Electronic Journal of Statistics
Vol. 9 (2015) 608–642
ISSN: 1935-7524
DOI: 10.1214/15-EJS1013

On model selection consistency of regularized M-estimators

Jason D. Lee, Yuekai Sun

Institute for Computational and Mathematical Engineering
Stanford University
e-mail: jdl17@stanford.edu; yuekai@stanford.edu

and

Jonathan E. Taylor

Department of Statistics
Stanford University
e-mail: jonathan.taylor@stanford.edu

Abstract: Regularized M-estimators are used in diverse areas of science and engineering to fit high-dimensional models with some low-dimensional structure. Usually the low-dimensional structure is encoded by the presence of the (unknown) parameters in some low-dimensional model subspace. In such settings, it is desirable for estimates of the model parameters to be *model selection consistent*: the estimates also fall in the model subspace. We develop a general framework for establishing consistency and model selection consistency of regularized M-estimators and show how it applies to some special cases of interest in statistical learning. Our analysis identifies two key properties of regularized M-estimators, referred to as geometric decomposability and irrepresentability, that ensure the estimators are consistent and model selection consistent.

AMS 2000 subject classifications: 62F10.
Keywords and phrases: Regularized M-estimator, geometrically decomposable penalties, lasso, generalized lasso, group lasso, nuclear norm minimization.

Received October 2014.

1. Introduction

The principle of parsimony is used in many areas of science and engineering to promote "simple" models over more complex ones. In machine learning, signal processing, and high-dimensional statistics, this principle motivates the use of sparsity inducing penalties for model selection and signal recovery from incomplete/noisy measurements. Usually the "simplicity" of the model is encoded by the presence of the (unknown) parameters in some low-dimensional model subspace, and it is desirable for estimates of the parameters to fall in the model subspace. This notion of correctness is termed *model selection consistency.* In this work, we consider regularized M-estimators of the form

$$\operatorname*{minimize}_{\theta \in E} \ \ell(\theta) + \lambda\rho(\theta), \tag{1.1}$$

where ℓ is a convex, twice continuously differentiable loss, ρ is a penalty function, and $E \subseteq \mathbf{R}^p$ is a subspace. We identify two key properties of regularized M-estimators, referred to as geometric decomposability and irrepresentability, that ensure the estimators are consistent and model selection consistent. We also develop a general framework for analyzing the consistency and model selection consistency of M-estimators with geometrically decomposable penalties. When specialized to various statistical models, our framework yields some known and some new model selection consistency results.

The article is organized as follows: First, we review existing work on consistency and model selection consistency of regularized M-estimators. Then, in Section 2, we describe geometrically decomposable penalties. Section 3 is devoted to our main result and some discussion of its consequences. converse results on the necessity of the irrepresentable condition in Section 5. The final section, Section 4, is devoted to applications of our main result to various statistical models, including sparse regression and low-rank multivariate regression.

1.1. Notation

Given a set $S \subset \mathbf{R}^p$ and a point $x \in \mathbf{R}^p$, we use $P_S(x)$ to denote the *projector* of x on $\operatorname{span}(S)$:

$$P_S(x) = \arg\min_{y \in \operatorname{span}(S)} \tfrac{1}{2} \left\| x - y \right\|_2^2 .$$

Since $P_S(x)$ is a linear mapping, we write $P_S x = P_S(x)$, where $P_S \in \mathbf{R}^{p \times p}$ We use B_q to denote the q norm ball $\{x \in \mathbf{R}^p \mid \|x\|_q \leq 1\}$. For a semi-norm ρ, we use ρ^* to denote its *dual semi-norm:*

$$\rho(x)^* = \sup_{\rho(x) \leq 1} y^T x.$$

Finally, given a matrix $X \in \mathbf{R}^{p_1 \times p_1}$, we use $X^\dagger$ to denote its *Moore-Penrose pseudoinverse.*

1.2. Consistency of regularized M-estimators

A large body of work in high-dimensional statistics focuses on obtaining sufficient conditions for consistency of regularized M-estimation. A recurring theme in this avenue of research is the notion of *restricted strong convexity.* We refer to the past work section in Negahban et al. (2012) and Bühlmann and van de Geer (2011) for a comprehensive treatment of recent work on this topic.

Negahban et al. (2012) proposes a unified framework for establishing consistency and convergence rates for M-estimators with penalties ρ that are *decomposable* with respect to a pair of subspaces $M, \bar{M}$:

$$\rho(x+y) = \rho(x) + \rho(y), \text{for all } x \in M, y \in \bar{M}^\perp.$$

Many common penalties such as the lasso, group lasso, and nuclear norm are decomposable in this sense. Negahban et al. (2012) also develop a general notion of restricted strong convexity and prove a general result that establishes the consistency of M-estimators with decomposable penalties. Using their framework, they establish estimation consistency results for different statistical models including sparse and group sparse linear regression. Our results propose a unified framework for model selection consistency in a similar setting.

More recently, van de Geer (2012) proposes the notion of *weakly decomposability*. A penalty ρ is weakly decomposable if there is some norm $\rho_{\mathcal{S}^c}$ on $\mathbf{R}^{p-|\mathcal{S}|}$ such that ρ is superior to the sum of ρ and $\rho_{\mathcal{S}^c}$; i.e.

$$\rho(x) \geq \rho(x_{\mathcal{S}}) + \rho_{\mathcal{S}^c}(x_{\mathcal{S}^c}), \text{for all } x \in \mathbf{R}^p,$$

where $\mathcal{S} \subset [p]$ and $x_{\mathcal{S}} \in \mathbf{R}^{|\mathcal{S}|}, x_{\mathcal{S}^c} \in \mathbf{R}^{p-|\mathcal{S}|}$. Many common sparsity inducing penalties, including the ℓ_2/ℓ_1-norm (with possibly overlapping groups), are weakly decomposable. van de Geer (2012) shows oracle inequalities for the ℓ_1 penalty generalizes to weakly decomposable penalties.

In the parallel world of signal processing, there is a rich literature on constrained M-estimators of the form

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \; \|\theta\|_{\mathcal{A}} \text{ subject to } \theta \in C, \tag{1.2}$$

where $C \subset \mathbf{R}^p$ is a convex set. Candès and Recht (2012) proposed a unified analysis of (1.2) when $\|\cdot\|_{\mathcal{A}}$ is decomposable. By Candès and Recht (2012), Definition 1, $\|\cdot\|_{\mathcal{A}}$ is decomposable at $\theta^\star \in \mathbf{R}^p$ if $\partial\|\theta^\star\|_{\mathcal{A}}$ has the form

$$\partial \|\theta^\star\|_{\mathcal{A}} = \{z \in \mathbf{R}^p \mid P_T(z) = e, \rho(P_{T^\perp}(z))^* \leq 1\}$$

for some subspace $T \subset \mathbf{R}^p$ and a point $e \in T$. Above, P_T (resp. $P_{T^\perp}$) is the orthogonal projection onto T (resp. $T^\perp$). We emphasize the notion of decomposability by Candès and Recht (2012) is different from the notion by Negahban et al. (2012). Candès and Recht (2012) show exact recovery results for sparse/block-sparse vectors and low-rank matrices from random linear measurements depend upon the decomposability of the $\ell_1, \ell_2/\ell_1$, and nuclear norms. Bach (2010) show support recovery results for polyhedral norms arising from non-decreasing submodular functions including the ℓ_1 and ℓ_∞/ℓ_1 norms.

Recently, Chandrasekaran et al. (2012) proposed the notion of an *atomic norm*:

$$\|x\|_A = \inf\{t > 0 \mid x \in t\,\text{conv}(A)\} \text{ for a set of atoms } A.$$

They develop a general framework for deriving both exact (in the noise-free case) and robust (in the noisy case) recovery results from random Gaussian measurements by solving convex optimization problems of the form (1.2).

The model selection consistency of regularized M-estimators has also been extensively studied. The most commonly studied problems are

1. sparse regression (including generalized linear models): Zhao and Yu (2006); Bunea (2008); Wainwright (2009); Obozinski, Wainwright and Jordan (2011); Vaiter et al. (2013)
2. sparse covariance estimation and (more generally) structure learning: Meinshausen and Bühlmann (2006); Kolar et al. (2010); Ravikumar, Wainwright and Lafferty (2010); Jalali et al. (2011); Loh and Wainwright (2012).

In addition to restricted strong convexity, these results also rely upon the notion of *irrepresentability* originally proposed by Zhao and Yu (2006)[1]. Despite extensive work on this area, there is no general framework for establishing model selection consistency of commonly used M-estimators.

2. Geometrically decomposable penalties

Let $C \subset \mathbf{R}^p$ be a closed convex set. Then the *gauge function* on C is

$$\gamma_C(x) = \inf_x \{\lambda \in \mathbf{R}_+ \mid x \in \lambda C\},$$

and the *support function* on C is

$$h_C(x) = \sup_y \{y^T x \mid y \in C\}. \tag{2.1}$$

Both gauge support functions are sublinear and should be thought of as seminorms. If C is a norm ball, i.e. $C = \{x \mid \|x\| \leq 1\}$, then γ_C is the norm and h_C is the dual norm given by

$$\|y\|_* = \sup_x \{x^T y \mid \|x\| \leq 1\}.$$

The support function is a supremum of linear functions, hence the subdifferential consists of the linear functions that attain the supremum:

$$\partial h_C(x) = \{y \in C \mid y^T x = h_C(x)\}. \tag{2.2}$$

The support function (as a function of the convex set C) is also additive over Minkowski sums, i.e. if C and D are convex sets, then

$$h_{C+D}(x) = h_C(x) + h_D(x).$$

We use this property to express penalty functions as sums of support functions. E.g. if ρ is a norm and the dual norm ball can be expressed as a (Minkowski) sum of convex sets $C_1, \dots, C_k$, then ρ can be expressed as a sum of support functions:

$$\rho(x) = h_{C_1}(x) + \cdots + h_{C_k}(x).$$

If a penalty ρ can be expressed as

$$\rho(\theta) = h_A(\theta) + h_I(\theta) + h_{E^\perp}(\theta), \tag{2.3}$$

where $A, I \subset \mathbf{R}^p$ are closed convex sets and $E \subset \mathbf{R}^p$ is a subspace, then we say ρ is a *geometrically decomposable* penalty. This form is general; if ρ can be expressed as a sum of support functions, i.e.

$$\rho(\theta) = h_{C_1}(\theta) + \cdots + h_{C_k}(\theta),$$

then we can set A, I, and $E^\perp$ to be sums of the sets $C_1, \dots, C_k$ to express ρ in geometrically decomposable form (2.3). In many cases of interest, $A + I$ is a norm ball and $h_{A+I} = h_A + h_I$ is the dual norm. In our analysis, we further assume

[1] An equivalent notion, called *neighborhood stability*, was proposed by Meinshausen and Bühlmann (2006).

1. A, I are bounded.
2. I contains a relative neighborhood of the origin, i.e. $0 \in \text{relint}(I)$.

To allow for unregularized parameters, we do not assume $A + I$ contains a neighborhood of the origin. Thus ρ is not necessarily a norm. We summarize the form of geometrically decomposable penalties in a definition.

Definition 2.1. A regularizer is *geometrically decomposable* in terms of convex sets $A, I \subset \mathbf{R}^p$ and a subspace $E \subset \mathbf{R}^p$ if

$$\rho(\theta) = h_A(\theta) + h_I(\theta) + h_{E^\perp}(\theta).$$

We assume A, I are bounded and $0 \in \text{relint}(I)$.

The notation A, I should be as read as "active" and "inactive": $\text{span}(A)$ should contain the (unknown) parameter vector and $\text{span}(I)$ should contain deviations that we want to penalize. For example, if we know the sparsity pattern of the unknown parameter vector, then A should span the subspace of all vectors with the correct sparsity pattern.

The third term enforces a subspace constraint $\theta \in E$ because the support function of a subspace is the (convex) indicator function of the orthogonal complement:

$$h_{E^\perp}(\theta) = \mathbf{1}_E(\theta) = \begin{cases} 0 & \theta \in E \\ \infty & \text{otherwise.} \end{cases}$$

Such subspace constraints arise in many problems, either naturally (e.g. the constrained lasso by James, Paulson and Rusmevichientong (2012)) or after reformulation (e.g. group lasso with overlapping groups).

Before we state our theoretical results, we note that regularizers of the form $\rho(D\theta)$ for some $D \in \mathbf{R}^{m\times p}$ are geometrically decomposable, as long as ρ is geometrically decomposable. By the geometric decomposability of ρ,

$$\begin{aligned} \rho(D\theta) &= h_A(D\theta) + h_I(D\theta) + h_{E^\perp}(D\theta) \\ &= h_{D^T A}(\theta) + h_{D^T I}(\theta) + h_{D^T E^\perp}(\theta). \end{aligned}$$

In signal processing, regularizing with $\rho(D\theta)$ for some dictionary D is called *analysis regularization.* We give some examples of M-estimators with geometrically decomposable penalties in Section 3.

3. Main results

3.1. Problem setup

We begin with a description of the problem at hand. Let $X^{(n)} = \{X_1, \dots, X_n\}$ be n identically distributed observations of some random variable with marginal distribution $\mathbf{P}$. We seek to estimate some (unknown) parameters $\theta^\star \in M \subset \mathbf{R}^p$ of $\mathbf{P}$, where M is the *model subspace.* The model subspace is usually low-dimensional and captures the simple structure of the model. For example, M may be the subspace of vectors with a particular support or a subspace of low-rank matrices. We focus on the high-dimensional setting, i.e. when $n > p$.

Let ℓ be a convex and twice-continuously differentiable loss that assigns a cost $\ell(\theta)$ to any parameter $\theta \in E$. To estimate $\theta^\star$ from the data $X^{(n)}$, we solve the convex optimization problem

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}}\ \ell(\theta) + \lambda(h_A(\theta) + h_I(\theta) + h_{E^\perp}(\theta)), \tag{3.1}$$

where ρ is geometrically decomposable in terms of convex sets $A, I \subset \mathbf{R}^p$ and a subspace $E \subset \mathbf{R}^p$. The sets A, I, E are chosen such that $M = E \cap \operatorname{span}(I)^\perp$. Intuitively, $\operatorname{span}(I) \subset M^\perp$ contains deviations from M that we wish to kill. Many common regularized M-estimators possess the decomposable structure given by (3.1). To gain some intuition, we give three examples, beginning with sparse regression.

3.1.1. Sparse linear regression

Consider the linear model

$$y = X\theta + \epsilon, \tag{3.2}$$

where $X \in \mathbf{R}^{n \times p}$ is the design matrix, and $y \in \mathbf{R}^n$ are the responses. We assume the coefficients $\theta \in \mathbf{R}^p$ are sparse, i.e. most of the coefficients are zero. Let $\mathcal{S} \subset [\,p\,]$ be the support of θ, and $\mathcal{S}^c$ be the complementary subset of $[\,p\,]$. The model subspace is $\{\theta \in \mathbf{R}^p \mid \theta_{\mathcal{S}^c} = 0\}$.

The *lasso* by Tibshirani (1996) (also known as *basis pursuit denoising* by Chen, Donoho and Saunders (2001)) estimates $\theta^\star$ by the solution of:

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}}\ \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1. \tag{3.3}$$

The ℓ_1 norm is geometrically decomposable: $\|\theta\|_1 = h_{B_{\infty,\mathcal{S}}}(\theta) + h_{B_{\infty,\mathcal{S}^c}}(\theta)$, where $h_{B_{\infty,\mathcal{S}}}$ and $h_{B_{\infty,\mathcal{S}^c}}$ are support functions of the sets

$$B_{\infty,\mathcal{S}} = \left\{\theta \in \mathbf{R}^p \mid \|\theta\|_\infty \le 1, \theta_{\mathcal{S}^c} = 0\right\}$$
$$B_{\infty,\mathcal{S}^c} = \left\{\theta \in \mathbf{R}^p \mid \|\theta\|_\infty \le 1, \theta_{\mathcal{S}} = 0\right\}.$$

It is straightforward to check $\operatorname{span}(B_{\infty,\mathcal{S}^c})^\perp = M$. Thus the lasso possesses the structure given by (3.1). There is a well-developed theory of the lasso that says, under suitable assumptions on X, the lasso estimator is (consistent and) model selection consistent. In fact, under a stronger beta-min condition, the lasso is *sign consistent*. As we shall see, the aforementioned structure is the key to the performance of the lasso.

Given an estimate $\hat\theta$, there are various ways to assess its performance. We consider two notions: consistency and model selection consistency. An estimate $\hat\theta$ is *consistent* (in the ℓ_2 norm) if the estimation error in the ℓ_2 norm decays to zero in probability as sample size grows:

$$\left\|\hat\theta - \theta^\star\right\|_2 \xrightarrow{p} 0 \text{ as } n \to \infty.$$

An estimate is *model selection consistent* if $\hat\theta$ is in the *model subspace*:

$$\mathbf{Pr}(\hat\theta \in M) \to 1 \text{ as } n \to \infty. \tag{3.4}$$

3.2. The main result

Before we state our main result, we state our assumptions on the problem. Our two main assumptions are on the (sample) *Fisher information matrix*: $Q = \nabla^2 \ell(\theta^\star)$. The first is *restricted strong convexity (RSC)* and the second is *irrepresentability.*

Assumption 3.1 (Restricted strong convexity (RSC))**.** *Let $C \subset \mathbf{R}^p$ be some (a priori) known convex set containing $\theta^\star$. The loss function ℓ is RSC (on $C \cap M$) when*

$$\Delta^T \nabla^2 \ell(\theta)\Delta \geq m \left\|\Delta\right\|_2^2, \theta \in C \cap M, \Delta \in (C \cap M) - (C \cap M) \tag{3.5}$$

$$\|\nabla^2 \ell(\theta) - Q\|_2 \leq L \left\|\theta - \theta^\star\right\|_2, \theta \in C, \tag{3.6}$$

for some $m > 0$ and $L < \infty$.

The set C is usually taken to be a compact set (see Section 4.2 for an example). In these cases, restricted strong smoothness (3.6) holds by the continuity of $\nabla^2 \ell$. Similar notions of restricted strong convexity/smoothness are common in the literature on high-dimensional statistics. For example, the unified framework by Negahban et al. (2012) requires a (slightly stronger) notion of restricted strong convexity.

For a concrete example, we consider the sparse linear regression problem described in Section 3.1.1. When the rows of $X \in \mathbf{R}^{n \times p}$ are *i.i.d.* Gaussian random vectors with mean zero and covariance Σ, Raskutti, Wainwright and Yu (2010) showed there are constants $m_1, m_2 > 0$ such that

$$\frac{1}{n} \left\|X\Delta\right\|_2^2 \geq m_1 \left\|\Delta\right\|_2^2 - m_2 \frac{\log p}{n} \left\|\Delta\right\|_1^2 \text{ for any } \Delta \in \mathbf{R}^p$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Their result implies RSC over $\operatorname{span}(B_{\infty,\mathcal{S}})$ with constants $L = 0$ and $m = \frac{m_1}{2}$ as long as $n > 2\frac{m_2}{m_1}|\mathcal{S}| \log p$. Thus sparse regression with random Gaussian designs satisfy RSC, even when there are dependencies among the predictors. The result was extended to subgaussian designs by Rudelson and Zhou (2013), also allowing for dependencies among the predictors.

Assumption 3.2 (Irrepresentability)**.** *There is $\tau \in (0, 1)$ such that*

$$\sup\nolimits_{z \,\in\, \partial h_A(M)} V(P_{M^\perp}(QP_M(P_M Q P_M)^\dagger P_M z - z)) < 1 - \tau, \tag{3.7}$$

where $V(z) = \inf_y \{\gamma_I(y) + \mathbf{1}_{E^\perp}(z - y)\}$ and $\partial h_A(M) = \bigcup_{\theta \in M} \partial h_A(\theta)$.

As we shall see, V is a semi-norm: it measures the size of the component of z in I. In particular, $V(z) < 1$ implies

$$z = z_I + z_{E^\perp} \text{ for some } z_I \in \operatorname{relint}(I) \text{ and } z_{E^\perp} \in E^\perp.$$

Lemma 3.3. *V is finite and sublinear.*

To interpret the irrepresentable condition, consider again the sparse regression problem. Since $E = \mathbf{R}^p$ and $Q = \frac{1}{n}X^T X$, (3.7) simplifies to

$$\left\| X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T \right)^\dagger \operatorname{sign}(\theta_{\mathcal{S}}^\star) \right\|_\infty \leq 1 - \tau.$$

By the properties of support functions, $\partial h_A(M) \subset B_\infty$. Thus it is sufficient to assume

$$\left\| X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T \right)^\dagger \right\|_\infty \leq 1 - \alpha \text{ for some } \alpha \in (0,1). \tag{3.8}$$

The rows of $X_{\mathcal{S}^c}^T (X_{\mathcal{S}}^T)^\dagger$ are the regression coefficients of $x_j, j \in \mathcal{S}^c$ on $X_{\mathcal{S}}$. Thus (3.8) says the active predictors (columns of $X_{\mathcal{S}}$) are not overly well-aligned with the inactive predictors. Ideally, we would like the inactive predictors to be orthogonal to active predictors: $\|X_{\mathcal{S}^c}^T (X_{\mathcal{S}}^T)^\dagger\|_\infty = 0$. Unfortunately, orthogonality is impossible in the high-dimensional setting. The irrepresentable condition relaxes orthogonality to "near orthogonality".

As we shall see, the main result requires the regularization parameter λ to be larger than the "empirical process" part of the problem. Known results on the convergence rates of regularized M-estimators usually require $\lambda = \Omega(\rho^*(\nabla \ell(\theta^\star)))$. However, when ρ is not a norm (e.g. when there are unregularized parameters), $\rho^*(\nabla \ell(\theta^\star))$ is usually infinite. To allow for unregularized parameters, we relax the requirement to $\lambda = \Omega(\varrho^*(\nabla \ell(\theta^\star)))$ for some norm ϱ such that $\rho(\theta) \leq \varrho(\theta)$ for any $\theta \in \mathbf{R}^p$.

Before we state our main result, we describe some constant that appear in the result. Let B_2 be the 2-norm ball. We use κ_ρ (resp. $\kappa_\varrho, \kappa_{\varrho^*}$) to denote the *compatibility constant* between ρ (resp. ϱ, ϱ^*) and the ℓ_2-norm on M :

$$\kappa_\rho = \sup\nolimits_\theta \left\{ \rho(\theta) \mid \theta \in B_2 \cap M \right\}$$

(resp. $\kappa_\varrho, \kappa_{\varrho^*}$). Similarly, we use κ_{IC} to denote the compatibility constant between the irrepresentable term and ϱ^* :

$$\kappa_{\mathrm{IC}} = \sup\nolimits_{\varrho^*(z) \leq 1} V(P_{M^\perp}(Q P_M (P_M Q P_M)^\dagger P_M z - z)).$$

The constants κ_ρ and κ_{IC} are finite because ρ and ϱ^* are finite.

Theorem 3.4. *Assume ℓ and ρ satisfy RSC (on $C \cap M$) and irrepresentability (Assumptions 3.1 and 3.2). For any*

$$\tfrac{4\kappa_{\mathrm{IC}}}{\tau} \varrho^*(\nabla \ell(\theta^\star)) < \lambda < \tfrac{m^2}{2L} \left(2\kappa_\rho + \tfrac{\kappa_\varrho}{\kappa_{\mathrm{IC}}} \tfrac{\tau}{2} \right)^{-2} \tfrac{\tau}{\kappa_{\varrho^*} \kappa_{\mathrm{IC}}}, \tag{3.9}$$

the optimal solution to (3.1) *is unique,*

1. *consistent:* $\|\hat{\theta} - \theta^\star\|_2 \leq \frac{2}{m}(\kappa_\rho + \frac{\tau}{4} \frac{\kappa_\varrho}{\kappa_{\mathrm{IC}}})\lambda$,
2. *model selection consistent:* $\hat{\theta} \in M$.

Theorem 3.4 makes a *deterministic* statement about the optimal solution to (3.1). To use this result to derive consistency and model selection consistency results for a statistical model, we must first verify the loss and penalty satisfies restricted strong convexity/smoothness and irrepresentability. Then, we must select a penalty parameter that satisfies (3.9) (for some error norm). We know $\varrho^*(\nabla\ell(\theta^\star)) = O_p(\frac{1}{\sqrt{n}})$ for most problems of interest, so, for n large enough, there exist λ that satisfies (3.9).

Proof. The proof of Theorem 3.4 consists of three main steps:

1. Show the solution to a restricted problem (3.10) is unique and consistent (Lemma 3.5).
2. Establish a *primal-dual witness (PDW) condition* that ensures all solutions to the original problem are also solutions to the restricted problem (Lemma 3.6).
3. Construct a primal-dual pair for the original problem from the restricted primal-dual pair that satisfies the dual certificate condition.

Let $(\bar{\theta}, \bar{z}_A, \bar{z}_{M^\perp})$ be a primal-dual pair to the restricted problem:

$$\underset{\theta\in\mathbf{R}^p}{\text{minimize}}\ \ell(\theta) + \lambda(h_A(\theta) + h_{M^\perp}(\theta)). \tag{3.10}$$

The restricted primal-dual pair satisfies the first order optimality condition

$$\nabla\ell(\bar{\theta}) + \lambda\bar{z}_A + \lambda\bar{z}_{M^\perp} = 0 \tag{3.11}$$

$$\bar{z}_A \in \partial h_A(\bar{\theta}), \quad \bar{z}_{M^\perp} \in M^\perp. \tag{3.12}$$

First, we show the solution to the restricted problem is consistent.

Lemma 3.5. *Assume ℓ and ρ satisfy RSC (on $C \cap M$). For any $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\varrho^*(P_M\nabla\ell(\theta^\star))$, the optimal solution to the restricted problem (3.10) is unique and consistent: $\|\bar{\theta} - \theta^\star\|_2 \le \frac{2}{m}(\kappa_\rho + \frac{\tau}{4}\frac{\kappa_\varrho}{\kappa_{\mathrm{IC}}})\lambda$.*

Next, we establish the PDW condition that ensures all solutions to the original problem are also solutions to the restricted problem.

Lemma 3.6. *Suppose $\hat{\theta}$ is a primal solution to (3.1), and $\hat{z}_A, \hat{z}_I, \hat{z}_{E^\perp}$ are dual solutions; i.e. $(\hat{\theta}, \hat{z}_A, \hat{z}_I, \hat{z}_{E^\perp})$ satisfy*

$$\nabla\ell(\hat{\theta}) + \lambda(\hat{z}_A + \hat{z}_I + \hat{z}_{E^\perp}) = 0$$

$$\hat{z}_I \in \partial h_I(\hat{\theta}), \quad \hat{z}_A \in \partial h_A(\hat{\theta}), \quad \hat{z}_{E^\perp} \in E^\perp.$$

If $\hat{z}_I \in \mathrm{relint}(I)$, then all primal solutions to (3.1) satisfy $h_I(\theta) = 0$.

Finally, we use the restricted primal-dual pair to construct a feasible primal-dual pair for the original problem (3.1). The optimality conditions of the original problem are

$$\nabla\ell(\hat{\theta}) + \lambda(\hat{z}_A + \hat{z}_I + \hat{z}_{E^\perp}) = 0 \tag{3.13}$$

$$\hat{z}_I \in \partial h_I(\hat{\theta}), \quad \hat{z}_A \in \partial h_A(\hat{\theta}), \quad \hat{z}_{E^\perp} \in E^\perp. \tag{3.14}$$

Let

$$\begin{gathered}\hat{z}_I = \arg\min_z \gamma_I(z) + \mathbf{1}_{E^\perp}(\bar{z}_{M^\perp} - z) \\ \hat{z}_{E^\perp} = \bar{z}_{M^\perp} - \hat{z}_I.\end{gathered}$$

The pair $(\bar{\theta}, \bar{z}_A, \hat{z}_I, \hat{z}_{E^\perp})$ satisfies (3.14) by construction. Thus $\bar{\theta}$ is a solution to the original problem. To show $\bar{\theta}$ is the *unique* solution to the original problem, we show $\hat{z}_I$ is PDW feasible: $\hat{z}_I \in \operatorname{relint}(I)$.

The restricted primal-dual pair $(\bar{\theta}, \bar{z}_A, \bar{z}_{M^\perp})$ satisfies (3.12) and thus the zero reduced gradient condition:

$$P_M \nabla \ell(\bar{\theta}) + \lambda P_M \bar{z}_A = 0.$$

We Taylor expand $\nabla \ell$ around $\theta^\star$ (component-wise) to obtain

$$P_M W + P_M Q P_M (\bar{\theta} - \theta^\star) + P_M R + \lambda P_M \bar{z}_A = 0,$$

where $W = \nabla \ell(\theta^\star)$ and

$$R = \nabla \ell(\bar{\theta}) - \nabla \ell(\theta^\star) - Q(\bar{\theta} - \theta^\star).$$

Since $P_M Q P_M$ is invertible on M, we solve for $\bar{\theta}$ to obtain

$$\bar{\theta} = \theta^\star - (P_M Q P_M)^\dagger P_M (W + \lambda \bar{z}_A + R). \tag{3.15}$$

We Taylor expand $\nabla \ell$ in (3.12) around $\theta^\star$ to obtain

$$W + Q(\bar{\theta} - \theta^\star) + R + \lambda(\bar{z}_A + \bar{z}_{M^\perp}) = 0.$$

We substitute (3.15) into this expression to obtain

$$0 = W - Q(P_M Q P_M)^\dagger P_M (W + \lambda \bar{z}_A + R) + R + \lambda(\bar{z}_A + \bar{z}_{M^\perp}). \tag{3.16}$$

Rearranging, we obtain

$$\begin{aligned}\bar{z}_{M^\perp} &= \frac{1}{\lambda}\left(Q(P_M Q P_M)^\dagger P_M (W + \lambda \bar{z}_A + R) - W - R - \lambda \bar{z}_A)\right) \\ &= Q P_M (P_M Q P_M)^\dagger P_M \bar{z}_A - \bar{z}_A \\ &\quad + \frac{1}{\lambda}\left(Q P_M (P_M Q P_M)^\dagger P_M (W + R) - W + R\right).\end{aligned}$$

Finally, we take V's to obtain

$$V(\bar{z}_{M^\perp}) \le V(P_{M^\perp}(Q P_M (P_M Q P_M)^\dagger P_M \bar{z}_A - \bar{z}_A)) \tag{3.17}$$

$$+ \frac{1}{\lambda} V(P_{M^\perp}(Q P_M (P_M Q P_M)^\dagger W - W)) \tag{3.18}$$

$$+ \frac{1}{\lambda} V(P_{M^\perp}(Q P_M (P_M Q P_M)^\dagger P_M R - R)). \tag{3.19}$$

The irrepresentable condition (3.7) implies the first term is small:

$$V(P_{M^\perp}(QP_M(P_MQP_M)^\dagger P_M\bar{z}_A - \bar{z}_A)) \leq 1 - \tau.$$

Since V is a semi-norm on $M^\perp$, there is some κ_{IC} such that

$$V(P_{M^\perp}(QP_M(P_MQP_M)^\dagger W - W)) \leq \kappa_{\mathrm{IC}}\varrho^*(W).$$

We substitute these expressions into (3.19) to obtain

$$V(\bar{z}_{M^\perp}) \leq 1 - \tau + \kappa_{\mathrm{IC}}\left(\frac{\varrho^*(W)}{\lambda} + \frac{\varrho^*(R)}{\lambda}\right).$$

If we have $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\varrho^*(W)$, then $\frac{\kappa_{\mathrm{IC}}}{\lambda}\varrho^*(W) \leq \frac{\tau}{4}$ and

$$V(\bar{z}_{M^\perp}) < 1 - \tau + \frac{\tau}{4} + \frac{\kappa_{\mathrm{IC}}}{\lambda}\rho^*(R). \tag{3.20}$$

Lemma 3.7. *Assume ℓ and ρ satisfy RSC (over $C \cap M$). For any $\lambda < \frac{m^2}{2L}(2\kappa_\rho + \frac{\kappa_\varrho}{\kappa_{\mathrm{IC}}}\frac{\tau}{2})^{-2}\frac{\tau}{\kappa_{\varrho^*}\kappa_{\mathrm{IC}}}$, $\frac{\kappa_{\mathrm{IC}}}{\lambda}\varrho^*(R) < \frac{\tau}{4}$.*

We substitute this bound into (3.20) to obtain

$$V(\bar{z}_{M^\perp}) < 1 - \tau + \frac{\tau}{4} + \frac{\tau}{4} \leq 1 - \frac{\tau}{2} < 1.$$

Thus $\hat{z}_I$ is PDW feasible. By Lemma 3.6 and the uniquenss of the solution to the restricted problem, $\bar{\theta}$ is also the unique solution to the original problem. □

3.3. (Partial) converse results

Although the irrepresentable condition (3.7) seems cryptic and hard to verify, Zhao and Yu (2006) and Wainwright (2009) showed it is necessary for sign consistency of the lasso.[2] In this section, we give necessary conditions for an M-estimator with a geometrically decomposable penalty to be both consistent and model selection consistent.

Theorem 3.8. *Assume ℓ and ρ satisfy RSC (on $C \cap M$) and irrepresentability (Assumptions 3.1 and 3.2). Further, assume the optimal solution to (3.1) is unique, consistent, and model selection consistent, i.e.*

$$\hat{\theta} \in (\theta^\star + rB_2) \cap M.$$

We must have

$$\begin{aligned} P_{M^\perp}QP_M(P_MQP_M)^\dagger(W + \lambda\hat{z}_A + R) \\ \in P_{M^\perp}(W + R + \lambda(\hat{z}_A + I + E^\perp)) \\ \hat{z}_A \in \partial h_A((\theta^\star + rB_2) \cap M), \end{aligned}$$

where $W = \nabla\ell(\theta^\star)$ and $R = \nabla\ell(\hat{\theta}) - W - Q(\hat{\theta} - \theta^\star)$.

[2] Zhao and Yu (2006) and Wainwright (2009) refer to the (slightly) stronger condition (3.8) as irrepresentability. Thus their converse results are often summarized as irrepresentability is "almost" necessary for model selection consistency of the lasso.

Proof of Theorem 3.8. The proof proceeds like the proof of Theorem 3.4. The optimal solution to (3.1) satisfies

$$\nabla\ell(\hat{\theta}) + \lambda(\hat{z}_A + \hat{z}_I + \hat{z}_{E^\perp}) = 0 \tag{3.21}$$

$$\hat{z}_I \in \partial h_I(\hat{\theta}), \quad \hat{z}_A \in \partial h_A(\hat{\theta}), \quad \hat{z}_{E^\perp} \in E^\perp. \tag{3.22}$$

Since $\hat{\theta}$ is consistent and model selection consistent (by assumption), $\hat{\theta} \in (\theta^\star + rB_2) \cap M$. We solve for the error like we did to prove Theorem 3.4:

$$\hat{\theta} - \theta^\star = -(P_M Q P_M)^\dagger P_M (W + \lambda\hat{z}_A + R).$$

We plug in the expression for the error to (3.22) to obtain

$$0 = W - Q(P_M Q P_M)^\dagger P_M (W + \lambda\hat{z}_A + R) + R + \lambda(\hat{z}_A + \hat{z}_I + \hat{z}_{E^\perp}).$$

We project onto $M^\perp$ to obtain the desired result. □

Theorem 3.8 is a deterministic statement concerning the solution to (3.1). It says the (random) term

$$P_{M^\perp}(W + R) - P_{M^\perp} Q P_M (P_M Q P_M)^\dagger (W + R) \tag{3.23}$$

falls in the set

$$\begin{aligned} &P_{M^\perp}(\partial h_A((\theta^\star + rB_2) \cap M) + I + E^\perp) \\ &\quad - P_{M^\perp} Q P_M (P_M Q P_M)^\dagger \partial h_A((\theta^\star + rB_2) \cap M). \end{aligned} \tag{3.24}$$

To deduce the necessity of irrepresentability, we must show when irrepresentability is violated, the claims of Theorem 3.8 are invalid with positive probability. Although the distribution of (3.23) is generally hard to characterize, we do not need to completely characterize its distribution. As we shall see, showing the it is symmetric, i.e.

$$\mathbf{Pr}((3.23) \in B) = \mathbf{Pr}((3.23) \in -B) \text{ for any measurable set } B,$$

is enough to deduce the necessity of irrepresentability.

Corollary 3.9. *Assume ℓ and ρ satisfy RSC (over $C \cap M$) and A is a polyhedral set. Further, assume the distribution of (3.23) is symmetric, and*

$$\theta^\star \in \textstyle\bigcup_{\theta \in \mathrm{ext}(A)} \mathrm{relint}(N_A(\theta)).$$

When irrepresentability is violated—say

$$\inf\nolimits_{z \in \partial h_A(\theta^\star)} V(P_{M^\perp}(Q P_M (P_M Q P_M)^\dagger P_M z - z)) \geq 1,$$

$$\mathbf{Pr}(\hat{\theta} \in (\theta^\star + rB_2) \cap M) \leq \tfrac{1}{2}$$

for any r small enough such that $\theta^\star + rB_2 \subset \bigcup_{x \in \mathrm{ext}(A)} \mathrm{relint}(N_A(x))$.

Proof. Since $\theta^\star \in \bigcup_{x\in\mathrm{ext}(A)} \operatorname{relint}(N_A(x))$, $\partial h_A(\theta^\star)$ is a point. For any r small enough such that

$$\theta^\star + rB_2 \subset \bigcup_{x\in\mathrm{ext}(A)} \operatorname{relint}(N_A(x)),$$

$\partial h_A((\theta^\star + rB_2) \cap M)$ is also the point $\partial h_A(\theta^\star)$. Thus (3.24) is given by

$$P_{M^\perp}(\partial h_A(\theta^\star) + I + E^\perp) - P_{M^\perp}QP_M(P_MQP_M)^\dagger \partial h_A(\theta^\star). \tag{3.25}$$

When irrepresentability is violated, (3.25) is a convex set that does not contain a relative neighborhood of the origin. Thus there is a halfspace (through the origin) that contains (3.25). Since the distribution of (3.23) is symmetric, $\mathbf{Pr}((3.23) \in (3.24)) \leq \frac{1}{2}$. □

4. Examples

We use Theorem 3.4 to establish the consistency and model selection consistency of the lasso, the generalized lasso, and the regularized maximum likelihood estimator for exponential families in the high-dimensional setting. Our results are nonasymptotic, i.e. we obtain bounds in terms n and p that hold with high probability.

4.1. Sparse linear regression

We return to the sparse linear regression setup described in Section 3.1.1. The Fisher information is $\hat{\Sigma} = \frac{1}{n}X^TX$. We assume

1. RSC (over $\operatorname{span}(B_{\infty,\mathcal{S}})$ and let $C = \mathbf{R}^p$) and (3.8). Since ℓ is quadratic, it satisfies the smoothness condition with $L = 0$.
2. the components of ϵ are *i.i.d.* subgaussian random variables with mean zero and subgaussian norm σ.

The assumption (3.8) is a stronger condition than irrepresentability. It implies irrepresentability with $\tau = \alpha$:

$$\left\|X_{\mathcal{S}^c}^T\left(X_{\mathcal{S}}^T\right)^\dagger \operatorname{sign}(\theta_{\mathcal{S}}^\star)\right\|_\infty \leq \left\|X_{\mathcal{S}^c}^T\left(X_{\mathcal{S}}^T\right)^\dagger\right\|_\infty \left\|\operatorname{sign}(\theta^\star)_{\mathcal{S}}\right\|_\infty \leq 1-\alpha. \tag{4.1}$$

Corollary 4.1. *Assume $\hat{\Sigma}$ is RSC (on $\operatorname{span}(B_{\infty,\mathcal{S}})$) and (3.8). For $\lambda = \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{\log p}{n}}$, the lasso estimator is unique,*

1. *consistent: $\|\hat{\theta}-\theta^\star\|_2 \leq \frac{4}{m}(1+\frac{4(2-\alpha)}{\alpha})\sigma\sqrt{\frac{|\mathcal{S}|\log p}{n}}$,*
2. *model selection consistent: $\hat{\theta}_{\mathcal{S}^c} = 0$*

with probability at least $1-2p^{-1}$.

Further, if $\min_{a\in\mathcal{S}} |\theta_a^\star| > \frac{4}{m}(1+\frac{4(2-\alpha)}{\alpha})\sigma\sqrt{\frac{|\mathcal{S}|\log p}{n}}$, then the lasso estimator is also sign consistent: $\operatorname{sign}(\hat{\theta}_{\mathcal{S}}) = \operatorname{sign}(\theta_{\mathcal{S}}^\star)$.

Proof. Before we apply Theorem 3.4, we compute the constants $\kappa_\rho, \kappa_\varrho$ and κ_{IC}. Since the regularizer is finite (it's a norm), its dual semi-norm is finite. To keep things simple, we let $\varrho = \|\cdot\|_1$. The constant $\kappa_\rho = \kappa_\varrho$ is

$$\kappa_\rho = \sup_\theta \left\{\|\theta\|_1 \mid \theta \in B_2 \cap \operatorname{span}(B_{\infty,\mathcal{S}})\right\} = \sqrt{|\mathcal{S}|}.$$

Similarly, the constant κ_{IC} is given by

$$\begin{aligned}&\left\|P_{B_{\infty,\mathcal{S}^c}}(\hat{\Sigma}P_{B_{\infty,\mathcal{S}}}(P_{B_{\infty,\mathcal{S}}}\hat{\Sigma}P_{B_{\infty,\mathcal{S}}})^\dagger P_{B_{\infty,\mathcal{S}}}z - z)\right\|_\infty \\ &\quad\le \left\|X_{\mathcal{S}^c}^T\left(X_{\mathcal{S}}^T\right)^\dagger z_{\mathcal{S}}\right\|_\infty + \|z_{\mathcal{S}^c}\|_\infty \le (2-\alpha)\,\|z\|_\infty\end{aligned}$$

is at most $2-\alpha$.

To apply Theorem 3.4, we check $\lambda = \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{\log p}{n}}$ satisfies the assumptions. Since the loss function is quadratic, it satisfies the smoothness condition (3.6) with $L = 0$. Thus any $\lambda < \infty$ satisfies the upper bound in (3.9). We check our choice also satisfies the lower bound in (3.9). By Vershynin (2010), Proposition 5.10 and a union bound,

$$\mathbf{Pr}\left(\left\|\nabla\ell(\theta^\star)\right\|_\infty > t\right) = \mathbf{Pr}\Big(\left\|X^T\epsilon\right\|_\infty > nt\Big) \le 2\exp\Big(-\tfrac{nt^2}{2\sigma^2} + \log p\Big).$$

Thus

$$\begin{aligned}&\mathbf{Pr}\left(\frac{4(2-\alpha)}{\alpha}\left\|\nabla\ell(\theta^\star)\right\|_\infty > \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{\log p}{n}}\right) \\ &\quad\le 2\exp(-2\log p + \log p) = 2p^{-1}.\end{aligned}$$

Consequently, the claims of Theorem 3.4 are valid with probability at least $1-2p^{-1}$:

1. $\|\hat{\theta} - \theta^\star\|_2 \le \frac{2}{m}(1 + \frac{\alpha}{4(2-\alpha)})\sqrt{|\mathcal{S}|}\lambda = \frac{4}{m}(1 + \frac{4(2-\alpha)}{\alpha})\sigma\sqrt{\frac{|\mathcal{S}|\log p}{n}}$,
2. $\hat{\theta} \in \operatorname{span}(B_{\infty,\mathcal{S}}) = \{\theta \in \mathbf{R}^p \mid \theta_{\mathcal{S}^c} = 0\}$.

An easy consequence of (1) is $\|\hat{\theta}_a - \theta_a^\star\|_\infty \le \theta_{\min}$. Thus $\hat{\theta}$ is sign consistent: $\operatorname{sign}(\hat{\theta}_a) = \operatorname{sign}(\theta_a^\star)$ for any $a \in \mathcal{S}$ such that $|\theta_a^\star| > \theta_{\min}$. □

As we saw in Section 2, analysis regularizers of the form $\rho(D\theta)$ are geometrically decomposable. A prominent example of ℓ_1 analysis regularization is the *generalized lasso:*

$$\operatorname*{minimize}_{\theta\in\mathbf{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|D\theta\|_1\,. \tag{4.2}$$

The underlying (statistical) model is a straightforward modification of the linear model (3.2): we assume $D\theta^\star$ (instead of $\theta^\star$) is sparse. The sparsity of $D\theta$ usually translates to some desirable structural or geometric property of θ. We refer to Section 2 in Tibshirani and Taylor (2011) for some examples.

The model subspace is $\{\theta \in \mathbf{R}^p \mid (D\theta)_{\mathcal{S}^c} = 0\}$, where $\mathcal{S} \subset [\,m\,]$ is the support of $D\theta$. It's straightforward to check $\operatorname{span}(D^T B_{\infty,\mathcal{S}^c})^\perp = M$. Thus the generalized lasso possesses the structure given by (3.1). To study the model selection properties of the generalized lasso, we assume

$$\left\| D_{\mathcal{S}^c} X^T \left(D_{\mathcal{S}} X^T \right)^{\dagger} \operatorname{sign}(\theta^\star_{\mathcal{S}}) \right\|_\infty \le 1 - \tau \tag{4.3}$$

in lieu of (3.8). The assumption (4.3) is equivalent to irrepresentability. It is usually referred to as an *identifiability criterion (IC)*. Given IC (4.3), we derive the analog of Corollary (4.1) for the generalized lasso.

Corollary 4.2. *Assume $\hat{\Sigma}$ is RSC (on $\operatorname{span}(D^T B_{\infty,\mathcal{S}^c})^\perp$) and (4.3). For $\lambda = \frac{8\kappa_{\mathrm{IC}}}{\tau}\sigma\sqrt{\frac{\log p}{n}}$, the generalized lasso estimator is unique,*

1. *consistent: $\|\hat{\theta} - \theta^\star\|_2 \le \frac{4}{m}(\kappa_\varrho + \frac{4\kappa_{\mathrm{IC}}}{\tau}\kappa_\rho)\sigma\sqrt{\frac{\log p}{n}}$,*
2. *model selection consistent: $(D\theta)_{\mathcal{S}^c} = 0$*

with probability at least $1 - 2p^{-1}$.

Proof. Before we apply Theorem 3.4, we compute the constants κ_ρ and κ_ϱ. When D has a (nontrivial) null space, the regularizer is not a norm. To allow for the possibility, we let $\varrho = \|\cdot\|_1$ and set $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\|\nabla\ell(\theta^\star)\|_\infty$. The constants $\kappa_\rho, \kappa_\varrho$ are

$$\kappa_\rho = \sup_\theta \left\{ \|D\theta\|_1 \mid \theta \in B_2 \cap \operatorname{span}(D^T B_{\infty,\mathcal{S}^c})^\perp \right\}$$
$$\kappa_\varrho = \sup_\theta \left\{ \|\theta\|_1 \mid \theta \in B_2 \cap \operatorname{span}(D^T B_{\infty,\mathcal{S}^c})^\perp \right\}.$$

By an argument similar to the argument in the proof of Corollary 4.1, $\lambda = \frac{8\kappa_{\mathrm{IC}}}{\tau}\sigma\sqrt{\frac{\log p}{n}}$ satisfies the assumptions of Theorem 3.4 with probability at least $1 - 2p^{-1}$. Thus, with probability at least $1 - 2p^{-1}$,

1. $\|\hat{\theta} - \theta^\star\|_2 \le \frac{2}{m}(\kappa_\rho + \frac{\tau}{4}\frac{\kappa_\varrho}{\kappa_{\mathrm{IC}}})\lambda = \frac{2}{m}(2\kappa_\varrho + \frac{8\kappa_{\mathrm{IC}}}{\tau}\kappa_\rho)\sigma\sqrt{\frac{\log p}{n}}$,
2. $\hat{\theta} \in \operatorname{span}(D^T B_{\infty,\mathcal{S}^c})^\perp = \{\theta \in \mathbf{R}^p \mid (D\theta)_{\mathcal{S}^c} = 0\}$. □

4.2. Learning exponential families

We turn our attention to a problem with a non-quadratic loss function. Recall an exponential family is a distribution of the form

$$\mathbf{Pr}(x;\theta) = h(x)\exp\left(\theta^T\phi(x) - A(\theta)\right),$$

where θ are the *natural parameters*, $\phi(X) \in \mathbf{R}^p$ are *sufficient statistics*. We assume $\theta^\star$ is *group-sparse*, i.e. the components of $\theta^\star$ are organized in (possibly overlapping) groups and only a few groups are active. Let $\mathcal{G}$ be the collection of groups and $\mathcal{S}$ be the subset of active groups. The model subspace is $M = \{\theta \in \mathbf{R}^p \mid \theta_g = 0 \text{ for any } g \in \mathcal{S}^c\}$.

Given independent observations $X^{(n)} = \{X_1, \dots, X_n\}$, we seek to estimate $\theta^\star$ by the regularized *maximum likelihood estimator (MLE):*

$$\operatorname*{minimize}_{\theta\,\in\,E\,\subset\,\mathbf{R}^p} -\frac{1}{n}\sum_{i=1}^n \phi(x^{(i)})^T\theta + A(\theta) + \lambda\,\|\theta\|_{2/1}. \tag{4.4}$$

The ℓ_1/ℓ_2 norm is geometrically decomposable:

$$\|\theta\|_{2/1} = \sum_{g\in\mathcal{G}} \|\theta_g\|_2 = h_{B_{\infty/2,\mathcal{S}}}(\theta) + h_{B_{\infty/2,\mathcal{S}^c}}(\theta),$$

where $h_{B_{\infty/2,\mathcal{S}}}$ and $h_{B_{\infty/2,\mathcal{S}^c}}$ are support functions of the sets

$$\begin{aligned} B_{\infty/2,\mathcal{S}} &= \left\{\theta \in \mathbf{R}^p \mid \max_{g\in\mathcal{G}} \|\theta_g\|_2 \le 1, \theta_g = 0 \text{ for any } g \in \mathcal{S}^c\right\} \\ B_{\infty/2,\mathcal{S}^c} &= \left\{\theta \in \mathbf{R}^p \mid \max_{g\in\mathcal{G}} \|\theta_g\|_2 \le 1, \theta_g = 0 \text{ for any } g \in \mathcal{S}\right\}. \end{aligned}$$

It is easy to check $\operatorname{span}(B_{\infty/2,\mathcal{S}}) = M$. Thus (4.4) has the structure given by (3.1). The Fisher information is $Q = \nabla^2 A(\theta^\star)$. We assume Q satisfies RSC (over $\operatorname{span}(B_{\infty/2,\mathcal{S}})$) and irrepresentability.

First, we establish two auxiliary results: (i) a concentration result for W and (ii) the optimal solution to (4.4) is contained in some compact subset of the model subspace.

Lemma 4.3. *The random variable W satisfies*

$$\mathbf{Pr}\left(\left|\frac{\partial \ell}{\partial \theta_j}(\theta^\star)\right| > t\right) \le 2\exp\left(-cn\min\left(\frac{t^2}{\max_{g\in\mathcal{G}}|g|K^2}, \frac{t}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right)$$

$$\begin{aligned} &\mathbf{Pr}\left(\max_{g\in\mathcal{G}} \left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > t\right) \\ &\le 2\exp\left(\log|\mathcal{G}| - cn\min\left(\frac{t^2}{\max_{g\in\mathcal{G}}|g|K^2}, \frac{t}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right) \end{aligned}$$

for some absolute constant $c > 0$ and a constant K that is independent of n.

Lemma 4.4. *The optimal solution to (4.4) satisfies*

$$\begin{aligned} \|\hat\theta\|_{2/1} &\le \frac{1}{(\lambda - \|\phi^n - \phi^\star\|_{2,\infty})}\left(\lambda\|\theta^\star\|_{2/1} + \|\phi^n - \phi^\star\|_{2,\infty}\|\theta^\star\|_{2/1}\right) \\ A(\hat\theta) &\le \|\theta^\star\|_{2/1}\|\phi^n\|_{2,\infty} + \|\hat\theta\|_{2/1}\|\phi^n\|_{2,\infty} + A(\theta^\star) + \lambda\|\theta^\star\|_{2/1} \end{aligned}$$

where $\phi^n = \frac{1}{n}\sum_{i=1}^n \phi(x^{(i)})$ and $\phi^\star = \mathbf{E}_{\theta^\star}[\phi(X)]$.

We use these two results to establish the consistency and model selection consistency of the regularized MLE.

Corollary 4.5. *Suppose we are given samples $x^{(1)}, \dots, x^{(n)}$ drawn* i.i.d. *from a regular* exponential family *with unknown parameters $\theta^\star$, $\kappa_{\mathrm{IC}} \ge \tau$, and Assumption 3.2 is satisfied. Select*

$$\lambda = \frac{3\kappa_{\mathrm{IC}}}{\tau}\max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}}$$

and the sample size

$$n > \max\left(\frac{36}{c}\frac{\kappa_{\mathrm{IC}}^4}{\tau^4}\max_{g\in G}|g|K^2\log|\mathcal{G}|\frac{L^2}{m^4}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^4, \left(\frac{3}{2}\right)^2\frac{\log|\mathcal{G}|}{c}\right)$$

where $C := \{\theta \mid \|\theta\|_{2/1} \le 4\|\theta^\star\|_{2/1}$ and $A(\theta) \le R\}$, $c > 0$ is an absolute constant, and K is a constant independent of n defined in Lemma B.1. With probability at least $1 - 2|\mathcal{G}|^{-5/4}$, the optimal solution to (4.4) satisfies

1. $\|\hat{\theta} - \theta^\star\|_2 \le \frac{6}{m}\frac{\kappa_{\mathrm{IC}}}{\tau} \max_{g\in G} \sqrt{|g|}K(\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|})\sqrt{\frac{\log|\mathcal{G}|}{cn}}$
2. $\hat{\theta}_g = 0, g \in \mathcal{S}^c$.

Furthermore if we assume the beta-min condition

$$\left\|\theta_g^\star\right\|_2 > \frac{6}{m}\frac{\kappa_{\mathrm{IC}}}{\tau}\max_{g\in G}\sqrt{|g|}K\left(\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)\sqrt{\frac{\log|\mathcal{G}|}{cn}}$$

for all $g \in \mathcal{S}$, then all groups $g \in \mathcal{S}$ are correctly estimated as non-zero, $\|\hat{\theta}_g\|_2 > 0$.

5. Model selection properties of regularized M-estimators with weakly decomposable penalties

5.1. Background and problem setup

Geometric decomposability, although general, excludes some common regularizers. An important example is the nuclear norm:

$$\|\Theta\|_* = \textstyle\sum_{j=1}^r \sigma_j(\Theta),$$

where r is the rank of $\Theta \in \mathbf{R}^{p_1\times p_2}$ and $\sigma_j(\Theta), j = 1,\ldots,r$ are its singular values. The motivating example we have in mind is low-rank multivariate regression. Consider the (multivariate) generalization of the linear model:

$$Y = X\Theta^\star + W, \tag{5.1}$$

where the rows of $Y \in \mathbf{R}^{n\times p_2}$ are (multivariate) responses. We assume the matrix of coefficients $\Theta^\star \in \mathbf{R}^{p_1\times p_2}$ has rank $r \ll \min\{p_1, p_2\}$. Given observations $\{(x_i, y_i)\}_{i=1}^n$, a standard approach to estimating the unknow $\Theta^\star$ is *nuclear norm minimization:*

$$\underset{\Theta\in\mathbf{R}^{p_1\times p_2}}{\text{minimize}}\ \frac{1}{2n}\|Y - X\Theta\|_{\mathrm{F}}^2 + \lambda\|\Theta\|_*. \tag{5.2}$$

Bach (2008) showed that nuclear norm minimization is *rank consistent*, i.e.

$$\mathbf{Pr}\big(\mathrm{rank}(\hat{\Theta}) = r\big) \to 1 \text{ as } n \to \infty, \tag{5.3}$$

subject to irrepresentability. Although rank consistency does not fit into our notion of model selection consistency because the set of rank r matrices is not a subspace, our results may be used to derive a non-asymptotic form of Bach's rank consistency result.

To study the rank consistency of nuclear norm minimization, we consider an alternative notion of decomposability: *weak decomposability.*

Definition 5.1. A regularizer is *weakly decomposable* at $\theta^\star \in \mathbf{R}^p$ in terms of convex sets $A, I \subset \mathbf{R}^p$ if it is sublinear and

$$\partial\rho(\theta^\star) = \partial h_A(\theta^\star) + \partial h_I(\theta^\star).$$

We assume A, I are bounded and $0 \in \operatorname{relint}(I)$.

Weak decomposability is more general than geometric decomposability. However, the structure of the subdifferential of a weakly decomposable penalty at $\theta^\star$ is very similar to that of a geometrically decomposable penalty. Consequently, the directional derivative of ρ at $\theta^\star$ along Δ is geometrically decomposable:

$$\bar\rho(\theta^\star, \Delta) = h_{\partial h_A(\theta^\star)}(\Delta) + h_{\partial h_I(\theta^\star)}(\Delta).$$

As we shall see, the geometric decomposability of $\bar\rho(\theta^\star, \Delta)$ is the key to the model selection properties of weakly decomposable penalties.

The problem setup is similar to the setup in Section 3.1. To keep things simple, we focus on regularized least squares. Given n identically distributed observations of some random variable, we estimate some parameters $\theta^\star \in M \subset \mathbf{R}^p$ of its distribution by

$$\operatorname*{minimize}_{\theta\in\mathbf{R}^p} \frac{1}{2}\theta^T Q\theta - q^T\theta + \lambda\rho(\theta), \tag{5.4}$$

where ρ is weakly decomposable in terms of convex sets $A, I \subset \mathbf{R}^p$. The sets A, I are chosen such that $M = \operatorname{span}(I)^\perp$. As we shall see, the nuclear norm minimization problem has the form given by (5.4).

5.2. Dual consistency of regularized M-estimators

To study the model selection properties of (5.4), we compare the its optimal solution to the optimal solution to a linearized problem

$$\operatorname*{minimize}_{\theta\in\mathbf{R}^p} \frac{1}{2}\theta^T Q\theta - q^T\theta + \lambda(\rho(\theta^\star) + \bar\rho(\theta^\star, \theta - \theta^\star)). \tag{5.5}$$

Since the objective functions of (5.5) and (5.4) are similar, we expect the (optimal) solutions are close. Unfortunately, due to the lack of strong convexity, we cannot conclude the solutions are close. However, as we shall see, the dual solutions are close.

After a change of variables, the linearized problem is

$$\operatorname*{minimize}_{\Delta\in\mathbf{R}^p} \frac{1}{2}\Delta^T Q\Delta + (Q\theta^\star - q)^T\Delta + \lambda(h_{\partial h_A(\theta^\star)}(\Delta) + h_I(\Delta)). \tag{5.6}$$

We recognize (5.6) possesses the decomposable structure given by (3.1). By Theorem 3.4, a primal-dual pair $(\bar\Delta, \bar z_A, \bar z_I)$ that satisfies

$$\begin{gathered} Q(\theta^\star + \bar\Delta) - q + \lambda(\bar z_A + \bar z_I) = 0 \\ \bar z_A \in \partial h_A(\theta^\star), \quad \bar z_I \in I \end{gathered} \tag{5.7}$$

is unique. Further, $\bar\Delta$ is consistent, and $\bar z_I$ is PDW feasible. We summarize the properties of $(\bar\Delta, \bar z_A, \bar z_I)$ in a lemma.

Lemma 5.2. *Assume Q and $\bar{\rho}$ satisfy RSC (on $\operatorname{span}(I)^{\perp}$) and irrepresentability. For any $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\bar{\varrho}^*(Q\theta^\star - q)$, the unique primal-dual pair for (5.6) $(\bar{\Delta}, \bar{z}_A, \bar{z}_I)$ is*

1. *consistent: $\|\bar{\Delta}\|_2 \leq \frac{2}{m}(\kappa_{\bar{\rho}} + \frac{\tau}{4}\frac{\kappa_{\bar{\varrho}}}{\kappa_{\mathrm{IC}}})\lambda$.*
2. *PDW feasible: $\gamma_I(\bar{z}_I) \leq 1 - \frac{\tau}{2}$.*

The main result shows the dual solutions to (5.4) and (5.6) are close.

Theorem 5.3. *Assume Q and $\bar{\rho}$ satisfy RSC (on $\operatorname{span}(I)^{\perp}$) and irrepresentability. For any $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\bar{\varrho}^*(Q\theta^\star - q)$, the optimal dual solutions to (5.5) and (5.4) satisfy*

$$\|\bar{z}_A + \bar{z}_I - \hat{z}\|_2^2 \leq \frac{\|Q\|_2}{\lambda}\big(R(\bar{\Delta}) - R(\hat{\Delta})\big),$$

where $R(\Delta) = \rho(\theta^\star + \Delta) - \rho(\theta^\star) - \bar{\rho}(\theta^\star, \Delta)$.

Proof. After a change of variables, the original problem is

$$\operatorname*{minimize}_{\Delta \in \mathbf{R}^p} \ \frac{1}{2}\Delta^T Q \Delta + (Q\theta^\star - q)^T\Delta + \lambda\rho(\theta^\star + \Delta). \tag{5.8}$$

Its optimality conditions are

$$\begin{gathered} Q(\theta^\star + \hat{\Delta}) - \gamma + \lambda\hat{z} = 0 \\ \hat{z} \in \partial\rho(\theta^\star + \hat{\Delta}). \end{gathered} \tag{5.9}$$

Let $\bar{\Delta}$ and $\hat{\Delta}$ be the solutions to (5.6) and (5.8). By Fermat's rule, $\bar{z}_A + \bar{z}_I$ and $\hat{z}_A + \hat{z}_I$ are also the dual solutions to (5.5) and (5.4). We subtract (5.9) from (5.7) to obtain

$$Q(\hat{\Delta} - \bar{\Delta}) = \lambda(\bar{z}_A + \bar{z}_I - \hat{z}). \tag{5.10}$$

To complete the proof, we show $\|Q(\hat{\Delta} - \bar{\Delta})\|_2^2$ is small. By inspection of the optimality conditions (5.9) and (5.7), $\bar{\Delta}$ and $\hat{\Delta}$ are also the solutions to

$$\operatorname*{minimize}_{\Delta \in E} \bar{\Delta}^T Q\Delta + (Q\theta^\star - q)^T\Delta + \lambda\bar{\rho}(\theta^\star, \Delta)$$

$$\operatorname*{minimize}_{\Delta \in E} \hat{\Delta}^T Q\Delta + (Q\theta^\star - q)^T\Delta + \lambda(\rho(\theta^\star + \Delta) - \rho(\theta^\star)).$$

Since $\bar{\Delta}$ and $\hat{\Delta}$ are their respective optimal solutions, we know

$$\begin{aligned} &\bar{\Delta}^T Q\bar{\Delta} + (Q\theta^\star - q)^T\bar{\Delta} + \lambda\bar{\rho}(\theta^\star, \bar{\Delta}) \\ &\quad \leq \bar{\Delta}^T Q\hat{\Delta} + (Q\theta^\star - q)^T\hat{\Delta} + \lambda\bar{\rho}(\theta^\star, \hat{\Delta}), \\ &\hat{\Delta}^T Q\hat{\Delta} + (Q\theta^\star - q)^T\hat{\Delta} + \lambda(\rho(\theta^\star + \hat{\Delta}) - \rho(\theta^\star)) \\ &\quad \leq \hat{\Delta}^T Q\bar{\Delta} + (Q\theta^\star - q)^T\bar{\Delta} + \lambda\left(\rho(\theta^\star + \bar{\Delta}) - \rho(\theta^\star)\right). \end{aligned}$$

We add the inequalities and rearrange to obtain

$$(\bar{\Delta} - \hat{\Delta})^T Q(\bar{\Delta} - \hat{\Delta}) = \|\Delta\|_Q^2 \leq \lambda\big(R(\bar{\Delta}) - R(\hat{\Delta})\big),$$

where $R(\Delta) = \rho(\theta^\star + \Delta) - \rho(\theta^\star) - \bar{\rho}(\theta^\star, \Delta)$. Since $\|Q\Delta\|_2^2 \leq \|Q\|_2\|\Delta\|_Q^2$,

$$\big\|Q(\hat{\Delta} - \bar{\Delta})\big\|_2^2 \leq \|Q\|_2\big\|\hat{\Delta} - \bar{\Delta}\big\|_Q^2 \leq \|Q\|_2\,\lambda\big(R(\bar{\Delta}) - R(\hat{\Delta})\big).$$

We plug in (5.10) to reach the stated conclusion. □

5.3. Rank consistency of low-rank multivariate regression

We return to the low-rank multivariate regression problem described in Section 5.1. The nuclear norm is weakly decomposable. Let $\Theta^\star = U\Sigma V^T$ be the (full) SVD of $\Theta^\star$ and define the sets

$$\begin{aligned} A &= \left\{\Theta \in B_{\mathrm{sp}} \subset \mathbf{R}^{p_1\times p_2} \mid \Theta = U_r D V_r^T \text{ for some diagonal } D\right\} \\ I &= \left\{\Theta \in B_{\mathrm{sp}} \subset \mathbf{R}^{p_1\times p_2} \mid \Theta = U_{p_1-r} D V_{p_2-r}^T \text{ for some diagonal } D\right\}, \end{aligned}$$

where U_r, U_{p_1-r} (resp. V_r, V_{p_2-r}) are submatrices of U (resp. V) consisting of the top r and bottom $p_1 - r$ left (resp. $p_2 - r$ right) singular vectors of $\Theta^\star$. It's straightforward to check the nuclear norm is weakly decomposable at $\Theta^\star$ in terms of A, I. Since $A + I \subset B_{\mathrm{sp}}$,

$$\|\Theta\|_* = h_{B_{\mathrm{sp}}}(\Theta) \geq h_A(\Theta) + h_I(\Theta).$$

Before we delve into the rank consistency of low-rank multivariate regression, we state our assumptions on the problem. To keep notation manageable, we adopt operator theoretic notation. Let $\vec{X} \in \mathbf{R}^{p_1 p_2}$ be the vectorized form of $X \in \mathbf{R}^{p_1\times p_2}$. In operator notation, the model is

$$\vec{Y} = \mathcal{X}(\Theta^\star) + \vec{W}, \tag{5.11}$$

where $\mathcal{X} : \mathbf{R}^{p_1\times p_2} \to \mathbf{R}^n$ is a linear operator. Since $\mathcal{X}$ is linear, we abuse notation by writing $\mathcal{X}\Theta = \mathcal{X}(\Theta)$. The Fisher information $Q : \mathbf{R}^{p_1\times p_2} \to \mathbf{R}^{p_1\times p_2}$ is given by $\frac{1}{n}\mathcal{X}^*\mathcal{X}$. We assume

1. RSC and
$$\sup_{Z\in B_{\mathrm{sp}}} \left\|U_{p_1-r}^T \left[P_I Q P_{I^\perp} (P_{I^\perp} Q P_{I^\perp})^\dagger Z\right] V_{p_2-r}\right\|_{\mathrm{sp}} \leq 1-\alpha, \tag{5.12}$$
where $P_I : \mathbf{R}^{p_1\times p_2} \to \mathbf{R}^{p_1\times p_2}$ (resp. $P_{I^\perp}$) is the projector onto $\mathrm{span}(I)$ (resp. $\mathrm{span}(I)^\perp$).
2. the entries of W are *i.i.d.* subgaussian random variables with mean zero and subgaussian norm σ.

The assumption (5.12) is stronger than irrepresentability. It implies irrepresentability with $\tau = \alpha$:

$$\begin{aligned} &\left\|U_{p_1-r}^T \left[P_I\left(Q P_{I^\perp}(P_{I^\perp} Q P_{I^\perp})^\dagger U_r V_r^T - U_r V_r^T\right)\right] V_{p_2-r}\right\|_{\mathrm{sp}} \\ &\quad = \left\|U_{p_1-r}^T \left[P_I\left(Q P_{I^\perp}(P_{I^\perp} Q P_{I^\perp})^\dagger U_r V_r^T\right)\right] V_{p_2-r}\right\|_{\mathrm{sp}} \\ &\quad \leq \sup_{Z\in B_{\mathrm{sp}}} \left\|U_{p_1-r}^T \left[P_I Q P_{I^\perp}(P_{I^\perp} Q P_{I^\perp})^\dagger Z\right] V_{p_2-r}\right\|_{\mathrm{sp}}. \end{aligned} \tag{5.13}$$

We make the stronger assumption to obtain an explicit expression for the constant κ_{IC} (in terms of the constant α).

The final ingredient we require is a "Taylor's theorem" for the nuclear norm that says the nuclear norm is well-approximated by its linearization.

Lemma 5.4. *Let s_r be smallest nonzero singular value of $\Theta^\star$. For any $\Delta \in \operatorname{span}(I)^\perp, \|\Delta\|_{\mathrm{sp}} < \frac{s_r}{2}$, we have*

$$\|\Theta^\star + \Delta\|_* - \|\Theta^\star\|_* - \operatorname{tr}\big(U_r^T \Delta V_r\big) \le \frac{4}{3s_r}\|\Delta\|_{\mathrm{F}}^2 .$$

We put the pieces togther to conclude low-rank multivariate regression is rank consistent.

Corollary 5.5. *Assume Q is RSC (over $\operatorname{span}(I)^\perp$) and (5.12). For $\lambda = \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{p_1+p_2}{n}}$, the optimal solution to (5.2) is unqiue and rank consistent when*

$$n > \max\left\{\frac{128^2}{9s_r^2}\frac{M^2}{m^4}\frac{(\sqrt{2}+\alpha')^4}{\alpha^4\alpha'^2}r^2, \frac{16}{m^2}(\sqrt{2}+\alpha')^2 r\right\}\sigma^2(p_1+p_2)$$

with probability at least $1 - c_1 e^{-c_2(p_1+p_2)}$. The constants M and α' are given by $\sup_{\|\Delta\|_{\mathrm{F}}\le 1}\|Q\Delta\|_{\mathrm{F}}$ and $\frac{4(2-\alpha)}{\alpha}$.

Proof. To show $\hat{\Theta}$ has rank at most r, it suffices to show the optimal dual solution $\hat{U}\hat{V}^T$ has no more than r unit singular values. At a high level, the proof consists of three steps:

1. Show the unique (feasible) primal-dual pair to a linearized problem ($\bar{\Delta}$, $U_r V_r^T, \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T$) is consistent and PDW feasible (Lemma 5.6).
2. Invoke Theorem 5.3 to show $\hat{U}\hat{V}^T$ is close to the optimal dual solution to the linearized problem $U_r V_r^T + \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T$. Since $\bar{U}_{p_1-r}\bar{V}_{p_2-r}^T$ is PDW feasible, its singular values are bounded away from one.
3. Apply a singular value perturbation result to conclude $\hat{U}\hat{V}^T$ has (no more than) r unit singular values.

Consider the linearized problem

$$\underset{\Delta\in\mathbf{R}^{p_1\times p_2}}{\text{minimize}}\ \frac{1}{2n}\|Y - X(\Theta^\star+\Delta)\|_{\mathrm{F}}^2 + \lambda\left(\operatorname{tr}\big(U_r^T\Delta V_r\big) + \|U_{p_1-r}^T\Delta V_{p_2-r}\|_*\right). \quad (5.14)$$

We apply Lemma 5.2 to deduce a primal-dual pair $(\bar{\Delta}, U_r V_r^T, \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T)$ that satisfies

$$\hat{\Sigma}(\Theta^\star + \bar{\Delta}) - q + \lambda(U_r V^T + \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T) = 0$$
$$\bar{U}_{p_1-r}\bar{V}_{p_2-r}^T \in I$$

is unique, consistent, and PDW feasible.

Lemma 5.6. *Assume the linearized problem (5.14) satisfies RSC (over $\operatorname{span}(I)^\perp$) and (5.12). For $\lambda = \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{p_1+p_2}{n}}$, the unique primal-dual pair for (5.14) $(\bar{\Delta}, U_r V_r^T, \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T)$ is*

1. *consistent: $\|\bar{\Delta}\|_{\mathrm{F}} \le \frac{4}{m}(\sqrt{2} + \frac{4(2-\alpha)}{\alpha})\sigma\sqrt{\frac{r(p_1+p_2)}{n}}$.*
2. *PDW feasible: $\|\bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\|_{\mathrm{sp}} \le 1 - \frac{\tau}{2}$.*

By Theorem 5.3 (and the convexity of the nuclear norm),

$$\begin{aligned}&\big\|\hat{U}\hat{V}^T - U_rV_r^T - \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\big\|_{\mathrm{sp}}^2\\ &\quad\le \|\hat{U}\hat{V}^T - U_rV_r^T - \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\|_{\mathrm{F}}^2\\ &\quad\le \frac{M}{\lambda}\big(R(\bar{\Delta}) - R(\hat{\Delta})\big) \le \frac{M}{\lambda}R(\bar{\Delta}),\end{aligned}$$

where $M = \sup_{\|\Delta\|_{\mathrm{F}}\le 1}\|Q\Delta\|_{\mathrm{F}}$. Since $\bar{\Delta}\in\mathrm{span}(I)^\perp$, we may apply Lemma 5.4 to obtain

$$\big\|\hat{U}\hat{V}^T - U_rV_r^T - \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\big\|_{\mathrm{sp}}^2 \le \frac{4}{3s_r}\frac{M}{\lambda}\|\bar{\Delta}\|_{\mathrm{F}}^2$$

as long as $\|\bar{\Delta}\|_{\mathrm{sp}} \le \frac{s_r}{2}$. By the consistency of the linearized problem

$$\|\bar{\Delta}\|_{\mathrm{sp}} \le \|\bar{\Delta}\|_{\mathrm{F}} \le \frac{4}{m}(\sqrt{2}+\alpha')\sigma\sqrt{\frac{r(p_1+p_2)}{n}},$$

where $\alpha' = \frac{4(2-\alpha)}{\alpha}$. We put the pieces together to obtain

$$\begin{aligned}&\big\|\hat{U}\hat{V}^T - U_rV_r^T - \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\big\|_{\mathrm{sp}}^2\\ &\quad\le \frac{32}{3s_r}\frac{M}{m^2}\frac{(\sqrt{2}+\alpha')^2}{\alpha'}\sigma r\sqrt{\frac{(p_1+p_2)}{n}},\end{aligned} \tag{5.15}$$

when $n > \frac{16}{m^2}\frac{\sigma^2}{s_r^2}(\sqrt{2}+\alpha')^2 r(p_1+p_2)$.

By Lemma 5.2, $\bar{U}_{p_1-r}\bar{V}_{p_2-r}^T$ is PDW feasible. Thus it has at most r unit singular values. Its $\min\{p_1,p_2\}-r$ remaining singular values are smaller than $1-\frac{\alpha}{2}$. By Weyl's inequality, it suffices to ensure

$$\big\|\hat{U}\hat{V}^T - U_rV_r^T - \bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\big\|_{\mathrm{sp}} \le \frac{\alpha}{2}, \tag{5.16}$$

to ensure $\bar{U}\bar{V}^T$ has no more than r unit singular values. We combine (5.15) and (5.16) to deduce the requirement on n. □

To our knowledge, Theorem 5.5 is the first non-asymptotic rank consistency result for the multivariate regression problem. Further, the proof technique generalizes readily to M-estimators with other loss functions and regularizers. In Chandrasekaran, Parrilo and Willsky (2012), the authors demonstrated rank and sign consistency in the problem of graphical model estimation with latent variables. For the multivariate regression problem, Negahban et al. (2011) showed an operator norm consistency bound. Operator norm consistency by itself does not give rank consistency, but hard-thresholding the singular values of $\hat{\Theta}$ at the appropriate level does give rank consistency. In contrast, Corollary 5.5 is a statement regarding $\hat{\Theta}$, the regularized M-estimator 5.2 without any post-processing.

6. Computational experiments

We show some consequences of Corollary 4.5 with experiments on two models from structure learning of networks that are motivated by bioinformatics

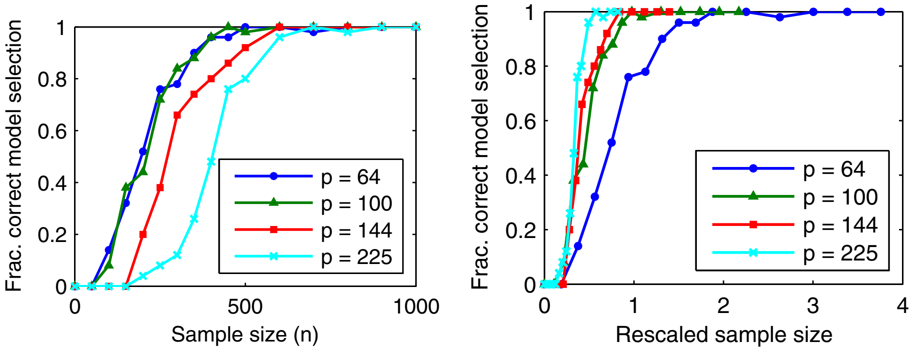


Fig 1. *Fraction of correct model selection versus sample size n and rescaled sample size $n/((\max_{g\in\mathcal{G}}|g|)\log|\mathcal{G}|)$ with the grouped graphical lasso. Each point represents the fraction of 100 trials when the grouped graphical lasso correctly estimated the true group structure.*

applications. We select λ proportional to $\sqrt{(\max_{g\in\mathcal{G}}|g|)\frac{\log|\mathcal{G}|}{n}}$ and use a proximal Newton-type method Lee, Sun and Saunders (2009) to solve the likelihood maximization problem.

6.1. The graphical lasso

Suppose we are given samples drawn *i.i.d.* from a normal distribution, and we seek to estimate the inverse covariance matrix. If $p > n$, we cannot simply invert the sample covariance matrix $\hat{\Sigma}$ to estimate $\Theta^\star$. However, we can obtain a maximum likelihood estimate of $\Theta^\star$:

$$\underset{\Theta}{\text{minimize}}\ \operatorname{tr}\left(\hat{\Sigma}\Theta\right) - \log\det(\Theta) + \lambda\sum_{s,t\in\mathcal{G}}\|\Theta_{st}\|_2 . \tag{6.1}$$

The group lasso penalty to promotes block sparse inverse covariance matrices, and λ trades-off goodness-of-fit and group sparsity. This estimator is a group sparse variant of the *graphical lasso* Friedman, Hastie and Tibshirani (2008).

We estimate the probability of correct model selection using the fraction of 100 trials when the graphical lasso correctly estimates the true group structure. Figure 1 shows the fraction of correct group structure selection versus the sample size n for four graphs consisting of 64, 100, 144, and 225 nodes. In these experiments, we varied the sample size n from 100 to 1000.

The fraction of correct model selection is small for small sample sizes but grows to one as the sample size increases. Intuitively, more samples are required to learn a larger model, hence the curves for larger graphs are to the right of curves for smaller graphs. If we plot these curves with the x-axis rescaled by $1/((\max_{g\in\mathcal{G}}|g|)\log|\mathcal{G}|)$, then the curves align. This is consistent with Corollary 4.5 that say the effective sample size scales logarithmically with $|\mathcal{G}|$.

6.2. Learning mixed graphical models

The pairwise mixed graphical model was developed to model data that contain both categorical and continuous features Lee and Hastie (2012); Cheng, Levina and Zhu (2013) e.g., two features about a person are weight (continuous) and gender (categorical). The model is a natural pairwise extension of the Gaussian MRF and a pairwise discrete MRF:

$$\mathbf{Pr}(x, y; (\beta, \theta, \gamma)) \propto \exp\Big(\sum_{s,t} -\tfrac{1}{2}\beta_{st}x_s x_t + \sum_{s,j} \theta_{sj}(y_j)x_s + \sum_{j,r} \gamma_{rj}(y_r, y_j)\Big). \quad (6.2)$$

x_s, $s = 1, \dots, p$ and y_j, $j = 1, \dots, q$'s are continuous and discrete variables and $\beta_{st}, \theta_{sj}, \gamma_{rj}$ are continuous-continuous, continuous-discrete, and discrete-discrete edge potentials. We seek maximum likelihood and pseudolikelihood estimates of the parameters (β, θ, γ)

$$\underset{(\beta,\theta,\gamma)}{\text{minimize}} \ -\ell^{(n)}((\beta, \theta, \gamma)) + \lambda\rho((\beta, \theta, \gamma)). \quad (6.3)$$

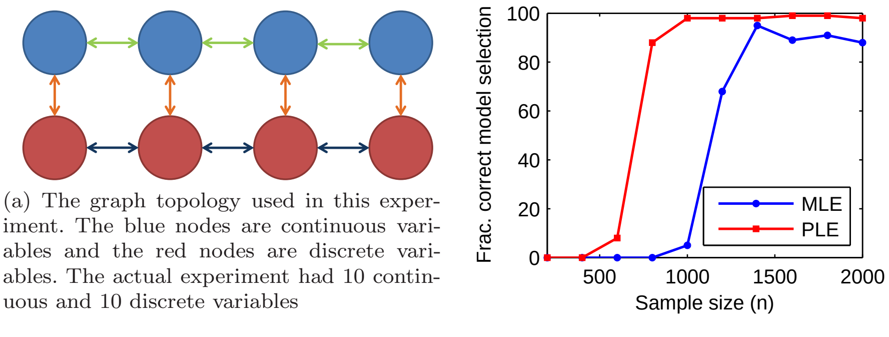
ρ is the group lasso penalty:

$$\rho((\beta, \theta, \gamma)) = \sum_{s,t} |\beta_{st}| + \sum_{s,j} \|\theta_{sj}\|_2 + \sum_{j,r} \|\gamma_{rj}\|_F .$$

To make sure the model is identifiable, we enforce linear constraints on γ_{rj}:

$$\textstyle\sum_{x_r,x_j} \gamma_{rj}(x_r, x_j) = 0,\ j, r = 1, \dots, q.$$

We estimate the probability of correct model selection using the fraction of 100 trials when (6.3) correctly estimates the true group structure. Figure 2 shows the fraction of correct group structure selection versus the sample size n. In these experiments, we varied the sample size from 300 to 2000.



(a) The graph topology used in this experiment. The blue nodes are continuous variables and the red nodes are discrete variables. The actual experiment had 10 continuous and 10 discrete variables

Fig 2. *Fraction of correct model selection versus sample size n of the penalized MLE and PLE on a mixed graphical model. Each point represents the fraction of 100 trials when the grouped graphical lasso correctly estimated the true group structure.*

The fraction of correct model selection is small for small sample sizes but grows with the sample size. For the penalized PLE, the fraction grows to one, but, for the penalized MLE, the fraction plateaus around 0.9. This can be explained by the penalized MLE violating the irrepresentable condition. We refer to Section 3.1.1 in Ravikumar et al. (2011) for a similar example where the the irrepresentable condition holds for a neighborhood-selection estimator but fails for the penalized MLE.

7. Conclusion

We proposed the notion of geometric decomposablility and showed it is key to the model selection properties of regularized M-estimators. Our notion of decomposability builds on the notions by Candès and Recht (2012) and van de Geer (2012) and readily admits a notion of irrepresentability.

We also developed a general framework for establishing consistency and model selection consistency of regularized M-estimators. Our main result (Theorem 3.4 gives deterministic conditions on the problem that guarantee consistency and model selection consistency. We combined our main result with probabilistic analysis to study the model selection properties of the lasso, generalized lasso, and nuclear norm minimization. To our knowledge the non-asymptotic result on rank-consistency of nuclear norm minimization is novel.

Acknowledgements

We thank Trevor Hastie and three anonymous reviewers for their insightful comments. J. Lee was supported by a National Defense Science and Engineering Graduate Fellowship (NDSEG), a NSF Graduate Fellowship, and a Stanford Graduate Fellowship. Y. Sun was supported by the NIH, grant 1U01GM102098-01. J.E. Taylor was supported by the NSF, grant DMS 1208857, and by the AFOSR, grant 113039.

Appendix A: Proof of lemmas in Section 3

Proof of Lemma 3.3. V is the infimal convolution of $\gamma_I = h_{I^\circ}$ and $\mathbf{1}_E = h_{E^\perp}$. By the properties of support functions, $V = h_{E \cap I^\circ}$. Since I° is bounded and support functions of bounded sets are finite and sublinear, V is finite and sublinear. □

Proof of Lemma 3.5. Since $\bar{\theta}$ solve the restricted problem, we have

$$\ell(\bar{\theta}) + \lambda h_A(\bar{\theta}) \leq \ell(\theta^\star) + \lambda h_A(\theta^\star).$$

Since $\hat{\theta} \in C$ and the objective is strongly convex over C, $\hat{\theta}$ is the unique solution to (3.1). By Assumption 3.1, we have

$$W^T P_M(\bar{\theta} - \theta^\star) + \frac{m}{2}\|\bar{\theta} - \theta^\star\|_2^2 + \lambda(\rho(\bar{\theta}) - \rho(\theta^\star)) \leq 0,$$

where $W = \nabla \ell(\theta^\star)$. We take norms to obtain

$$0 \geq -\varrho^*(P_M W)\varrho\big(\bar{\theta} - \theta^\star\big) + \frac{m}{2}\big\|\bar{\theta} - \theta^\star\big\|_2^2 - \lambda\rho(\bar{\theta} - \theta^\star) \tag{A.1}$$

$$\geq -\kappa_\varrho \varrho^*(P_M W)\big\|\bar{\theta} - \theta^\star\big\|_2 + \frac{m}{2}\big\|\bar{\theta} - \theta^\star\big\|_2^2 - \lambda\rho(\bar{\theta} - \theta^\star). \tag{A.2}$$

Since $\hat{\theta} - \theta^\star \in M$, we have

$$h_A(\bar{\theta} - \theta^\star) = \rho(\bar{\theta} - \theta^\star) \leq \kappa_\rho \big\|\bar{\theta} - \theta^\star\big\|_2.$$

We substitute this bound into (A.2) to obtain

$$0 \geq -\kappa_\varrho \varrho^*(P_M W)\big\|\bar{\theta} - \theta^\star\big\|_2 + \frac{m}{2}\big\|\bar{\theta} - \theta^\star\big\|_2^2 - \kappa_\rho \lambda \big\|\bar{\theta} - \theta^\star\big\|_2.$$

This means

$$\big\|\bar{\theta} - \theta^\star\big\|_2 \leq \frac{2}{m}\left(\kappa_\varrho \varrho^*(P_M W) + \kappa_\rho \lambda\right).$$

Plugging in the choice of $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\varrho^*(P_M W)$ gives our conclusion. □

Proof of Proposition 3.6. Suppose there are two primal dual solution pairs, $(\theta_1, z_{A,1}, z_{I,1}, z_{E^\perp,1})$ and $(\theta_2, z_{A,2}, z_{I,2}, z_{E^\perp,2})$, i.e.

$$\nabla \ell(\theta_1) + \lambda(z_{A,1} + z_{I,1} + z_{E^\perp,1}) = 0 \tag{A.3}$$

$$\nabla \ell(\theta_2) + \lambda(z_{A,2} + z_{I,2} + z_{E^\perp,2}) = 0. \tag{A.4}$$

Since the original problem (3.1) is convex, the optimal value is unique:

$$\begin{aligned}\ell(\theta_1) + P(\theta_1) &= \ell(\theta_1) + \lambda(z_{A,1} + z_{I,1} + z_{E^\perp,1})^T\theta_1 \\ &= \ell(\theta_2) + P(\theta_2) = \ell(\theta_2) + \lambda(z_{A,2} + z_{I,2} + z_{E^\perp,2})^T\theta_2.\end{aligned}$$

We subtract $\lambda(z_{A,1} + z_{I,1} + z_{E^\perp,1})^T\theta_2$ from both sides to obtain

$$\begin{aligned}&\ell(\theta_1) + \lambda(z_{A,1} + z_{I,1} + z_{E^\perp,1})^T(\theta_1 - \theta_2) \\ &\quad = \ell(\theta_2) + \lambda(z_{A,2} + z_{I,2} + z_{E^\perp,2} - z_{A,1} - z_{I,1} - z_{E^\perp,1})^T\theta_2.\end{aligned}$$

We rearrange this expression to obtain

$$\begin{aligned}&\ell(\theta_1) - \ell(\theta_2) + \lambda(z_{A,1} + z_{I,1} + z_{E^\perp,1})^T(\theta_1 - \theta_2) \\ &\quad = \lambda(z_{A,2} + z_{I,2} + z_{E^\perp,2} - z_{A,1} - z_{I,1} - z_{E^\perp,1})^T\theta_2.\end{aligned}$$

We substitute in (A.3) to obtain

$$\begin{aligned}&\ell(\theta_1) - \ell(\theta_2) - \nabla\ell(\theta_1)^T(\theta_1 - \theta_2) \\ &\quad = \lambda(z_{A,2} + z_{I,2} + z_{E^\perp,2} - z_{A,1} - z_{I,1} - z_{E^\perp,1})^T\theta_2.\end{aligned}$$

Since ℓ is convex, the left side is non-positive and

$$(z_{A,2} + z_{I,2} + z_{E^\perp,2})^T\theta_2 \leq (z_{A,1} + z_{I,1} + z_{E^\perp,1})^T\theta_2.$$

Since θ_1 and θ_2 are in S, we can ignore the terms $z_{E^\perp,2}^T\theta_2$ and $z_{E^\perp,1}^T\theta_2$ to obtain

$$(z_{A,2}+z_{I,2})^T\theta_2 \le (z_{A,1}+z_{I,1})^T\theta_2.$$

But we also know

$$\begin{aligned}(z_{A,1}+z_{I,1})^T\theta_2 &\le \sup_u \{u^T\theta_2 \mid u\in A\} + \sup_u \{u^T\theta_2 \mid u\in I\}\\ &= z_{A,2}^T\theta_2 + z_{I,2}^T\theta_2.\end{aligned}$$

We combine these two inequalities to obtain

$$(z_{A,2}+z_{I,2})^T\theta_2 = (z_{A,1}+z_{I,1})^T\theta_2 \le z_{A,2}^T\theta_2 + z_{I,1}^T\theta_2$$

This simplifies to $z_{I,2}^T\theta_2 \le z_{I,1}^T\theta_2$. If $z_{I,1}\in \operatorname{relint}(I)$, then

$$\begin{aligned}z_{I,1}^T\theta_2 &= z_{I,2}^T\theta_2 \text{ if } \theta_2 \text{ has no component in } \operatorname{span}(I)\\ z_{I,1}^T\theta_2 &< z_{I,2}^T\theta_2 \text{ if } \theta_2 \text{ has a component in } \operatorname{span}(I).\end{aligned}$$

But we also know $z_{I,2}^T\theta_2 \le z_{I,1}^T\theta_2$. Thus θ_2 has no component in $\operatorname{span}(I)$. □

Proof of Lemma 3.7. The Taylor remainder term is simply

$$R = \nabla\ell(\bar\theta) - \nabla\ell(\theta^\star) - Q(\bar\theta-\theta^\star).$$

By mean value theorem (along $\bar\theta-\theta^\star$), we have

$$R = \int_0^1 \big(\nabla^2\ell(\theta^\star+\alpha(\bar\theta-\theta^\star)) - Q\big)(\bar\theta-\theta^\star)\,d\alpha.$$

Since $\nabla^2\ell$ is Lipschitz continuous with constant L over C,

$$\begin{aligned}\|R\|_2 &= \left\|\int_0^1 \big(\nabla^2\ell(\theta^\star+\alpha(\bar\theta-\theta^\star)) - Q\big)(\bar\theta-\theta^\star)\,d\alpha\right\|_2\\ &\le \int_0^1 \big\|\nabla^2\ell(\theta^\star+\alpha(\bar\theta-\theta^\star)) - Q\big\|\big\|\bar\theta-\theta^\star\big\|_2\,d\alpha\\ &\le \int_0^1 L\big\|\bar\theta-\theta^\star\big\|_2^2\alpha\,d\alpha\\ &\le \frac{L}{2}\big\|\bar\theta-\theta^\star\big\|_2^2.\end{aligned}$$

By Lemma 3.5, we have

$$\|R\|_2 \le \frac{2L}{m^2}\Big(\kappa_\rho + \frac{\tau}{4}\frac{\kappa_\varrho}{\kappa_{\mathrm{IC}}}\Big)^2\lambda^2.$$

To ensure $\frac{\kappa_{\mathrm{IC}}}{\lambda}\varrho^*(R) \le \frac{\tau}{4}$, it sufficies to ensure $\frac{\kappa_{\mathrm{IC}}}{\lambda}\|R\|_2 \le \frac{\tau}{4\kappa_{\varrho^*}}$. Plugging in the choice of λ gives the desired conclusion. □

Appendix B: Proofs of lemmas in Section 4

Proof of Lemma 4.3. First, we prove an auxiliary result: the sufficient statistics of an exponential family are subexponential random variables.

Lemma B.1. *Let* $\phi_i^\star = \mathbf{E}_{\theta^\star}[\phi_i(x)]$ *and* $\phi_i(x)$ *be a sufficient statistics of a regular exponential family. The random variable* $\phi_i(x) - \phi_i^\star$ *is subexponential:*

$$\mathbf{E}_{\theta^\star}\left[\exp s_i(\phi_i(x) - \phi_i^\star)\right] \le \exp(1)$$

for some $s_i > 0$.

Proof.

$$\begin{aligned}
&\mathbf{E}_{\theta^\star}\left[\exp s_i(\phi_i(x) - \phi_i^\star)\right]\\
&\int dx\; h(x)\exp\left(\theta^{\star T}\phi(x) + s_i\phi_i(x) - A(\theta^\star) - s_i\phi_i^\star\right)\\
&\exp\left(-A(\theta^\star) - s_i\phi_i^\star\right)\int dx\; h(x)\exp\left((\theta^\star + s_ie_i)^T\phi(x)\right)\\
&\exp\left(-A(\theta^\star) - s_i\phi_i^\star + A(\theta^\star + s_ie_i)\right)\\
&\exp(-s_i\phi_i^\star)\exp\left(A(\theta^\star + s_ie_i) - A(\theta^\star)\right).
\end{aligned}$$

Using continuity and the regular exponential family, $|A(\theta^\star + s_ie_i) - A(\theta^\star)| < \epsilon$ and $|-s_i\phi_i^\star| < \epsilon$ for small enough s_i. Thus

$$\exp\left(A(\theta^\star + s_ie_i) - A(\theta^\star) - s_i\phi_i^\star\right) \le \exp(1)$$

for small enough s_i. □

We have $\frac{\partial \ell}{\partial \theta_j}(\theta^\star) = \frac{1}{n}\sum_{i=1}^n(-\phi_j(x^{(i)}) + \mathbf{E}_{\theta^\star}[\phi_j(x)])$. Thus $\frac{\partial \ell}{\partial \theta_j}(\theta^\star)$ is a sum of *i.i.d.* subexponential random variables (Lemma B.1) and applying (Vershynin, 2010, Corollary 5.17) gives

$$\mathbf{Pr}\left(\left|\frac{\partial \ell}{\partial \theta_j}(\theta^\star)\right| > t\right) \le 2\exp\left(-cn\min(t^2/K_j^2, t/K_j)\right)$$

where K_j is the Orlicz 1-norm of $-\phi_j(X) + \mathbf{E}_{\theta^\star}[\phi_j(X)]$ (Vershynin, 2010, Definition 5.13). Let $K = \max_j K_j$. By the union bound,

$$\begin{aligned}
\mathbf{Pr}\left(\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > t\right) &\le \mathbf{Pr}\left(\text{for some } j \in g,\ \left|\frac{\partial \ell}{\partial \theta_j}(\theta^\star)\right| > t/\sqrt{|g|}\right)\\
&\le \sum_{j\in g}\mathbf{Pr}\left(\left|\frac{\partial \ell}{\partial \theta_j}(\theta^\star)\right| > t/\sqrt{|g|}\right)\\
&\le 2\exp\left(-cn\min\left(\frac{t^2}{|g|K^2}, \frac{t}{\sqrt{|g|}K}\right)\right)\\
&\le 2\exp\left(-cn\min\left(\frac{t^2}{\max_{g\in\mathcal{G}}|g|K^2}, \frac{t}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{Pr}\left(\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > t\right) \le \mathbf{Pr}\left(\text{for some } g\in\mathcal{G},\ \left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > t\right)\\
&\le \sum_{g\in\mathcal{G}}\mathbf{Pr}\left(\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > t\right)\\
&\le |\mathcal{G}| 2\exp\left(-cn\min\left(\frac{t^2}{\max_{g\in\mathcal{G}}|g|K^2},\frac{t}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right)\\
&= 2\exp\left(\log|\mathcal{G}| - cn\min\left(\frac{t^2}{\max_{g\in\mathcal{G}}|g|K^2},\frac{t}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right). \qquad \square
\end{aligned}$$

Proof of Lemma 4.4. By optimality of $\hat{\theta}$,

$$\begin{aligned}
\ell(\hat{\theta}) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le \ell(\theta^\star) + \lambda\left\|\theta^\star\right\|_{2/1}\\
-\hat{\theta}^T\phi^n + A(\hat{\theta}) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le -\theta^{\star T}\phi^n + A(\theta^\star) + \lambda\left\|\theta^\star\right\|_{2/1}\\
-\hat{\theta}^T\phi^n + \nabla A(\theta^\star)^T(\hat{\theta}-\theta^\star) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le -\theta^{\star T}\phi^n + \lambda\left\|\theta^\star\right\|_{2/1}\\
-\hat{\theta}^T\phi^n + \phi^{\star T}(\hat{\theta}-\theta^\star) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le -\theta^{\star T}\phi^n + \lambda\left\|\theta^\star\right\|_{2/1}\\
\lambda\left\|\hat{\theta}\right\|_{2/1} &\le \lambda\left\|\theta^\star\right\|_{2/1} + (\hat{\theta}-\theta^\star)^T(\phi^n-\phi^\star).
\end{aligned}$$

We now bound $\|\hat{\theta}\|_{2/1}$,

$$\begin{aligned}
\lambda\left\|\hat{\theta}\right\|_{2/1} &\le \lambda\left\|\theta^\star\right\|_{2/1} + (\hat{\theta}-\theta^\star)^T(\phi^n-\phi^\star)\\
&\le \lambda\left\|\theta^\star\right\|_{2/1} + \left\|\hat{\theta}-\theta^\star\right\|_{2/1}\left\|\phi^n-\phi^\star\right\|_{2,\infty}\\
&\le \lambda\left\|\theta^\star\right\|_{2/1} + \left\|\hat{\theta}\right\|_{2/1}\left\|\phi^n-\phi^\star\right\|_{2,\infty} + \left\|\theta^\star\right\|_{2/1}\left\|\phi^n-\phi^\star\right\|_{2,\infty}.
\end{aligned}$$

Rearranging gives us,

$$\left\|\hat{\theta}\right\|_{2/1} \le \frac{1}{(\lambda - \left\|\phi^n-\phi^\star\right\|_{2,\infty})}\left(\lambda\left\|\theta^\star\right\|_{2/1} + \left\|\phi^n-\phi^\star\right\|_{2,\infty}\left\|\theta^\star\right\|_{2/1}\right). \tag{B.1}$$

For the second part,

$$\begin{aligned}
\ell(\hat{\theta}) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le \ell(\theta^\star) + \lambda\left\|\theta^\star\right\|_{2/1}\\
-\hat{\theta}^T\phi^n + A(\hat{\theta}) + \lambda\left\|\hat{\theta}\right\|_{2/1} &\le -\theta^{\star T}\phi^n + A(\theta^\star) + \lambda\left\|\theta^\star\right\|_{2/1}\\
A(\hat{\theta}) &\le (\hat{\theta}-\theta^\star)^T\phi^n + A(\theta^\star) + \lambda\left\|\theta^\star\right\|_{2/1} - \lambda\left\|\hat{\theta}\right\|_{2/1}\\
A(\hat{\theta}) &\le \left\|\hat{\theta}\right\|_{2/1}\left\|\phi^n\right\|_{2,\infty} + \left\|\theta^\star\right\|_{2/1}\left\|\phi^n\right\|_{2,\infty} + A(\theta^\star)\\
&\quad + \lambda\left\|\theta^\star\right\|_{2/1}. \qquad \square
\end{aligned}$$

Proof of Corollary 4.5. Let $f(n, |\mathcal{G}|, |g|)$ be a function that inverts the concentration inequality of Lemma 4.3 in the sense

$$\begin{aligned}&\mathbf{Pr}\left(\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > f(n,|\mathcal{G}|,|g|)\right)\\&\le 2\exp\left(\log|\mathcal{G}| - cn\min\left(\frac{f(n,|\mathcal{G}|,|g|)^2}{\max_{g\in\mathcal{G}}|g|K^2}, \frac{f(n,|\mathcal{G}|,|g|)}{\max_{g\in\mathcal{G}}\sqrt{|g|}K}\right)\right)\\&= 2\exp(0).\end{aligned}$$

Thus f is chosen so

$$\log|\mathcal{G}| - cn\min\left(\left(\frac{f}{\max_{g\in G}\sqrt{|g|}K}\right)^2, \frac{f}{\max_{g\in G}\sqrt{|g|}K}\right) = 0. \tag{B.2}$$

Let

$$f(n,|\mathcal{G}|,|g|) := \max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}}.$$

For $n > (\frac{3}{2})^2\frac{\log|\mathcal{G}|}{c}$ the first term in the min is active, so the choice of f satisfies (B.2).

By the following computation, the choice $\lambda = \frac{3\kappa_{\mathrm{IC}}}{\tau}f(n,|\mathcal{G}|,|g|)$ ensures that

$$\mathbf{Pr}\left(\frac{2\kappa_{\mathrm{IC}}}{\tau}\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > \lambda\right) < 2|\mathcal{G}|^{-5/4}$$

$$\begin{aligned}&\mathbf{Pr}\left(\frac{2\kappa_{\mathrm{IC}}}{\tau}\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > \lambda\right)\\&= \mathbf{Pr}\left(\frac{2\kappa_{\mathrm{IC}}}{\tau}\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > \frac{3\kappa_{\mathrm{IC}}}{\tau}\max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}}\right)\\&= \mathbf{Pr}\left(\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > \frac{3}{2}\max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}}\right)\\&\le 2\exp\left(\log|\mathcal{G}| - cn\min\left(9/4\max_{g\in G}|g|K^2\frac{\log|\mathcal{G}|}{cn}, 3/2\frac{f}{\max_{g\in G}\sqrt{|g|}K}\right)\right)\\&= 2\exp\left(\log|\mathcal{G}| - cn\min\left(9/4\frac{\log|\mathcal{G}|}{cn}, 3/2\sqrt{\frac{\log|\mathcal{G}|}{cn}}\right)\right).\end{aligned}$$

Since $n > (\frac{3}{2})^2\frac{\log|\mathcal{G}|}{c}$, we have $9/4\frac{\log|\mathcal{G}|}{cn} < 3/2\frac{\log|\mathcal{G}|}{cn}$ and thus

$$\begin{aligned}&\mathbf{Pr}\left(\frac{2\kappa_{\mathrm{IC}}}{\tau}\max_{g\in\mathcal{G}}\left\|\nabla_{\theta_g}\ell(\theta^\star)\right\|_2 > \lambda\right)\\&\le 2\exp\left(\log|\mathcal{G}| - \frac{9}{4}\log|\mathcal{G}|\right)\\&= 2|\mathcal{G}|^{-5/4}.\end{aligned}$$

For the rest of this proof, we will assume the event $\{\lambda > \frac{2\kappa_{\mathrm{IC}}}{\tau}\max_{g\in\mathcal{G}}\|\nabla_{\theta_g}\ell(\theta^\star)\|_2\}$, so all the following statements hold with probability at least $1 - 2|\mathcal{G}|^{-5/4}$.

Lemma 4.4 shows

$$\begin{aligned}
\left\|\hat{\theta}\right\|_{2/1} &\leq \frac{\lambda \left\|\theta^\star\right\|_{2/1} + \left\|\phi^n - \phi^\star\right\|_{2,\infty} \left\|\theta^\star\right\|_{2/1}}{\lambda - \left\|\phi^n - \phi^\star\right\|_{2,\infty}} \\
&\leq \frac{\lambda \left\|\theta^\star\right\|_{2/1} + \frac{\tau}{2\kappa_{\mathrm{IC}}} \lambda \left\|\theta^\star\right\|_{2/1}}{\lambda - \frac{\tau}{2\kappa_{\mathrm{IC}}} \lambda} \\
&\leq \frac{2 \left\|\theta^\star\right\|_{2/1}}{1 - \frac{\tau}{2\kappa_{\mathrm{IC}}}} \\
&\leq 4 \left\|\theta^\star\right\|_{2/1}
\end{aligned}$$

where we used $\|\phi^n - \phi^\star\|_{2,\infty} = \max_{g \in \mathcal{G}} \|\nabla_{\theta_g} \ell(\theta^\star)\|_2$ and $\frac{\tau}{\kappa_{\mathrm{IC}}} \leq 1$. Lemma 4.4 also shows that

$$\begin{aligned}
A(\hat{\theta}) &\leq \|\theta^\star\|_{2/1} \|\phi^n\|_{2,\infty} + \left\|\hat{\theta}\right\|_{2/1} \|\phi^n\|_{2,\infty} + A(\theta^\star) + \lambda \|\theta^\star\|_{2/1} \\
&\leq 5 \|\theta^\star\|_{2/1} \|\phi^n\|_{2,\infty} + A(\theta^\star) + 3\frac{\kappa_{\mathrm{IC}}}{\tau} f(n, |\mathcal{G}|, |g|) \|\theta^\star\|_{2/1} \\
&\leq 5 \|\theta^\star\|_{2/1} \left(\|\phi^\star\|_{2,\infty} + \frac{3}{2} f(n, |\mathcal{G}|, |g|) \right) + A(\theta^\star) + 3\frac{\kappa_{\mathrm{IC}}}{\tau} f(n, |\mathcal{G}|, |g|) \|\theta^\star\|_{2/1} \\
&= 5 \|\theta^\star\|_{2/1} \left(\|\phi^\star\|_{2,\infty} + \frac{3}{2} \max_{g \in G} \sqrt{|g|} K \sqrt{\frac{\log |\mathcal{G}|}{cn}} \right) + A(\theta^\star) \\
&+ 3\frac{\kappa_{\mathrm{IC}}}{\tau} \max_{g \in G} \sqrt{|g|} K \sqrt{\frac{\log |\mathcal{G}|}{cn}} \|\theta^\star\|_{2/1} \\
&<= 5 \|\theta^\star\|_{2/1} \left(\|\phi^\star\|_{2,\infty} + \max_{g \in G} \sqrt{|g|} K \right) + A(\theta^\star) + 2\frac{\kappa_{\mathrm{IC}}}{\tau} \max_{g \in G} \sqrt{|g|} K \|\theta^\star\|_{2/1} \\
&=: R
\end{aligned}$$

by the triangle inequality, $\frac{3}{2} f(n, |\mathcal{G}|, |g|) > \max_{g \in \mathcal{G}}$, and $n > (\frac{3}{2})^2 \frac{\log |\mathcal{G}|}{c}$.

Thus from the above arguments we know that

$$\hat{\theta} \in C := \{\theta \mid \|\theta\|_{2/1} \leq 4 \left\|\theta^\star\right\|_{2/1} \text{ and } A(\theta) \leq R\}.$$

The subset C is convex and compact. Since the exponential family is minimal on M, $v^T \nabla^2 A(\theta) v > 0$ for $v \in M$ Wainwright and Jordan (2008) and thus strongly convex over the compact subset $C \cap M$ with strong convexity constant m (Assumption 3.1). By the extreme value theorem applied to $\frac{\|\nabla^2 A(\theta) - \nabla^2 A(\theta^\star)\|_2}{\|\theta - \theta^\star\|_2}$, the function $\nabla^2 \ell(\theta)$ has a finite Lipschitz constant L over C.

Before we apply Theorem 3.4, we compute the constants $\kappa_{\rho^*}, \kappa_{\rho^*}$. Since the regularizer is finite (it's a norm), its dual semi-norm is finite. To keep things simple, we let $\varrho = \| \cdot \|_{2/1}$. The constants $\kappa_\rho = \kappa_\varrho, \kappa_{\rho^*}$ are

$$\begin{aligned}
\kappa_\rho &= \sup_\theta \left\{ \|\theta\|_{2/1} \mid \theta \in B_2 \cap \operatorname{span}(B_{2/\infty, \mathcal{S}}) \right\} = \sqrt{|\mathcal{S}|}, \\
\kappa_{\rho^*} &= \sup_x \left\{ \max_{g \in \mathcal{G}} \mid \theta \in B_2 \cap \operatorname{span}(B_{2/\infty, \mathcal{S}}) \right\} \leq 1.
\end{aligned}$$

To apply Theorem 3.4, we need to verify that the choice of λ satisfies

$$\lambda < \frac{m^2}{2L}\frac{\tau}{\kappa_{\mathrm{IC}}\kappa_{\rho^*}}\left(2\kappa_\rho + \frac{\kappa_\rho}{\kappa_{\mathrm{IC}}}\frac{\tau}{2}\right)^{-2}\frac{\tau}{2\kappa_{\mathrm{IC}}}.$$

Substituting the expressions for the compatibility constants into the expression above gives

$$\lambda < \frac{m^2}{2L}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^{-2}\frac{\tau}{2\kappa_{\mathrm{IC}}}$$

or equivalently

$$\frac{3\kappa_{\mathrm{IC}}}{\tau}f(n,|\mathcal{G}|,|g|) < \frac{m^2}{L}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^{-2}\frac{\tau}{2\kappa_{\mathrm{IC}}}. \tag{B.3}$$

Using $f(n,|\mathcal{G}|,|g|) = \max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}}$,

$$\begin{aligned}
\frac{3\kappa_{\mathrm{IC}}}{\tau}\max_{g\in G}\sqrt{|g|}K\sqrt{\frac{\log|\mathcal{G}|}{cn}} &< \frac{m^2}{L}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^{-2}\frac{\tau}{2\kappa_{\mathrm{IC}}} \\
\sqrt{cn} &> 6\frac{\kappa_{\mathrm{IC}}^2}{\tau^2}\max_{g\in G}\sqrt{|g|}K\sqrt{\log|\mathcal{G}|}\frac{L}{m^2}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^{2} \\
n &> \frac{36}{c}\frac{\kappa_{\mathrm{IC}}^4}{\tau^4}\max_{g\in G}|g|K^2\log|\mathcal{G}|\frac{L^2}{m^4}\left(2\sqrt{|\mathcal{S}|} + \frac{\tau}{2\kappa_{\mathrm{IC}}}\sqrt{|\mathcal{S}|}\right)^{4}.
\end{aligned}$$

This completes the proof. We have verified all the assumptions of Theorem 3.4 and applying the theorem for the chosen value of λ gives the desired result. □

Appendix C: Proof of lemmas in Section 5

Proof of Lemma 5.6. Before we apply Lemma 5.2, we compute the constants $\kappa_{\bar\rho}, \kappa_{\bar\varrho}, \kappa_{\mathrm{IC}}$. Since the regularizer is not a norm, we let $\bar\varrho = \|\cdot\|_*$ and check $\lambda > \frac{4\kappa_{\mathrm{IC}}}{\tau}\|\nabla\ell(\Theta^\star)\|_{\mathrm{sp}}$. It's straighforward to check

$$\mathrm{tr}(U_r^T\Delta V) + \left\|U_{p_1-r}^T\Delta V_{p_2-r}\right\|_* \leq \|\Delta\|_*.$$

The "model subspace" M is given by

$$\mathrm{span}(I)^\perp = \left\{U_rX + (V_rY)^T \mid \text{ for any } X\in\mathbf{R}^{r\times p_2}, Y\in\mathbf{R}^{r\times p_1}\right\},$$

and the constants $\kappa_{\bar\rho}, \kappa_{\bar\varrho}$ are given by

$$\begin{gathered}
\kappa_{\bar\rho} = \sup\nolimits_{X,Y}\left\{\mathrm{tr}\left(U_r^T\Delta V_r\right) \mid \|\Delta\|_{\mathrm{F}} \leq 1\right\} = \sqrt{r} \\
\kappa_{\bar\varrho} = \sup\nolimits_{X,Y}\left\{\left\|XV_r + U_rY^T\right\|_* \mid \left\|U_rX + (V_rY)^T\right\|_{\mathrm{F}} \leq 1\right\} = \sqrt{2r}.
\end{gathered}$$

Similarly, the constant κ_{IC} is given by

$$\begin{aligned}
&\left\| U_{p_1-r}^T \left[P_I \left(Q P_{I^\perp} (P_{I^\perp} Q P_{I^\perp})^\dagger Z - Z \right) \right] V_{p_2-r} \right\|_{\mathrm{sp}} \\
&\quad \le \left\| U_{p_1-r}^T \left[P_I Q P_{I^\perp} (P_{I^\perp} Q P_{I^\perp})^\dagger Z \right] V_{p_2-r} \right\|_{\mathrm{sp}} + \left\| U_{p_1-r}^T Z V_{p_2-r} \right\|_{\mathrm{sp}} \qquad \text{(C.1)} \\
&\quad \le (2-\alpha) \left\| Z \right\|_{\mathrm{sp}}
\end{aligned}$$

is at most $2-\alpha$.

To apply Lemma 5.2, we check $\lambda = \frac{8(2-\alpha)}{\alpha}\sigma\sqrt{\frac{p_1+p_2}{n}}$ satisfies the assumptions. By Negahban et al. (2011), Lemma 3,

$$\mathbf{Pr}\left(\frac{8(2-\alpha)}{\alpha n} \left\| \mathcal{X}^*(\epsilon) \right\|_2 > \frac{8(2-\alpha)}{\alpha} \sigma \sqrt{\frac{p_1+p_2}{n}} \right) \le c_1 e^{-c_2(p_1+p_2)},$$

for some universal constants c_1, c_2. Thus the claims of Lemma 5.2 are valid with probability at least $1 - c_1 e^{-c_2(p_1+p_2)}$:

1. consistent: $\|\bar{\Delta}\|_2 \le \frac{2}{m}(\sqrt{2} + \frac{4(2-\alpha)}{\alpha})\sigma\sqrt{\frac{r(p_1+p_2)}{n}}$.
2. PDW feasible: $\|\bar{U}_{p_1-r}\bar{V}_{p_2-r}^T\|_2 \le 1 - \frac{\tau}{2}$. □

Proof of Lemma 5.4. For any $\Delta \in \operatorname{span}(I)^\perp$, we have

$$\begin{aligned}
&\left\| \Theta^\star + \Delta \right\|_* - \left\| \Theta^\star \right\|_* - \operatorname{tr}\left(V_r U_r^T \Delta \right) \\
&\quad = \operatorname{tr}\left(\tilde{V}_r \tilde{U}_r^T (\Theta^\star + \Delta) \right) - \operatorname{tr}\left(V_r U_r^T \Theta^\star \right) - \operatorname{tr}\left(V_r U_r^T \Delta \right),
\end{aligned}$$

where $\tilde{U} \in \mathbf{R}^{p_1 \times r}$ and $\tilde{V} \in \mathbf{R}^{p_2 \times r}$ are the left and right singular factors of $\Theta^\star + \Delta$. Since $\operatorname{tr}(\tilde{V}_r \tilde{U}_r^T \Theta^\star) \le \operatorname{tr}(V_r U_r^T \Theta^\star)$,

$$\begin{aligned}
&\left\| \Theta^\star + \Delta \right\|_* - \left\| \Theta^\star \right\|_* - \operatorname{tr}\left(V_r U_r^T \Delta \right) \le \operatorname{tr}\left(\left(\tilde{U}_r \tilde{V}_r^T - U_r V_r^T \right)^T \Delta \right) \\
&\quad \le \left\| \tilde{U}_r \tilde{V}_r^T - U_r V_r^T \right\|_{\mathrm{F}} \left\| \Delta \right\|_{\mathrm{F}} .
\end{aligned}$$

By Li and Sun (2002), Theorem 2.4,

$$\left\| \tilde{U}_r \tilde{V}_r^T - U_r V_r^T \right\|_{\mathrm{F}} \le \frac{4}{3 s_r} \left\| \Delta \right\|_F$$

for any Δ such that $\|\Delta\|_2 \le \frac{s_r}{2}$. We put the pieces together to obtain the desired bound. □

References

Bach, F. R. (2008). Consistency of trace norm minimization. *The Journal of Machine Learning Research* **9** 1019–1048. MR2417263

Bach, F. R. (2010). Structured sparsity-inducing norms through submodular functions. In *Adv. Neural Inf. Process. Syst. (NIPS)* 118–126.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. MR2807761

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2** 1153–1194. MR2461898

Candès, E. and Recht, B. (2012). Simple bounds for recovering low-complexity models. *Math. Prog. Ser. A* 1–13.

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statis.* **40** 1935–1967. MR3059067

Chandrasekaran, V., Recht, B., Parrilo, P. A. and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics* **12** 805–849. MR2989474

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review* **43** 129–159. MR1854649

Cheng, J., Levina, E. and Zhu, J. (2013). High-dimensional mixed graphical models. *arXiv:1304.2810*.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Jalali, A., Ravikumar, P., Vasuki, V., Sanghavi, S. and ECE, U. (2011). On learning discrete graphical models using group-sparse regularization. In *Int. Conf. Artif. Intell. Stat. (AISTATS).*

James, G. M., Paulson, C. and Rusmevichientong, P. (2012). The constrained lasso. Technical Report, University of Southern California.

Kolar, M., Song, L., Ahmed, A. and Xing, E. (2010). Estimating time-varying networks. *Ann. Appl. Stat.* **4** 94–123. MR2758086

Lee, J. D. and Hastie, T. (2012). Learning mixed graphical models. *arXiv:1205.5012*.

Lee, J. D., Sun, Y. and Saunders, M. A. (2009). Proximal Newton-type methods for minimizing composite functions. In *Adv. Neural Inf. Process. Syst. (NIPS)* 827–835.

Li, W. and Sun, W. (2002). Perturbation bounds of unitary and subunitary polar factors. *SIAM Journal on Matrix Analysis and Applications* **23** 1183–1193. MR1920940

Loh, P. L. and Wainwright, M. J. (2012). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *arXiv:1212.0478*. MR3161456

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statis.* **34** 1436–1462. MR2278363

Negahban, S., Wainwright, M. J. et al. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097. MR2816348

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statis.* **39** 1–47. MR2797839

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259. MR2719855

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statis.* **38** 1287–1319. MR2662343

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59** 3434–3447. MR3061256

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 267–288. MR1379242

Tibshirani, R. J. and Taylor, J. E. (2011). The solution path of the generalized lasso. *Ann. Statis.* **39** 1335–1371. MR2850205

Vaiter, S., Peyré, G., Dossal, C. and Fadili, J. (2013). Robust sparse analysis regularization. *IEEE Trans. Inform. Theory* **59** 2001–2016. MR3043779

van de Geer, S. (2012). Weakly decomposable regularization penalties and structured sparsity. *arXiv:1204.4813*. MR3181133

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*. MR2963170

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449