structured sparsity through convex optimization. *Statistical Science*, 27 (4):450–468, 2012.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. 37(4):1705–1732, 2009. Annals of Statistics.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

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

Donoho, D. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.

Fan, J. and Lv, J. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSSB)*, 70:849–911, 2008.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics*, 2007.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research (JMLR)*, 13:2107–2143, 2012.

Hsieh, C. J., Sustik, M., Dhillon, I., and Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. In *Neur. Info. Proc. Sys. (NIPS)*, 24, 2011.

Huang, J., Zhang, T., and Metaxas, D. Learning with structured sparsity. *Journal of Machine Learning Research (JMLR)*, 12: 3371–3412, 2011.

Jacob, L., Obozinski, G., and Vert, J. P. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pp. 433–440, 2009.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

Lozano, A. C., Swirszcz, G., and Abe, N. Group orthogonal matching pursuit for variable selection and prediction. In *Neur. Info. Proc. Sys (NIPS)*, 2009.

Mallat, S. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, December 1993.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34: 1436–1462, 2006.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37 (1):246–270, 2009.

Negahban, S. and Wainwright, M. J. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. Technical report, Department of Statistics, UC Berkeley, April 2009.

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

Obozinski, G., Wainwright, M. J., and Jordan, M. I. Union support recovery in high-dimensional multivariate regression. Technical report, Department of Statistics, UC Berkeley, August 2008.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5: 935–980, 2011.

Recht, B., Fazel, M., and Parrilo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, Vol 52(3):471–501, 2010.

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

Tropp, J. A., Gilbert, A. C., and Strauss, M. J. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

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

van de Geer, S., Bühlmann, P., and Zhou, S. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Varah, J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

Wille, A., Zimmermann, P., and Vranova, E. [and others]. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5, 2004.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

Zhang, J., Jeng, X. J., and Liu, H. Some two-step procedures for variable selection in high-dimensional linear regression. *Arxiv preprint arXiv:0810.1644*, 2008a.

Zhang, T. Sparse recovery with orthogonal matching pursuit under rip. Tech Report arXiv:1005.2249, May 2010.

Zhang, Z., Dolecek, L., Nikolic, B., Anantharam, V., and Wainwright, M. J. Lowering LDPC error floors by post-processing. In *Proc. IEEE GLOBECOM*, September 2008b.

Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

Zhao, P., Rocha, G., and Yu, B. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

Appendix

A. Proof of Theorem 1 and Theorem 2

The proofs of two theorems are almost identical with a single difference selecting initial parameter on which the soft-thresholding is performed. In the proof, we denote this initial parameter, i.e., $(X^\top X + \epsilon I)^{-1} X^\top y$ or $\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} \frac{X^\top y}{n}$ by $\bar{\theta}$.

Let Δ be the error vector, $\widehat{\theta} - \theta^*$. Since we choose λ_n greater than $\mathcal{R}^*(|\theta^* - \bar{\theta})$,

$$\begin{aligned}\mathcal{R}^*(\Delta) &= \mathcal{R}^*(\widehat{\theta} - \bar{\theta} + \bar{\theta} - \theta^*) \\ &\leq \mathcal{R}^*(\widehat{\theta} - \bar{\theta}) + \mathcal{R}^*(\theta^* - \bar{\theta}) \leq 2\lambda_n \end{aligned} \tag{9}$$

where we utilize the fact that $\widehat{\theta}$ is feasible.

For notational simplicity, we use (S, S^c) instead of an arbitrary subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$. Additionally, we use the notion Δ_S to represent the ℓ_2 projection onto the model space $\mathcal{M}$. Then, by the assumption of the statement that $\theta^*_{S^c} = \mathbf{0}$, and the decomposability of $\mathcal{R}(\cdot)$ with respect to (S, S^c),

$$\begin{aligned}\mathcal{R}(\theta^*) &= \mathcal{R}(\theta^*) + \mathcal{R}(\Delta_{S^c}) - \mathcal{R}(\Delta_{S^c}) \\ &= \mathcal{R}(\theta^* + \Delta_{S^c}) - \mathcal{R}(\Delta_{S^c}) \\ &\overset{(i)}{\leq} \mathcal{R}(\theta^* + \Delta_{S^c} + \Delta_S) + \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \\ &= \mathcal{R}(\theta^* + \Delta) + \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \end{aligned} \tag{10}$$

where the equality (i) holds by the triangle inequality, which is the basic property of norms. Since we minimize the objective $\mathcal{R}(\theta)$ in (4) or (6), we obtain the inequality of $\mathcal{R}(\theta^* + \Delta) = \mathcal{R}(\widehat{\theta}) \leq \mathcal{R}(\theta^*)$. Combining this inequality with (10), we have

$$0 \leq \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \tag{11}$$

Armed with inequalities (9) and (11), we utilize the Hölder's inequality and the decomposability of our regularizer $\mathcal{R}(\cdot)$ in order to derive the error bounds in terms of ℓ_2 norm:

$$\begin{aligned}\|\Delta\|_2^2 = \langle \Delta, \Delta \rangle &\leq \mathcal{R}^*(\Delta)\mathcal{R}(\Delta) \\ &\leq \mathcal{R}^*(\Delta)\left(\mathcal{R}(\Delta_S) + \mathcal{R}(\Delta_{S^c})\right).\end{aligned}$$

Since the error vector Δ satisfies the inequality (11),

$$\|\Delta\|_2^2 \leq 2\,\mathcal{R}^*(\Delta)\,\mathcal{R}(\Delta_S).$$

Combining all the pieces together yields

$$\|\Delta\|_2^2 \leq\ 4\Psi(S)\lambda_n \|\Delta_S\|_2 \tag{12}$$

where $\Psi(\mathcal{M})$ is the abbreviation for $\Psi(S, \|\cdot\|_2)$.

Notice that the projection operator is non-expansive, $\|\Delta_S\|_2^2 \leq \|\Delta\|_2^2$. Hence, we obtain $\|\Delta_S\|_2 \leq 4\Psi(S)\lambda_n$, and plugging it back into (12) yields the ℓ_2 error bounds.

Finally, the error bounds in terms of the regularizer itself are straightforward from the following reasoning:

$$\begin{aligned}\mathcal{R}(\Delta) &= \mathcal{R}(\Delta_S) + \mathcal{R}(\Delta_{S^c}) \leq 2\mathcal{R}(\Delta_S) \\ &\leq 2\Psi(S)\|\Delta_S\|_2 \leq 8[\Psi(S)]^2\lambda_n.\end{aligned}$$

B. Useful lemma(s)

Lemma 1 (Lemma 1 of (Ravikumar et al., 2011)). *Let $\mathcal{A}$ be the event that*

$$\left\|\frac{X^\top X}{n} - \Sigma\right\|_\infty \leq 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p'}{n}}$$

where $p' := \max\{n, p\}$ and τ is any constant greater than 2. Suppose that the design matrix X is i.i.d. sampled from Σ-Gaussian ensemble *with $n \geq 40 \max_i \Sigma_{ii}$. Then, the probability of event $\mathcal{A}$ occurring is at least $1 - 4/p'^{\tau-2}$.*

Lemma 2 (In the proof of Corollary 2 (Negahban et al., 2012)). *By the conditions of (C-OLS2), and the sub-Gaussian property of noise w,*

$$P\left(\frac{\|X^\top w\|_\infty}{n} \geq t\right) \leq 2\exp(-\frac{nt^2}{2\sigma^2} + \log p)$$

C. Proof of Proposition 1

By Lemma 1, we have the event $\mathcal{A}$:

$$\left\|\frac{X^\top X}{n} - \Sigma\right\|_\infty \leq 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p'}{n}}$$

with high probability specified in the statement of lemma. Conditioned on $\mathcal{A}$, $T_\nu\left(\frac{X^\top X}{n}\right)$ with the specific choice of ν in the statement, has larger diagonal entries and smaller off-diagonal entries than Σ. Therefore, on the $\mathcal{A}$, $T_\nu\left(\frac{X^\top X}{n}\right)$ is diagonally dominant, and hence invertible.

D. Proof of Corollary 2

In order to utilize Theorem 2, we need to derive the upper bound of $\left\|\theta^* - \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} \frac{X^\top y}{n}\right\|_\infty$:

$$\begin{aligned}&\|\theta^* - \bar{\theta}\|_\infty \\ &= \left\|\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} T_\nu\left(\frac{X^\top X}{n}\right)\theta^* - \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\frac{X^\top y}{n}\right\|_\infty \\ &\leq \left|\left|\left|\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\right|\right|\right|_\infty \left\|T_\nu\left(\frac{X^\top X}{n}\right)\theta^* - \frac{X^\top y}{n}\right\|_\infty\end{aligned}$$

We first control $\left|\left|\left|\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\right|\right|\right|_\infty$ term. We are going to show that $T_\nu\left(\frac{X^\top X}{n}\right)$ is diagonally dominant with high

probability hence the term we care about will be bound. By Lemma 1, if $n > 40 \max_i \Sigma_{ii}$, the event $\mathcal{A}$ occurs with probability at least $1 - 4/p'^{\tau-2}$ for $p' := \max\{n, p\}$ and any constant $\tau > 2$. Conditioned on $\mathcal{A}$, for all row index i,

$$\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ii} - \sum_{j\neq i}\left|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ij}\right|$$
$$\geq \left(\Sigma_{ii} - a\sqrt{\frac{\log p'}{n}} + \nu\right) - \sum_{j\neq i}\left(|\Sigma_{ij}| + a\sqrt{\frac{\log p'}{n}} - \nu\right).$$

where $a := 8(\max_i \Sigma_{ii})\sqrt{10\tau}$.

Therefore, provided $\nu := a\sqrt{\frac{\log p'}{n}}$,

$$\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ii} - \sum_{j\neq i}\left|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ij}\right|$$
$$\geq \Sigma_{ii} - \sum_{j\neq i}|\Sigma_{ij}| \geq \delta_i \geq \delta_{\min}.$$

Note that conditioned on $\mathcal{A}$, the matrix $T_\nu\big(\frac{X^\top X}{n}\big)$ is invertible since it is strictly diagonally dominant matrix, and $\Big\|\Big[T_\nu\big(\frac{X^\top X}{n}\big)\Big]^{-1}\Big\|_\infty \leq \frac{1}{\delta_{\min}}$ by Varah (1975).

Now consider the second term $\Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top y}{n}\Big\|_\infty$ in the equality:

$$\Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top y}{n}\Big\|_\infty$$
$$= \Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^* + \frac{X^\top X}{n}\theta^* - \frac{X^\top y}{n}\Big\|_\infty$$
$$\leq \Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^* - \frac{X^\top w}{n}\Big\|_\infty$$
$$\leq \Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^*\Big\|_\infty + \Big\|\frac{X^\top w}{n}\Big\|_\infty.$$

Since $\frac{\|X^\top w\|_\infty}{n}$ can be upper-bounded by $2\sigma\sqrt{\frac{\log p}{n}}$ as stated in Lemma 2, the only remaining term to control is $\Big\|\Big(T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}\Big)\theta^*\Big\|_\infty$. Each element of $T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}$ is upper-bounded by ν by construction, which is set $a\sqrt{\frac{\log p'}{n}}$. Therefore, for every entry of $\Big(T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}\Big)\theta^*$, we can apply Hölder inequality so that it is bound by $a\sqrt{\frac{\log p}{n}}\|\theta^*\|_1$.

Therefore, if we select λ_n as

$$\frac{1}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right),$$

the constraint $\|\theta^* - \bar{\theta}\|_\infty \leq \lambda_n$ with high probability, which completes the proof.

E. Proof of Corollary 3

For any $v \in \mathbb{R}^p$, the maximum absolute element of $\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1} v$ is bounded by

$$\Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1} v\Big\|_\infty \leq \Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\Big\|_\infty \|v\|_\infty.$$

Moreover, since the maximum group cardinality is m, we have

$$\Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1} v\Big\|^*_{\mathcal{G},\alpha} \leq \Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1} v\Big\|_\infty m^{1/\alpha^*}$$
$$\leq \Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\Big\|_\infty \|v\|_\infty m^{1/\alpha^*}$$

Now, we can derive the upper bound of $\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha}$:

$$\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha}$$
$$= \Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1} T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\frac{X^\top y}{n}\Big\|^*_{\mathcal{G},\alpha}$$
$$\leq \Big\|\Big[T_\nu\Big(\frac{X^\top X}{n}\Big)\Big]^{-1}\Big\|_\infty \Big\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top y}{n}\Big\|_\infty m^{1/\alpha^*}.$$

Finally, by the same reasoning and conditions as in Section D, we have, conditioned on the event $\mathcal{A}$,

$$\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha} \leq \frac{m^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right).$$

Therefore, given the choice of λ_n as in the statement, we have $\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha} \leq \lambda_n$ with high probability, and we can directly apply Theorem 2.

References

Bach, F. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Structured sparsity through convex optimization. *Statistical Science*, 27 (4):450–468, 2012.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. 37(4):1705–1732, 2009. Annals of Statistics.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

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

Donoho, D. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.

Fan, J. and Lv, J. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSSB)*, 70:849–911, 2008.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics*, 2007.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research (JMLR)*, 13:2107–2143, 2012.

Hsieh, C. J., Sustik, M., Dhillon, I., and Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. In *Neur. Info. Proc. Sys. (NIPS)*, 24, 2011.

Huang, J., Zhang, T., and Metaxas, D. Learning with structured sparsity. *Journal of Machine Learning Research (JMLR)*, 12: 3371–3412, 2011.

Jacob, L., Obozinski, G., and Vert, J. P. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pp. 433–440, 2009.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

Lozano, A. C., Swirszcz, G., and Abe, N. Group orthogonal matching pursuit for variable selection and prediction. In *Neur. Info. Proc. Sys (NIPS)*, 2009.

Mallat, S. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, December 1993.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34: 1436–1462, 2006.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37 (1):246–270, 2009.

Negahban, S. and Wainwright, M. J. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. Technical report, Department of Statistics, UC Berkeley, April 2009.

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

Obozinski, G., Wainwright, M. J., and Jordan, M. I. Union support recovery in high-dimensional multivariate regression. Technical report, Department of Statistics, UC Berkeley, August 2008.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5: 935–980, 2011.

Recht, B., Fazel, M., and Parrilo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, Vol 52(3):471–501, 2010.

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

Tropp, J. A., Gilbert, A. C., and Strauss, M. J. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

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

van de Geer, S., Bühlmann, P., and Zhou, S. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Varah, J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

Wille, A., Zimmermann, P., and Vranova, E. [and others]. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5, 2004.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

Zhang, J., Jeng, X. J., and Liu, H. Some two-step procedures for variable selection in high-dimensional linear regression. *Arxiv preprint arXiv:0810.1644*, 2008a.

Zhang, T. Sparse recovery with orthogonal matching pursuit under rip. Tech Report arXiv:1005.2249, May 2010.

Zhang, Z., Dolecek, L., Nikolic, B., Anantharam, V., and Wainwright, M. J. Lowering LDPC error floors by post-processing. In *Proc. IEEE GLOBECOM*, September 2008b.

Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

Zhao, P., Rocha, G., and Yu, B. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.