Supplementary Material for "Insights and algorithms for the multivariate square-root lasso"

1 Proofs of Proposition 1 and Theorem 1

We first provide a number of lemmas which will be useful for establishing results throughout the manuscript. For ease of notation, we use κ in place of $\kappa_{\mathcal{E},c}$. Recall $\bar{c} = (c+1)/(c-1)$ and $\|A\| = \varphi_1(A)$ where $\varphi_j(A)$ is the jth largest singular value of A. Let $\sum_{j,k} |\beta_{j,k}| \equiv |\beta|_1$, $\sum_{(j,k)\in\mathcal{S}} |\beta_{j,k}| \equiv |\beta_{\mathcal{S}}|_1$, and $\sum_{(j,k)\notin\mathcal{S}} |\beta_{j,k}| \equiv |\beta_{\mathcal{S}^c}|_1$.

We first state a result from Eaton (1989), which will be used throughout our proofs. This statement follows immediately from their proof of their Proposition 7.1.

Lemma 1 *(Eaton (1989)) Suppose $Z \in \mathbb{R}^{n\times q}$ is a random matrix which has q singular values almost surely and suppose Z is left-spherical, i.e., for any $n\times n$ orthogonal matrix $O_n^{(n)}$, $O_n^{(n)}Z$ has the same matrix variate distribution as Z. Let $U_Z D_Z V_Z' = \mathrm{svd}(Z)$. Then, $Z(Z'Z)^{-1/2} = U_Z V_Z'$ follows a uniform distribution of the set of $n \times q$ semiorthogonal matrices, i.e., the set of matrices $W_q^{(n)} \in \mathbb{R}^{n\times q}$ such that $W_q^{(n)\prime} W_q^{(n)} = I_q$.*

Lemma 2 *Assume A1 is true. Then, the subgradient of* $\|Y - X\beta_*\|_*$ *is*

$$\nabla_\beta \|Y - X\beta_*\|_* = -X'U_*V_*' = -X'(Y - X\beta_*)[(Y - X\beta_*)'(Y - X\beta_*)]^{-\frac{1}{2}}$$

where $U_*D_*V_*' = \mathrm{svd}(Y - X\beta_*)$.

Proof of Lemma 2. When $(Y - X\beta_*)$ has q nonzero singular values, we can write

$$(Y - X\beta_*)[(Y - X\beta_*)'(Y - X\beta_*)]^{-\frac{1}{2}} = U_*D_*V_*'(V_*D_*U_*'U_*D_*V_*')^{-\frac{1}{2}} = U_*V_*'$$

Thus, we only need to verify the first equality. We proceed with the following steps: first, we derive the subgradient for $\|Y - X\beta_*\|_*$, then we show that the subgradient contains only a single element when $(Y - X\beta_*)$ has q nonzero singular values. Let $\partial_X f(X)$ denote the subgradient of f with respect to X.

To establish the subgradient, we first apply the chain rule for subdifferentials:

$$\partial_\beta \|Y - X\beta\|_* = \left\{-X'H : H \in \partial_B \|B\|_* \mid_{B=Y-X\beta}\right\}. \tag{19}$$

From Watson (1992), letting $U_BD_BV_B' = \mathrm{svd}(B)$, we have

$$\partial_B \|B\|_* = \left\{UV' + W : W \in \mathbb{R}^{p\times q}, \|W\| \leq 1, U_B'W = 0, WV_B = 0\right\}. \tag{20}$$

Hence, combining (19) and (20),

$$\partial_\beta \|Y - X\beta\|_* = \left\{-X'U_\beta V_\beta' - X'W : \|W\| \leq 1, U_\beta'W = 0, WV_\beta = 0\right\}.$$

However, by the rank-nullity theorem, when $Y - X\beta_*$ has q nonzero singular values (i.e., is full column rank), the only such W which can satisfy both $U_\beta'W = 0$ and $WB_\beta = 0$ is $W = 0$.

Thus, in this case, the subgradient is a single point so can write $\nabla_\beta \|Y - X\beta_*\|_* = -X'U_*V_*'$ where $U_*D_*V_*' = \text{svd}(Y - X\beta_*)$. ■

Lemma 3 *Assume A1 is true. Then, for all $\Delta \in \mathcal{C}(\mathcal{S}, c)$,*

$$\frac{1}{\sqrt{n}}\|Y - X\beta_* - X\Delta\|_* - \frac{1}{\sqrt{n}}\|Y - X\beta_*\|_* \geq \kappa\|\Delta\|_F^2 - \frac{1}{\sqrt{n}}\text{tr}(\Delta'X'U_*V_*')$$

where $U_*D_*V_*' = \text{svd}(Y - X\beta_*)$.

Proof of Lemma 3. First, recall that the nuclear norm can be equivalently written

$$\|A\|_* = \sup_{\|Q\|\leq 1} \text{tr}(Q'A).$$

Also recall that for any convex function f, $\mathcal{L}(\beta_*, \Delta) = f(\beta_*+\Delta) - f(\beta_*) - \text{tr}\left\{\nabla_\beta f(\beta_*)'\Delta\right\} \geq 0$. Hence, letting $f(\beta_*) = \frac{1}{\sqrt{n}}\|Y - X\beta_*\|_*$ and using that $\nabla_\beta f(\beta_*) = -\frac{1}{\sqrt{n}}X'U_*V_*'$ under A1 by Lemma 2, we want to show

$$\mathcal{L}(\beta_*, \Delta) = \frac{1}{\sqrt{n}}\|Y - X\beta_* - X\Delta\|_* - \frac{1}{\sqrt{n}}\|Y - X\beta_*\|_* + \frac{1}{\sqrt{n}}\text{tr}(\Delta'XU_*V_*') \geq \kappa\|\Delta\|_F^2.$$

Notice, using the dual definition of the nuclear norm,

$$\mathcal{L}(\beta_*, \Delta) = \sup_{\|Q_1\|\leq 1} \frac{1}{\sqrt{n}}\text{tr}\left\{Q_1'(U_*D_*V_*' - X\Delta)\right\} - \sup_{\|Q_2\|\leq 1} \frac{1}{\sqrt{n}}\text{tr}(Q_2'U_*D_*V_*') + \frac{1}{\sqrt{n}}\text{tr}(\Delta'XU_*V_*')$$

and the Q_2 that maximizes the second term is $Q_2 = U_* V_*'$ by the definition of the nuclear norm, so

$$= \sup_{\|Q_1\| \le 1} \frac{1}{\sqrt{n}} \mathrm{tr}\left\{ Q_1'(U_* D_* V_*' - X\Delta) \right\} - \frac{1}{\sqrt{n}} \mathrm{tr}(V_* U_*' U_* D_* V_*') + \frac{1}{\sqrt{n}} \mathrm{tr}(\Delta' X U_* V_*')$$

$$= \sup_{\|Q_1\| \le 1} \frac{1}{\sqrt{n}} \mathrm{tr}\left\{ (Q_1 - U_* V_*')'(U_* D_* V_*' - X\Delta \right\} \tag{21}$$

where (21) the numerator in the definition of κ. Thus, we have

$$= \sup_{\|Q_1\| \le 1} \frac{1}{\sqrt{n}} \mathrm{tr}\left\{ (Q_1 - U_* V_*')'(U_* D_* V_*' - X\Delta \right\} \ge \kappa \|\Delta\|_F^2$$

for all $\Delta \in \mathcal{C}(\mathcal{S}, c)$, which establishes the result. ■

The following lemma is adapted from Lemma 3 of Negahban et al. (2012).

Lemma 4 *Suppose A1 is true and* $\lambda > \frac{c}{\sqrt{n}} \|X' U_* V_*'\|_{\max}$ *for a constant* $c > 1$*. Then,* $\hat{\Delta} = \hat{\beta} - \beta_*$ *belongs to the set*

$$\mathcal{C}(\mathcal{S}, c) = \left\{ \Delta \in \mathbb{R}^{p \times q} : \bar{c} |\Delta_{\mathcal{S}}|_1 \ge |\Delta_{\mathcal{S}^c}|_1 \right\}.$$

Proof of Lemma 4: Let $f(\beta) = \frac{1}{\sqrt{n}} \|Y - X\beta\|_* + \lambda |\beta|_1$. Since $\hat{\beta}$ is the minimizer of f and because f is convex, letting $\hat{\Delta} = \hat{\beta} - \beta_*$, we have

$$0 \ge f(\beta_* + \hat{\Delta}) - f(\beta_*) \ge \frac{1}{\sqrt{n}} \|Y - X\beta_* - X\hat{\Delta}\|_* - \frac{1}{\sqrt{n}} \|Y - X\beta_*\|_* + \lambda |\beta_* + \hat{\Delta}|_1 - \lambda |\beta_*|_1.$$

Recall, because the nuclear norm is convex, its first order Taylor expansion is nonnegative so under A1, by Lemma 2 we have $\frac{1}{\sqrt{n}}\|Y - X\beta_* - X\hat{\Delta}\|_* - \frac{1}{\sqrt{n}}\|Y - X\beta_*\|_* \geq -\frac{1}{\sqrt{n}}|\mathrm{tr}(\hat{\Delta}'X'U_*V_*')|$. In addition, by the same argument as in Rothman et al. (2008) to obtain equation (11), $\lambda|\beta_* + \hat{\Delta}|_1 - \lambda|\beta_*|_1 \geq \lambda(|\hat{\Delta}_{\mathcal{S}^c}|_1 - |\hat{\Delta}_{\mathcal{S}}|_1)$, so together we have

$$0 \geq -\frac{1}{\sqrt{n}}|\mathrm{tr}(\hat{\Delta}'X'U_*V_*')| + \lambda|\beta_* + \hat{\Delta}|_1 - \lambda|\beta_*|_1 \geq -\frac{1}{\sqrt{n}}|\mathrm{tr}(\hat{\Delta}'X'U_*V_*')| + \lambda(|\hat{\Delta}_{\mathcal{S}^c}|_1 - |\hat{\Delta}_{\mathcal{S}}|_1)$$

Finally, because $|\mathrm{tr}(\hat{\Delta}'X'U_*V_*')| \leq \|X'U_*V_*'\|_{\max}|\hat{\Delta}|_1$, and because $\lambda > \frac{c}{\sqrt{n}}\|X'U_*V_*'\|_{\max}$ by assumption

$$0 \geq -\frac{1}{\sqrt{n}}\|X'U_*V_*'\|_{\max}|\hat{\Delta}|_1 + \lambda(|\hat{\Delta}_{\mathcal{S}^c}|_1 - |\hat{\Delta}_{\mathcal{S}}|_1) \geq -\frac{\lambda}{c}(|\hat{\Delta}_{\mathcal{S}^c}|_1 + |\hat{\Delta}_{\mathcal{S}}|_1) + \lambda(|\hat{\Delta}_{\mathcal{S}^c}|_1 - |\hat{\Delta}_{\mathcal{S}}|_1)$$

which implies

$$\frac{c+1}{c}|\hat{\Delta}_{\mathcal{S}}|_1 \geq \frac{c-1}{c}|\hat{\Delta}_{\mathcal{S}^c}|_1,$$

the desired inequality. ■

Lemma 5 *Assume A1 is true. If* $\lambda = \frac{\kappa\epsilon}{\bar{c}\sqrt{s}}$, *then* $\frac{c}{\sqrt{n}}\|X'U_*V_*'\|_{\max} < \lambda$ *implies* $\|\hat{\beta} - \beta_*\|_F \leq \epsilon$.

Proof of Lemma 5. To prove Lemma 5, we follow the proof technique from Rothman et al. (2008), detailed in Negahban et al. (2012). Define $\mathcal{B}_{\epsilon,c} = \{\Delta \in \mathbb{R}^{p\times q} : \|\Delta\|_F = \epsilon, \bar{c}|\Delta_{\mathcal{S}}|_1 \geq |\Delta_{\mathcal{S}^c}|_1\}$. Let f be the objective function in (2). Because f is convex and $\hat{\beta}$ is its minimizer, and applying Lemma 4,

$$\inf\{f(\beta_* + \Delta) : \Delta \in \mathcal{B}_{\epsilon,c}\} > f(\beta_*) \implies \|\hat{\beta} - \beta_*\|_F \leq \epsilon.$$

Let $D(\Delta) = f(\beta_* + \Delta) - f(\beta_*)$ so that if we can show $D(\Delta) > 0$ for $\Delta \in \mathcal{B}_{\epsilon,c}$, the conclusion follows. First,

$$D(\Delta) = \underbrace{\frac{1}{\sqrt{n}}\|Y - X\beta_* - X\Delta\|_* - \frac{1}{\sqrt{n}}\|Y - X\beta_*\|_*}_{T_1} + \underbrace{\lambda\left(|\beta_* + \Delta|_1 - |\Delta|_1\right)}_{T_2}$$

so that using Lemma 3 to bound T_1 and applying the argument from Rothman et al. (2008) to obtain (11) to bound bound T_2, we have

$$\begin{aligned} &\geq \kappa\|\Delta\|_F^2 - \frac{1}{\sqrt{n}}|\mathrm{tr}(\Delta X' U_* V_*')| + \lambda\left(|\Delta_{\mathcal{S}^c}|_1 - |\Delta_{\mathcal{S}}|_1\right) \\ &\geq \kappa\|\Delta\|_F^2 - \frac{1}{\sqrt{n}}|\Delta|_1 \|X' U_* V_*'\|_{\max} + \lambda(|\Delta_{\mathcal{S}^c}|_1 - |\Delta_{\mathcal{S}}|_1) \qquad (22) \end{aligned}$$

where (22) follows from Hölder's inequality. Thus, since $\frac{c}{\sqrt{n}}\|X'UV'\|_{\max} < \lambda$ by assumption; and because $|\Delta|_1 = |\Delta_{\mathcal{S}}|_1 + |\Delta_{\mathcal{S}^c}|_1$ and $|\Delta_{\mathcal{S}}| \leq \sqrt{s}\|\Delta\|_F$, we have from (22)

$$D(\Delta) \geq \kappa\|\Delta\|_* - \lambda\left(\frac{c+1}{c}\right)\sqrt{s}\|\Delta\|_F \qquad (23)$$

so that finally, because $\|\Delta\|_F = \epsilon$ for $\Delta \in \mathcal{B}_{\epsilon,c}$, $\lambda = \frac{\kappa\epsilon}{\bar{c}\sqrt{s}}$ yields

$$= \kappa\epsilon^2\left\{1 - \lambda\left(\frac{c+1}{c}\right)\frac{\sqrt{s}}{\kappa\epsilon}\right\} = \kappa\epsilon^2\left(1 - \frac{c-1}{c}\right) > 0. \qquad \blacksquare$$

Now, an immediate application of Lemma 5 yields Proposition 1.

Proof of Proposition 1. To prove the first part of Proposition 1, set $\epsilon = \frac{\bar{c}}{\kappa}\sqrt{s}\lambda$ and apply Lemma 5. The second part follows immediately from Lemma 1, which states that A1 and A2 imply $\mathcal{E}(\mathcal{E}'\mathcal{E})^{-1/2}$ is uniformly distributed on the set of semi-orthogonal matrices. $\blacksquare$

Lemma 6 *Let* $\alpha \in (0,1)$ *and* $K > 1$ *be fixed constants. Let* ϕ_1 *and* ϕ_2 *be constants such that* $\phi_1^{-1} + \phi_2^{-1} = 1$ *and suppose* $n \geq \frac{4K^4}{(K^2-1)^2} \log(\phi_2 q/\alpha)$*. Suppose* $\eta \in \mathbb{R}^n$ *has entries which are independent and identically distributed* $\mathrm{N}(0,1)$*. Then, for* $X \in \mathbb{R}^{n \times p}$ *fixed with* $\|X_j\|_2 = 1$ *for* $j = 1, \dots, p$*, it follows that*

$$P\left(\frac{1}{\sqrt{n}} \frac{\|X'\eta\|_{\max}}{\|\eta\|_2} \geq K\sqrt{\frac{2\log(2\phi_1 pq/\alpha)}{n}}\right) \leq \alpha/q.$$

Proof of Lemma 6. To prove the inequality, we use the union bound:

$$P\left(\frac{1}{\sqrt{n}} \frac{\|X'\eta\|_{\max}}{\|\eta\|_2} \geq K\sqrt{\frac{2\log(2\phi_1 qp/\alpha)}{n}}\right) \leq P\left(\frac{1}{n}\|X'\eta\|_{\max} \geq \sqrt{\frac{2\log(2\phi_1 qp/\alpha)}{n}}\right) + P\left(\frac{\|\eta\|_2}{\sqrt{n}} \leq \frac{1}{K}\right).$$

Dealing with the first term, for normally distributed random variable η and fixed X such that for each j, $\|X_j\|_2 = 1$, for $t_1 > 0$,

$$P\left(\frac{1}{n}\|X'\eta\|_{\max} \geq \sqrt{\frac{2\log(2p) + 2t_1}{n}}\right) \leq \exp(-t_1),$$

see, for example, Lemma 17.5 of Van de Geer (2016). Thus, letting $t_1 = \log(\phi_1 q/\alpha)$,

$$P\left(\frac{1}{n}\|X'\eta\|_{\max} \geq \sqrt{\frac{2\log(2\phi_1 qp/\alpha)}{n}}\right) \leq \frac{\alpha}{\phi_1 q}.$$

In addition, we know from Lemma 1 of Laurent and Massart (2000) that for $t_2 > 0$

$$P\left(\frac{1}{n}\|\eta\|_2^2 \leq 1 - 2\sqrt{\frac{t_2}{n}}\right) \leq \exp(-t_2),$$

so that setting $t_2 = n\frac{(K^2-1)^2}{4K^4} > 0$, we have

$$P\left(\frac{1}{\sqrt{n}}\|\eta\|_2 \leq \frac{1}{K}\right) = P\left(\frac{1}{n}\|\eta\|_2^2 \leq \frac{1}{K^2}\right) \leq \exp\left\{-n\frac{(K^2-1)^2}{4K^4}\right\}.$$

Furthermore, by our assumption on n,

$$\exp\left\{-n\frac{(K^2-1)^2}{4K^4}\right\} \le \frac{\alpha}{\phi_2 q}.$$

Thus, by the union bound,

$$P\left(\frac{1}{\sqrt{n}}\frac{\|X'\eta\|_{\max}}{\|\eta\|_2} \ge K\sqrt{\frac{2\log(2\phi_1 qp)}{n}}\right) \le P\left(\frac{1}{n}\|X'\eta\|_{\max} \ge \sqrt{\frac{2\log(2\phi_1 qp/\alpha)}{n}}\right) + P\left(\frac{\|\eta\|_2}{\sqrt{n}} \le \frac{1}{K}\right)$$
$$\le \alpha/\phi_1 q + \alpha/\phi_2 q = \alpha/q. \qquad \blacksquare.$$

Lemma 7 *Suppose A1 and A2 are true. Let $\tilde{c}$ and c be fixed constants such that $\tilde{c} > c > 1$ and (ϕ_1, ϕ_2) be fixed constants such that $\phi_1^{-1} + \phi_2^{-1} = 1$. It follows that*

$$P\left(\frac{c}{\sqrt{n}}\|X'U_*V_*'\|_{\max} \le \tilde{c}\sqrt{\frac{2\log(2\phi_1 pq/\alpha)}{n}}\right) \ge 1-\alpha,$$

if $n \ge \frac{4K^4}{(K^2-1)^2}\log(\phi_2 q/\alpha)$ with $K = \tilde{c}/c$.

Proof of Lemma 7: Under Assumptions A1 and A2, by Lemma 1, $U_*V_*' \in \mathbb{R}^{n\times q}$ is a random matrix uniformly distributed on the set of $n \times q$ orthogonal matrices. Moreover, this fact implies that each column of $O = U_*V_*'$ has a marginal distribution which is uniform on the unit sphere. Thus, by the union bound, we have

$$P\left(\frac{1}{\sqrt{n}}\|X'U_*V_*'\|_{\max} \ge J\right) \le \sum_{k=1}^{q} P\left(\frac{1}{\sqrt{n}}\|X'O_k\|_{\max} \ge J\right). \tag{24}$$

Since the O_k's are each uniformly distributed on the unit sphere, we can write $O_k = \gamma_k/\|\gamma_k\|_2$ where $\gamma_k \in \mathbb{R}^n$ has entries which are independent and identically distributed $\mathrm{N}(0,1)$. Hence, setting $J = \frac{\tilde{c}}{c}\left\{2n^{-1}\log(2\phi_1 pq/\alpha)\right\}^{1/2}$ and applying Lemma 6 to the right hand side of (24), we have the result.

Proof of Theorem 1: Let $\dot{c} = \tilde{c}/c$, $\hat{c} = \frac{(c+1)\tilde{c}}{(c-1)}$ and let $\phi_1^{-1} + \phi_2^{-1} = 1$. Suppose that $(1 - \dot{c}^2)^2 n \geq 4\dot{c}^4 \log(\phi_2 q/\alpha)$. Set $\epsilon = \frac{\bar{c}}{\tilde{c}\kappa}\sqrt{\frac{2s\log(2\phi_1 pq/\alpha)}{n}}$ and $\lambda = \tilde{c}\left\{n^{-1}(2\log(2\phi_1 pq/\alpha)\right\}^{1/2}$. Then, applications of Lemma 5 and Lemma 7 implies

$$
\begin{aligned}
P\left(\|\hat{\beta} - \beta_*\|_F \leq \frac{\hat{c}}{\kappa}\sqrt{\frac{2\log(2\phi_1 pq/\alpha)}{n}}\right) &\geq P\left(\frac{c}{\sqrt{n}}\|X'U_*V_*'\|_{\max} \leq \tilde{c}\left[n^{-1}\left\{2s\log(2\phi_1 pq/\alpha)\right\}\right]^{1/2}\right) \\
&\geq 1 - \alpha.
\end{aligned}
$$

Finally, setting $\phi_1 = \phi/(\phi - 1)$ and $\phi_2 = \phi$ yields the result. ■

2 Additional simulation results

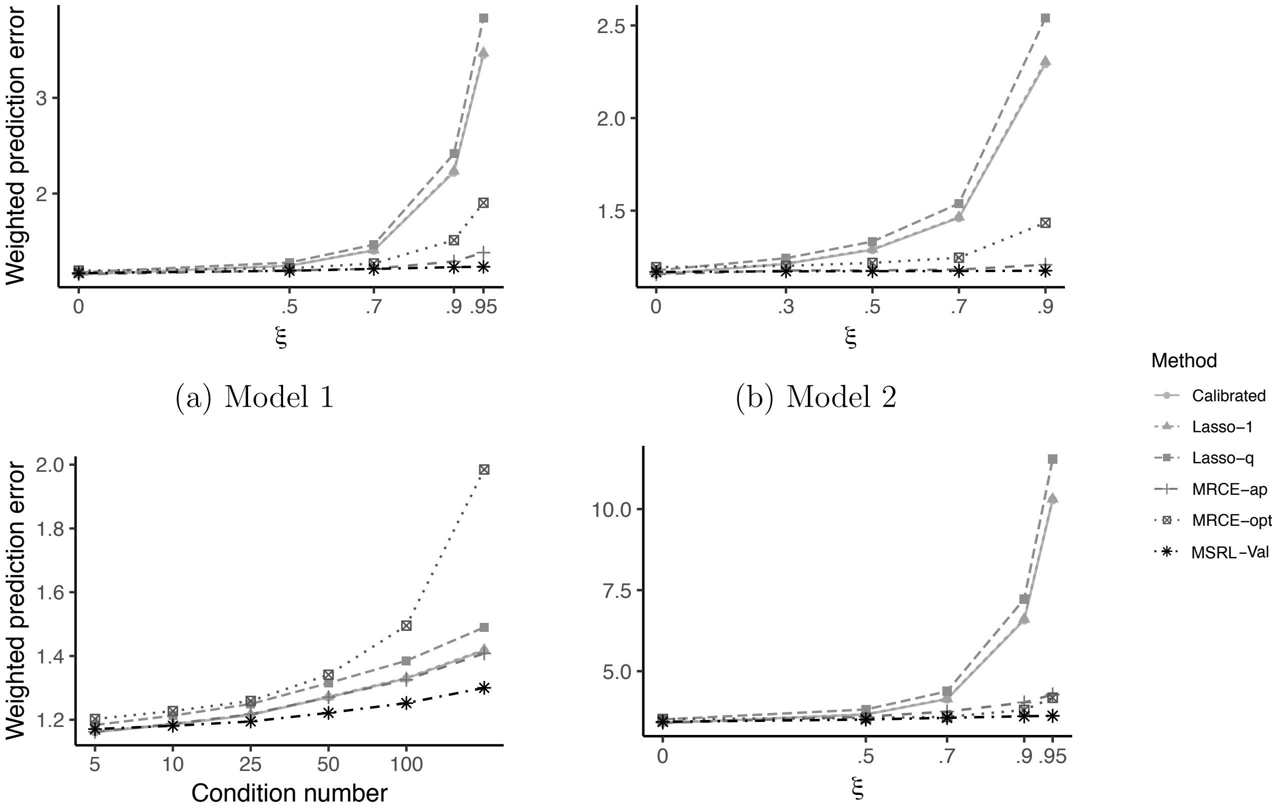
2.1 Weighted prediction error results

A second performance metric we consider is weighted prediction error:

$$
\|\Omega_*^{1/2}(Y_{\text{test}} - \hat{Y}_{\hat{\beta}})\|_F^2 / n_{\text{test}} q,
$$

where $Y_{\text{test}} \in \mathbb{R}^{n_{\text{test}} \times q}$ is the testing set of responses and $\hat{Y}_{\hat{\beta}}$ are the predicted values of Y_{test} based on estimate $\hat{\beta}$. The two error metrics (model error and weighted prediction error) are distinct: model error measures how well $\hat{\beta}$ estimates the conditional mean of the response, whereas weighted prediction error measures prediction accuracy in terms of realizations of the response.

In Figure 3, we display average weighted prediction errors for the six methods under the same settings as in Figure 1. As with model error, we saw that the methods were effectively indistinguishable when ξ was small or the condition number was less than 5, but as both ξ



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

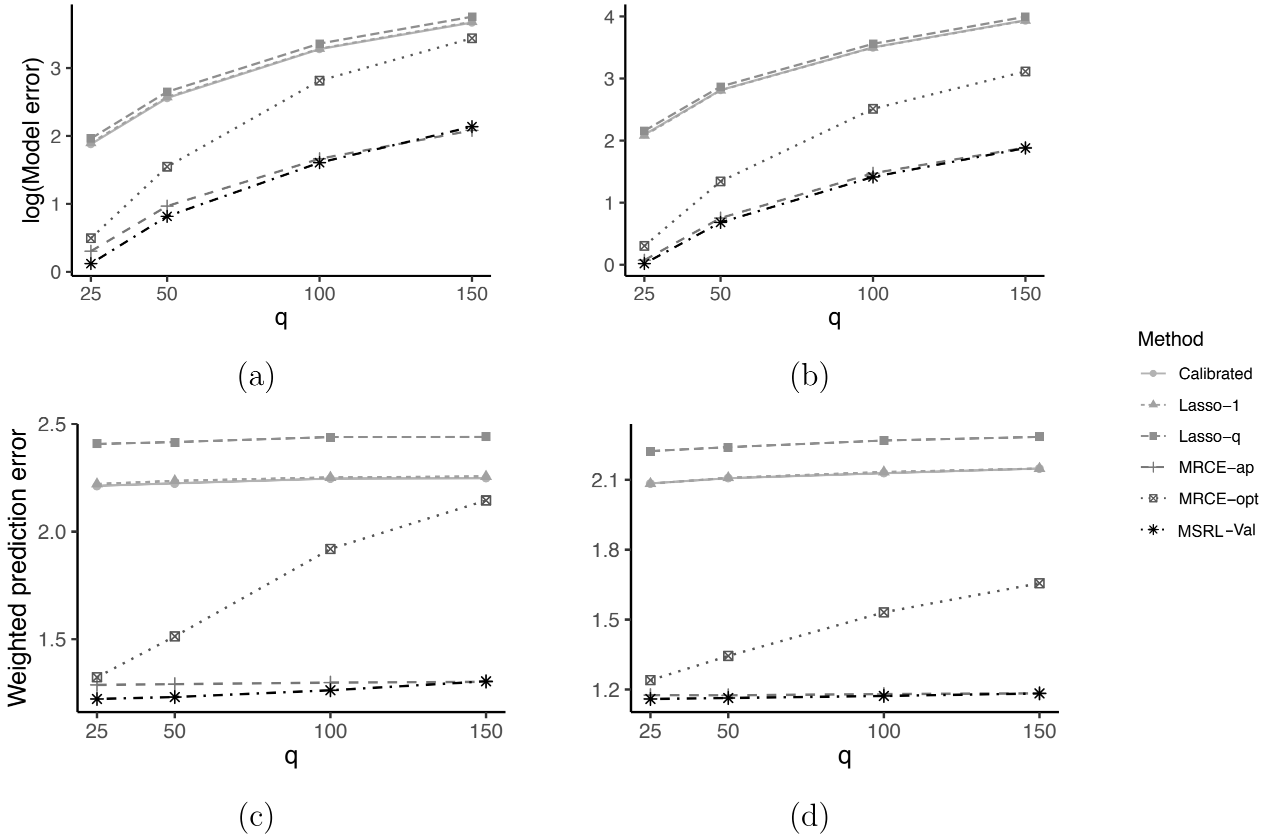
Figure 3: Average weighted prediction error over one hundred independent replications under Model 1 - 4 with ξ or (cond) varying and $q = 50$.

and the condition number increase, `MSRL-Val` tended to outperform competitors. As with model error, `MRCE-ap` also tended to perform relatively well under Models 1 and 2, but less so under Model 3 and 4.

2.2 Model 1 and 2 with q varying

In this section, we display additional results from the simulation studies described in Section 4. Under the settings of Model 1 and 2, we fit $\xi = 0.9$ and let $q \in \{25, 50, 100, 150\}$.

In Figure 4, we display average model errors and weighted prediction errors for the



(a)

(b)

(c)

(d)

Figure 4: (a) Model error and (c) weighted prediction error averaged over one hundred independent replications under Model 1 with q varying and $\xi = 0.9$. (b) Model error and (d) weighted prediction error averaged over one hundred independent replications under Model 2 with q varying and $\xi = 0.9$.

settings from Figure 1. We observed that when $q \leq 100$, `MSRL-Val` outperformed all other methods. When $q = 150$, `MSRL-Val` and `MRCE-ap` performedvery similarly. It is again worth pointing out that both Model 1 and Model 2 conform to the model assumptions of `MRCE-ap`, which requires the estimation of as many as 50^2 more parameters and the selection of two tuning parameters, making `MRCE-ap` a much more computationally burdensome procedure than `MSRL-Val`.

3 Computational details

3.1 Derivation of proximal ADMM β update

In this section, we show that

$$\begin{aligned}
&\underset{\beta\in\mathbb{R}^{p\times q}}{\arg\min}\ \mathcal{M}_{\rho,\eta}(\beta,\Omega^{(k+1)},\Gamma^{(k)};\beta^{(k)})\\
&\qquad= \mathrm{Prox}_{(\rho\eta)^{-1}\tilde{\lambda}\mathcal{P}}\left\{\beta^{(k-1)}+\eta^{-1}X'\left(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k)}-X\beta^{(k-1)}\right)\right\}\\
&\qquad= \underset{\beta\in\mathbb{R}^{p\times q}}{\arg\min}\left\{\frac{1}{2}\|\beta-\beta^{(k-1)}-\eta^{-1}X'\left(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k)}-X\beta^{(k-1)}\right)\|_F^2+(\rho\eta)^{-1}\tilde{\lambda}\mathcal{P}(\beta)\right\}.
\end{aligned}$$

By construction,

$$\begin{aligned}
&\mathcal{M}_{\rho,\eta}(\beta,\Omega^{(k+1)},\Gamma^{(k)};\beta^{(k)})\\
&\qquad= \frac{\rho}{2}\|Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k-1)}-X\beta\|_F^2+\lambda\mathcal{P}(\beta)+\frac{\rho\eta}{2}\|\beta-\beta^{(k-1)}\|_F^2\\
&\qquad\quad -\frac{\rho}{2}\mathrm{tr}\left[(\beta-\beta^{(k-1)})'X'X(\beta-\beta^{(k-1)})\right].
\end{aligned}$$

Then, expanding the first and last terms; and letting C_1, C_2 and C_3 denote constants not depending on β,

$$\begin{aligned}
&= -\rho\,\mathrm{tr}\left[(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k-1)})'X\beta\right]+\lambda\mathcal{P}(\beta)+\frac{\rho\eta}{2}\|\beta-\beta^{(k-1)}\|_F^2-\rho\,\mathrm{tr}\left[\beta'X'X\beta^{(k-1)}\right]+C_1\\
&= -\rho\,\mathrm{tr}\left[\beta'X'(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k-1)}-X\beta^{(k-1)})\right]+\lambda\mathcal{P}(\beta)+\frac{\rho\eta}{2}\|\beta-\beta^{(k-1)}\|_F^2+C_1\\
&\propto -\mathrm{tr}\left[\eta^{-1}\beta'X'(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k-1)}-X\beta^{(k-1)})\right]+\frac{\lambda}{\rho\eta}\mathcal{P}(\beta)+\frac{1}{2}\|\beta-\beta^{(k-1)}\|_F^2+C_2\\
&= \frac{1}{2}\|\beta-\beta^{(k-1)}-\eta^{-1}X'(Y+\rho^{-1}\Gamma^{(k-1)}-\Omega^{(k-1)}-X\beta^{(k-1)})\|_F^2+\frac{\lambda}{\rho\eta}\mathcal{P}(\beta)+C_3,
\end{aligned}$$

the desired equality. ■

3.2 Stochastic approximation of the gradient

Define $\mathcal{H}$ as a randomly selected subset of $\{1, \ldots, n\}$ such that the cardinality of $\mathcal{H}$, $\text{card}(\mathcal{H})$, is much larger than q but smaller than n. Then, let $[Y - X\beta]_{\mathcal{H}}$ and $X_{\mathcal{H}}$ be the submatrices $Y - X\beta$ and X containing only the rows indexed by $\mathcal{H}$. Similarly let $U_{\mathcal{H}(\beta)} D_{\mathcal{H}(\beta)} V'_{\mathcal{H}(\beta)} = \text{svd}([Y - X\beta]_{\mathcal{H}})$. Then, we can efficiently approximate the gradient from (14) since, when $n > \text{card}(\mathcal{H}) \gg q$, $X' U_\beta V'_\beta \approx X'_{\mathcal{H}} U_{\mathcal{H}(\beta)} V'_{\mathcal{H}(\beta)}$, reducing the per iteration complexity of Algorithm 2 to $O\left\{\text{card}(\mathcal{H})(q^2 + pq)\right\}$.

3.3 Method for refitting with unstructured covariance

To refit the estimators in Section 4.4 (i.e., to obtain the `-RF` estimates), we use a seemingly unrelated regressions-type penalized normal maximum likelihood estimator. Suppose we are given $\tilde{\beta}$, the estimate of β_* from which we want to obtain a refitted version. Define the set

$$G(\tilde{\beta}) = \left\{\beta : \beta \in \mathbb{R}^{p \times q}, \beta_{j,k} = 0 \ \forall (j,k) \text{ such that } \tilde{\beta}_{j,k} = 0\right\}.$$

To obtain the refitted version of $\tilde{\beta}$, we solve the following optimization problem:

$$\underset{\beta \in G(\tilde{\beta}), \Omega \in \mathbb{R}^q_+}{\arg\min} \left\{ n^{-1} \text{tr}\left[(Y - X\beta)\Omega(Y - X\beta)'\right] - \log\det(\Omega) + \frac{\alpha}{2}\|\Omega\|_F^2 \right\}, \tag{25}$$

where we fix $\alpha = 10^{-4}$. To solve (25), we use blockwise coordinate descent. Specifically, for $k = 1, \ldots,$ we iterate between the following two steps until convergence:

1. $\Omega^{(k+1)} = U \frac{1}{2\alpha}\left\{-D + (D^2 + 4\alpha I_q)^{1/2}\right\} U'$ where $n^{-1}(Y - X\beta^{(k)})'(Y - X\beta^{(k)}) = UDU'$

with U orthogonal and D diagonal.

2. $\beta^{(k+1)} = \arg\min_{\beta \in G(\tilde{\beta})} \operatorname{tr}\left\{(Y - X\beta)\Omega^{(k+1)}(Y - X\beta)'\right\}$

Step 1 is the well known solution for the ridge-penalized normal likelihood precision matrix estimation problem (e.g., see Witten and Tibshirani (2009)). Step 2 can be solved efficiently using an accelerated projected gradient descent algorithm. We terminate the algorithm when the objective functional value converges.