Supplementary Material for "An explicit mean-covariance parameterization for multivariate response linear regression"

Aaron J. Molstad[1]*, Guangwei Weng[2], Charles R. Doss[2], and Adam J. Rothman[2]
[1]Department of Statistics and Genetics Institute, University of Florida
[2]School of Statistics, University of Minnesota

1 Algorithm implementation and practical considerations

We recommend selecting tuning parameters by minimizing out-of-fold prediction error in V-fold cross-validation. In our implementation, we use the tuning parameter pair

$$\arg\min_{\tau,\lambda} \sum_{v=1}^{V} \|\mathbb{Y}_v - \mathbb{X}_v \hat{\beta}_{-v,\lambda,\tau}\|_F^2,$$

where $\mathbb{Y}_v$ are the responses in the vth fold centered by the responses outside the vth fold, $\mathbb{X}_v$ are the predictors in the vth fold centered by the predictors outside the vth fold, and $\hat{\beta}_{-v,\lambda,\tau}$ is the estimated regression coefficient matrix using the data outside the vth fold with candidate tuning parameters λ and τ.

The first order conditions, which can be derived from Proposition 1, can be used to select a set of reasonable candidate tuning parameters for λ. When $\text{Pen}(\beta) = |\beta|_1 \equiv \sum_{j,k} |\beta_{j,k}|$, if

$$\lambda \geq 2n^{-1}\text{max}_{i,j}\{[X'Y]_{i,j}\text{sign}([X'Y]_{i,j})\},$$

then $\hat{\beta} = 0$. Thus, for any set of candidate τ, we set $\lambda_{\text{max}} = 2n^{-1}\text{max}_{i,j}\{[X'Y]_{i,j}\text{sign}([X'Y]_{i,j})\}$ and $\lambda_{\text{min}} = \delta\lambda_{\text{max}}$. Following Gu et al. (2018), we then set $\lambda_m = \lambda_{\text{max}}^{\frac{M-m}{M-1}} \lambda_{\text{min}}^{\frac{m-1}{M-1}}, \quad m = 1, \ldots, M$, where M is the desired number of candidate tuning parameters. We recommend using a coarse grid of candidate tuning parameters to select τ. In our simulations, we used a subset of $\tau \in \{10^x : x = 4, 3.75, \ldots, -2.75, -3\}$ and $\delta = 10^{-1}$. For applications where a more refined grid is desired, we recommend running an initial cross-validation on a coarse grid for τ and then refining over the tuning parameters which yield small cross-validation prediction error.

*Correspondence: amolstad@ufl.edu

We also employ warm-start initializations to compute the entire solution path more efficiently and avoid local minima. With τ fixed, we first compute our estimator for λ_1 after initializing the algorithm at the matrix of zeros. Then, for λ_2, we initialize the algorithm at the solution obtained for λ_1, and so on.

2 Proofs

2.1 Proof of Theorem 1

We observe n independent realizations of the random pair $(\mathbf{Y}_i, \mathbf{X}_i)$ where each pair is generated according to the data generating model

$$\mathbf{Y}_i = \beta_*' z_i + \epsilon_i, \quad \mathbf{X}_i = z_i + \mathbf{U}_i, \quad i = 1, \ldots, n$$

where each $z_i \in \mathbb{R}^p$ is nonrandom. We assume that the $\mathbf{U}_i \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}^q$ are independent, mean zero random variables with (co)variance $\sigma_{*u}^2 I_p$ and $\gamma_*^2 I_q$ respectively. Let $\boldsymbol{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_n)' \in \mathbb{R}^{n\times q}$, $\boldsymbol{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)' \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)' \in \mathbb{R}^{n\times q}$, $Z = (z_1, \ldots, z_n) \in \mathbb{R}^{n\times p}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \in \mathbb{R}^{n\times q}$. It follows that

$$\mathrm{E}(\boldsymbol{X'Y}) = n\Sigma_Z\beta_*, \quad \mathrm{E}(\boldsymbol{X'X}) = n\Sigma_Z + n\sigma_{*u}^2 I_p, \quad \mathrm{E}(\boldsymbol{Y'Y}) = n\beta_*'\Sigma_Z\beta_* + n\gamma_*^2 I_q, \tag{1}$$

where $\Sigma_Z = n^{-1}Z'Z$.

We now prove the result of Theorem 1. We want to show $\mathrm{E}\left[\nabla\mathcal{F}_\tau(\beta)\right] = 0$. Note that the expectation is taken with respect to the joint distribution of $(\boldsymbol{Y}, \boldsymbol{X})$. Notice, letting $\Omega = \beta_*'\beta_* + \tau I_q$, using the result of Proposition 1,

$$\begin{aligned}
\mathrm{E}\left[\nabla\mathcal{F}_\tau(\beta_*)\right] &= -2n^{-1}\mathrm{E}\left[-\beta_*\Omega^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta_*)'(\boldsymbol{Y} - \boldsymbol{X}\beta_*)\Omega^{-1} - \boldsymbol{Y'X}\Omega^{-1} + \boldsymbol{X'X}\beta_*\Omega^{-1}\right] \\
&= -2n^{-1}\mathrm{E}\left[-\beta_*\Omega^{-1}(\boldsymbol{Y'Y} - \boldsymbol{Y'X}\beta_* - \beta_*'\boldsymbol{X'Y} + \beta_*'\boldsymbol{X'X}\beta_*)\Omega^{-1} - \boldsymbol{Y'X}\Omega^{-1} + \boldsymbol{X'X}\beta_*\Omega^{-1}\right]
\end{aligned}$$

and using the linearity of expectation and the expectations from (1)

$$\begin{aligned}
&\propto -\beta_*\Omega^{-1}\beta_*\Sigma_Z\beta_*\Omega^{-1} - \gamma_*^2\beta_*\Omega^{-1}\Omega^{-1} + \beta_*\Omega^{-1}\beta_*\Sigma_Z\beta_*\Omega^{-1} + \beta_*\Omega^{-1}\beta_*\Sigma_Z\beta_*\Omega^{-1} \\
&\quad - \beta_*\Omega^{-1}\beta_*'\Sigma_Z\beta_*\Omega^{-1} - \sigma_{*u}^2\beta_*\Omega^{-1}\beta_*'\beta_*\Omega^{-1} - \Sigma_Z\beta_*\Omega^{-1} + \Sigma_Z\beta_*\Omega^{-1} + \sigma_{*u}^2\beta_*\Omega^{-1} \\
&= -\gamma_*^2\beta_*\Omega^{-1}\Omega^{-1} - \sigma_{*u}^2\beta_*\Omega^{-1}\beta_*'\beta_*\Omega^{-1} + \sigma_{*u}^2\beta_*\Omega^{-1} \\
&= -\beta_*\Omega^{-1}(\gamma_*^2 I + \sigma_{*u}^2\beta_*'\beta_*)\Omega^{-1} + \sigma_{*u}^2\beta_*\Omega^{-1}
\end{aligned} \tag{2}$$

Thus, if we can select τ such that

$$(\gamma_*^2 I_q + \sigma_{*u}^2\beta_*'\beta_*)(\beta_*'\beta_* + \tau I)^{-1} = \sigma_{*u}^2 I_q, \tag{3}$$

then (2) is equal to zero, from which the conclusion follows. Right multiplying both sides of(3) by $(\beta_*'\beta_* + \tau I)$,

$$\sigma_{*u}^2\beta_*'\beta_* + \gamma_*^2 I_q = \sigma_{*u}^2\beta_*'\beta_* + \sigma_{*u}^2\tau I_q,$$

it follows that if $\tau = \frac{\gamma_*^2}{\sigma_{*u}^2}$, then (3) holds, from which the result follows.

2.2 Proof of Proposition 1

We first compute the gradient of $\mathcal{F}_\tau$ where

$$\mathcal{F}_\tau(\beta) = \operatorname{tr}\left\{\left[\beta'\beta + \tau I_q\right]^{-1} n^{-1}(Y - X\beta)'(Y - X\beta)\right\}.$$

In the first step, we apply the product rule for matrix-valued functions. Let $Q_1 = [\beta'\beta + \tau I_q]$ and $Q_2 = n^{-1}(Y - X\beta)'(Y - X\beta)$, so that we can write the differential

$$\begin{aligned}\mathrm{d}\operatorname{tr}\left\{\left[\beta'\beta + \tau I_q\right]^{-1} n^{-1}(Y - X\beta)'(Y - X\beta)\right\} &= \operatorname{tr}\left\{(\mathrm{d}Q_1^{-1})Q_2 + Q_1^{-1}(\mathrm{d}Q_2)\right\} \qquad (4)\\ &= \operatorname{tr}\left\{(\mathrm{d}Q_1^{-1})Q_2\right\} + \operatorname{tr}\left\{Q_1^{-1}(\mathrm{d}Q_2)\right\} \equiv T_1 + T_2\end{aligned}$$

where (4) follows from Chapter 8.1, (15) and (20), of Magnus and Neudecker (1988). Then, dealing first with T_1,

$$T_1 = \operatorname{tr}\left\{(\mathrm{d}Q_1^{-1})Q_2\right\} = \operatorname{tr}\left\{-Q_1^{-1}(\mathrm{d}Q_1)Q_1^{-1}Q_2\right\} = \operatorname{tr}\left\{-Q_1^{-1}Q_2Q_1^{-1}(\mathrm{d}Q_1)\right\} \equiv \operatorname{tr}\left\{-Q_3(\mathrm{d}Q_1)\right\}$$

where, letting $Q_3 = Q_1^{-1}Q_2Q_1^{-1}$, the second equality comes from Chapter 8.4, (1), of Magnus and Neudecker (1988). Expanding Q_1 and using that Q_3 is symmetric,

$$= \operatorname{tr}\left\{-Q_3(\mathrm{d}Q_1)\right\} = \operatorname{tr}\left\{-Q_3\mathrm{d}(\beta'\beta + \tau I_q)\right\} = \operatorname{tr}\left\{-2Q_3\beta'(\mathrm{d}\beta)\right\} + C_1 \qquad (5)$$

where C_1 is a constant such that $\frac{C_1}{\mathrm{d}\beta} = 0$. Thus, because $\frac{1}{\mathrm{d}X}\operatorname{tr}\{A(\mathrm{d}X)\} = A'$, combining (5) and the definition of Q_3, it follows that

$$\frac{T_1}{\mathrm{d}\beta} = \frac{\operatorname{tr}\left\{-2Q_3\beta'(\mathrm{d}\beta)\right\}}{\mathrm{d}\beta} = -2n^{-1}\beta(\beta'\beta + \tau I_q)^{-1}(Y - X\beta)'(Y - X\beta)(\beta'\beta + \tau I_q)^{-1}.$$

We now simplify T_2. Recalling that Q_1 is symmetric,

$$\begin{aligned}T_2 = \operatorname{tr}\left\{Q_1^{-1}(\mathrm{d}Q_2)\right\} &= \operatorname{tr}\left\{Q_1^{-1}\mathrm{d}\left\{(Y - X\beta)'(Y - X\beta)\right\}\right\}\\ &= \operatorname{tr}\left\{Q_1^{-1}\mathrm{d}(-Y'X\beta - \beta'X'Y + \beta'X'X\beta)\right\} + C_2\\ &= \operatorname{tr}\left\{-Q_1^{-1}Y'X(\mathrm{d}\beta) - Q_1^{-1}(\mathrm{d}\beta)'X'Y + Q_1(\mathrm{d}\beta)'X'X\beta + Q_1\beta'X'X(\mathrm{d}\beta)\right\} + C_2\\ &= \operatorname{tr}\left\{-2Q_1^{-1}Y'X(\mathrm{d}\beta) + 2Q_1\beta'X'X(\mathrm{d}\beta)\right\} + C_2 \qquad (6)\end{aligned}$$

where C_2 is a constant such that $\frac{C_2}{\mathrm{d}\beta} = 0$. Finally, using (6), it follows that

$$\frac{T_2}{\mathrm{d}\beta} = \frac{\operatorname{tr}\left\{-2Q_1^{-1}Y'X(\mathrm{d}\beta) + 2Q_1\beta'X'X(\mathrm{d}\beta)\right\}}{\mathrm{d}\beta} = -2n^{-1}X'Y\left[\beta'\beta + \tau I_q\right]^{-1} + 2n^{-1}X'X\beta\left[\beta'\beta + \tau I_q\right]^{-1}.$$

Thus,

$$\nabla \mathcal{F}_\tau(\beta) = \frac{T_1 + T_2}{\mathrm{d}\beta} \quad = -2n^{-1}\beta Q_1^{-1}(Y - X\beta)'(Y - X\beta)Q_1^{-1} - 2n^{-1}X'YQ_1^{-1} + 2n^{-1}X'X\beta Q_1^{-1},$$

which establishes the first part of Proposition 1.

We now prove the Lipschitz continuity of $\nabla \mathcal{F}_\tau$ over bounded sets $\mathcal{D}_\kappa = \{\beta : \|\beta\|_F \leq \kappa\}$ for $\kappa < \infty$. To establish Lipschitz continuity we show that for fixed κ, there exists a universal constant L_κ such that

$$\|\nabla \mathcal{F}_\tau(\beta) - \nabla \mathcal{F}_\tau(\tilde{\beta})\|_2 \leq L_\kappa \|\beta - \tilde{\beta}\|_2,$$

for all $\beta \in \mathcal{D}_\kappa$ and $\tilde{\beta} \in \mathcal{D}_\kappa$ where $\|\cdot\|_2$ denotes the spectral norm. Let $\tilde{Q}_1 = \tilde{\beta}'\tilde{\beta} + \tau I_q$. Throughout this section, let $\varphi_1(A)$ denote the largest singular value of A. Implicitly, we assume that $\|X'X\|_2 = \varphi_1(X'X)$, $\|X'Y\|_2 = \varphi_1(X'Y)$ and $\|Y'Y\|_2 = \varphi_1(Y'Y)$ are bounded. From the triangle inequality, it follows that

$$\begin{aligned}
&\|\nabla \mathcal{F}_\tau(\beta) - \nabla \mathcal{F}_\tau(\tilde{\beta})\|_2 \\
&\quad = \| -2n^{-1}\beta Q_1^{-1}(Y - X\beta)'(Y - X\beta)Q_1^{-1} + 2n^{-1}\tilde{\beta}\tilde{Q}_1^{-1}(Y - X\tilde{\beta})'(Y - X\tilde{\beta})\tilde{Q}_1^{-1} \\
&\qquad - 2n^{-1}X'Y(Q_1^{-1} - \tilde{Q}_1^{-1}) + 2n^{-1}X'X(\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1})\|_2 \\
&\quad \leq 2n^{-1}\| -\beta Q_1^{-1}(Y - X\beta)'(Y - X\beta)Q_1^{-1} + \tilde{\beta}\tilde{Q}_1^{-1}(Y - X\tilde{\beta})'(Y - X\tilde{\beta})\tilde{Q}_1^{-1}\|_2 \\
&\qquad + 2n^{-1}\|X'Y(Q_1^{-1} - \tilde{Q}_1^{-1})\|_2 + 2n^{-1}\|X'X(\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1})\|_2 \\
&\quad \equiv 2n^{-1}(A_1 + A_2 + A_3). \qquad (7)
\end{aligned}$$

We start by bounding A_3 from (7):

$$\begin{aligned}
A_3 &= \|X'X(\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1})\|_2 \\
&\leq \|X'X\|_2 \|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \qquad (8) \\
&\leq \|X'X\|_2 \|\beta Q_1^{-1} - \tilde{\beta}Q_1^{-1}\|_2 + \|X'X\|_2 \|\tilde{\beta}Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \qquad (9) \\
&\leq \|X'X\|_2 \|Q_1^{-1}\|_2 \|\beta - \tilde{\beta}\|_2 + \|X'X\|_2 \|\tilde{\beta}\|_2 \|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2. \qquad (10)
\end{aligned}$$

where (8) and (10) follow from submultiplicative property of the spectral norm, and (9) follows from the triangle inequality. We then bound $\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2$. Recalling that Q_1 and $\tilde{Q}_1$ are invertible, it follows that

$$\begin{aligned}
\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 &= \|Q_1^{-1}(Q_1 - \tilde{Q}_1)\tilde{Q}_1^{-1}\|_2 \\
&\leq \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|Q_1 - \tilde{Q}_1\|_2 \qquad (11) \\
&= \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\beta'\beta - \tilde{\beta}'\tilde{\beta}\|_2 \\
&\leq \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\beta'\beta - \tilde{\beta}'\beta\|_2 + \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\tilde{\beta}'\beta - \tilde{\beta}'\tilde{\beta}\|_2 \qquad (12) \\
&\leq \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\beta\|_2 \|\beta - \tilde{\beta}\|_2 + \|Q_1^{-1}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\tilde{\beta}\|_2 \|\beta - \tilde{\beta}\|_2 \qquad (13)
\end{aligned}$$

where (11) and (13) follow from the sub-multiplicative property of the spectral norm; and (12) follows from the triangle inequality. Notice, we can bound $\|Q_1^{-1}\|_2$:

$$\|Q_1^{-1}\|_2 \ = \varphi_1(Q_1^{-1}) = [\varphi_q(Q_1)]^{-1} = [\varphi_q(\beta'\beta + \tau I_q)]^{-1} \leq \tau^{-1}. \tag{14}$$

Hence, $\|Q_1^{-1}\|_2 \leq \tau^{-1}$ and by the same argument $\|\tilde{Q}_1^{-1}\|_2 \leq \tau^{-1}$; and by definition of $\mathcal{D}_\kappa$, $\|\beta\|_2 \leq \|\beta\|_F \leq \kappa$ and $\|\tilde{\beta}\|_2 \leq \|\tilde{\beta}\|_F \leq \kappa$. Thus, from (13),

$$\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 \leq 2\tau^{-2}\kappa\|\beta - \tilde{\beta}\|_2, \tag{15}$$

and consequently, combining (10) and (15),

$$\begin{aligned} A_3 \leq \ & \tau^{-1}\varphi_1(X'X)\|\beta - \tilde{\beta}\|_2 + 2\kappa^2\tau^{-2}\varphi_1(X'X)\|\beta - \tilde{\beta}\|_2 & (16)\\ = \ & \varphi_1(X'X)(\tau^{-1} + 2\kappa^2\tau^{-2})\|\beta - \tilde{\beta}\|_2 \equiv K_3\|\beta - \tilde{\beta}\|_2. & (17)\end{aligned}$$

The bound for $A_2 = \|X'Y(Q_1^{-1} - \tilde{Q}_1^{-1})\|_2$ follows immediately from the submultiplicative property of the spectral norm and (15):

$$\|X'Y(Q_1^{-1} - \tilde{Q}_1^{-1})\|_2 \leq \|X'Y\|_2\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 \leq 2\tau^{-2}\kappa\|X'Y\|_2\|\beta - \tilde{\beta}\|_2 \equiv K_2\|\beta - \tilde{\beta}\|_F \tag{18}$$

Then, to bound A_1, we start by using the triangle inequality:

$$\begin{aligned} A_1 = & \| - \beta Q_1^{-1}(Y - X\beta)'(Y - X\beta)Q_1^{-1} + \tilde{\beta}\tilde{Q}_1^{-1}(Y - X\tilde{\beta})'(Y - X\tilde{\beta})\tilde{Q}_1^{-1}\|_2 \\ & \leq \|\beta Q_1^{-1}Y'YQ_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}Y'Y\tilde{Q}_1^{-1}\|_2 + \|\beta Q_1^{-1}Y'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}Y'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 \\ & + \|\beta Q_1^{-1}\beta'X'YQ_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}'X'Y\tilde{Q}_1^{-1}\|_2 + \|\beta Q_1^{-1}\beta'X'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}'X'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 \\ & \equiv A_{11} + A_{12} + A_{13} + A_{14}. \end{aligned} \tag{19}$$

We bound each term in (19). Starting with A_{11},

$$\begin{aligned} A_{11} = & \|\beta Q_1^{-1}Y'YQ_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}Y'Y\tilde{Q}_1^{-1}\|_2 \\ & \leq \|\beta Q_1^{-1}Y'YQ_1^{-1} - \tilde{\beta}Q_1^{-1}Y'YQ_1^{-1}\|_2 + \|\tilde{\beta}Q_1^{-1}Y'YQ_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}Y'Y\tilde{Q}_1^{-1}\|_2 & (20)\\ & \leq \|\beta - \tilde{\beta}\|_2\|Q_1^{-1}\|_2^2\|Y'Y\|_2 + \|\tilde{\beta}\|_2\|Q_1^{-1}Y'YQ_1^{-1} - \tilde{Q}_1^{-1}Y'Y\tilde{Q}_1^{-1}\|_2 & (21)\\ & \leq \|\beta - \tilde{\beta}\|_2\|Q_1^{-1}\|_2^2\|Y'Y\|_2 + \|\tilde{\beta}\|_2\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2\|Y'Y\|_2\|Q_1^{-1}\|_2 \\ & \quad + \|\tilde{\beta}\|_2\|\tilde{Q}_1^{-1}\|_2\|Y'Y\|_2\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 & (22)\end{aligned}$$

where (20) and (22) follow from the triangle inequality; and (21) follows from the submultiplicative property of the spectral norm. Using the bound established for $\|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2$ in (15) along with (22), it follows that

$$A_{11} \leq \tau^{-2}\varphi_1(Y'Y)\|\beta - \tilde{\beta}\|_2 + 4\kappa^2\varphi_1(Y'Y)\tau^{-3}\|\beta - \tilde{\beta}\|_2 \equiv K_{11}\|\beta - \tilde{\beta}\|_2. \tag{23}$$

Similarly, bounding A_{12},

$$
\begin{aligned}
A_{12} &= \|\beta Q_1^{-1} Y'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1} Y'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 \\
&\leq \|\beta Q_1^{-1} Y'X\beta Q_1^{-1} - \tilde{\beta}Q_1^{-1} Y'X\beta Q_1^{-1}\|_2 + \|\tilde{\beta}Q_1^{-1} Y'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1} Y'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (24) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|Y'X\|_2 \|\beta\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1} Y'X\beta Q_1^{-1} - \tilde{Q}_1^{-1} Y'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (25) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|Y'X\|_2 \|\beta\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1} Y'X\beta Q_1^{-1} - \tilde{Q}_1^{-1} Y'X\beta Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1} Y'X\beta Q_1^{-1} - \tilde{Q}_1^{-1} Y'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (26) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|Y'X\|_2 \|\beta\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 \|Y'X\|_2 \|\beta\|_2 \|Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|Y'X\|_2 \|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (27)
\end{aligned}
$$

where (24) and (26) follow from the triangle inequality; and (25) and (27) follow from the submultiplicative property of the spectral norm. Thus, by the bound for A_3, e.g., from (8), $\|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \leq [\varphi_1(X'X)]^{-1} K_3 \|\beta - \tilde{\beta}\|_2$, the bound for (15), and the bound in (27),

$$
\begin{aligned}
A_{12} &\leq \tau^{-2}\varphi_1(Y'X)\kappa\|\beta - \tilde{\beta}\|_2 + 2\varphi_1(Y'X)\kappa^3\tau^{-3}\|\beta - \tilde{\beta}\|_2 + \kappa\tau^{-1}\varphi_1(X'Y)[\varphi_1(X'X)]^{-1}K_3\|\beta - \tilde{\beta}\|_2 \\
&\equiv K_{12}\|\beta - \tilde{\beta}\|_2. && (28)
\end{aligned}
$$

Next, we bound A_{13}:

$$
\begin{aligned}
A_{13} &= \|\beta Q_1^{-1}\beta' X'Y Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}' X'Y\tilde{Q}_1^{-1}\|_2 \\
&\leq \|\beta Q_1^{-1}\beta' X'Y Q_1^{-1} - \tilde{\beta}Q_1^{-1}\beta' X'Y Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}Q_1^{-1}\beta' X'Y Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}' X'Y\tilde{Q}_1^{-1}\|_2 && (29) \\
&\leq \|Q_1^{-1}\|_2^2 \|\beta\|_2 \|X'Y\|_2 \|\beta - \tilde{\beta}\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1}\beta' X'Y Q_1^{-1} - \tilde{Q}_1^{-1}\tilde{\beta}' X'Y\tilde{Q}_1^{-1}\|_2 && (30) \\
&\leq \|Q_1^{-1}\|_2^2 \|\beta\|_2 \|X'Y\|_2 \|\beta - \tilde{\beta}\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1}\beta' X'Y Q_1^{-1} - \tilde{Q}_1^{-1}\tilde{\beta}' X'Y Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1}\tilde{\beta}' X'Y Q_1^{-1} - \tilde{Q}_1^{-1}\tilde{\beta}' X'Y\tilde{Q}_1^{-1}\|_2 && (31) \\
&\leq \|Q_1^{-1}\|_2^2 \|\beta\|_2 \|X'Y\|_2 \|\beta - \tilde{\beta}\|_2 + \|\tilde{\beta}\|_2 \|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \|X'Y\|_2 \|Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\tilde{\beta}\|_2 \|X'Y\|_2 \|Q_1^{-1} - \tilde{Q}_1^{-1}\|_2 && (32)
\end{aligned}
$$

where (29) and (31) follow from the triangle inequality; and (30) and (32) follow from the submultiplicative property of the spectral norm. Then, using $\|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \leq \varphi_1(X'X)^{-1}K_2\|\beta - \tilde{\beta}\|_2$ and the bound in (15), it follows from (32) that

$$
\begin{aligned}
A_{13} &\leq \tau^{-2}\kappa\varphi_1(X'Y)\|\beta - \tilde{\beta}\|_2 + \kappa\varphi_1(X'Y)\tau^{-1}[\varphi_1(X'X)]^{-1}K_3\|\beta - \tilde{\beta}\|_2 + 2\kappa^3\tau^{-3}\varphi_1(X'Y)\|\beta - \tilde{\beta}\|_2 \\
&= K_{13}\|\beta - \tilde{\beta}\|_2. && (33)
\end{aligned}
$$

Finally, we bound A_{14}:

$$
\begin{aligned}
A_{14} &= \|\beta Q_1^{-1}\beta' X'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}' X'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 \\
&\leq \|\beta Q_1^{-1}\beta' X'X\beta Q_1^{-1} - \tilde{\beta}Q_1^{-1}\beta' X'X\beta Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}Q_1^{-1}\beta' X'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}' X'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (34) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|\beta\|_2^2 \|X'X\|_2 + \|\tilde{\beta}\|_2 \|Q_1^{-1}\beta' X'X\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\tilde{\beta}' X'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (35) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|\beta\|_2^2 \|X'X\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|Q_1^{-1}\beta' X'X\beta Q_1^{-1} - \tilde{Q}_1^{-1}\tilde{\beta}' X'X\beta Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1}\tilde{\beta}' X'X\beta Q_1^{-1} - \tilde{Q}_1^{-1}\tilde{\beta}' X'X\tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (36) \\
&\leq \|\beta - \tilde{\beta}\|_2 \|Q_1^{-1}\|_2^2 \|\beta\|_2^2 \|X'X\|_2 + \|\tilde{\beta}\|_2 \|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \|X'X\|_2 \|\beta\|_2 \|Q_1^{-1}\|_2 \\
&\quad + \|\tilde{\beta}\|_2 \|\tilde{Q}_1^{-1}\|_2 \|\tilde{\beta}\|_2 \|X'X\|_2 \|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 && (37)
\end{aligned}
$$

where (34) and (36) follow from the triangle inequality; and (35) and (37) follow from the submultiplicative property of the spectral norm. Hence, applying $\|\beta Q_1^{-1} - \tilde{\beta}\tilde{Q}_1^{-1}\|_2 \leq \varphi_1(X'X)^{-1}K_2\|\beta - \tilde{\beta}\|_2$ and (15),

$$
\begin{aligned}
&\leq \tau^{-2}\kappa^2\varphi_1(X'X)\|\beta - \tilde{\beta}\|_2 + \kappa^2\tau^{-1}K_3\|\beta - \tilde{\beta}\|_2\| + \kappa^2\tau^{-1}K_3\|\beta - \tilde{\beta}\|_2 \\
&\equiv K_{14}\|\beta - \tilde{\beta}\|_2. && (38)
\end{aligned}
$$

Thus, combining (23), (28), (33), (38), (18), and (16),

$$
\|\nabla\mathcal{F}_\tau(\beta) - \nabla\mathcal{F}_\tau(\tilde{\beta})\|_2 \leq 2n^{-1}(K_{11} + K_{12} + K_{13} + K_{14} + K_2 + K_3)\|\beta - \tilde{\beta}\|_2,
$$

where $K_{11}, K_{12}, K_{13}, K_{14}, K_2$, and K_3 are constants depending only on τ, κ, X and Y, which verifies the claim.

3 Handling incomplete or missing responses

We can also apply our method to the case that the response variable Y has entries missing at random. Fitting the multivariate response linear regression model with missing responses is especially difficult since standard estimators of Σ_* do not apply. However, under our proposed parametric link and proposed weighted residual sum of squares criterion, we can construct an approximation to our estimator.

For this scenario, we propose a modified version of $\mathcal{F}_\tau$ which uses only the observed data. Let $\mathcal{O}_i = \{j : Y_{ij} \text{ is non-missing}\}$ for $i = 1, \ldots, n$. We propose the observed-data weighted residual sum of squares criterion

$$
\mathcal{G}_{\tau,\mathcal{O}}(\beta) = n^{-1}\sum_{i=1}^{n}\left\{\sum_{j\in\mathcal{O}_i}\sum_{k\in\mathcal{O}_i}(Y_{ij} - \beta_j' X_i)(Y_{ik} - \beta_k' X_i)[(\beta'\beta + \tau I)^{-1}]_{j,k}\right\}, \tag{39}
$$

where, applying the Woodbury identity,

$$[(\beta'\beta + \tau I)^{-1}]_{j,k} = \begin{cases} -\tau^{-2}\beta_j'[I_p + \tau^{-1}\beta\beta']^{-1}\beta_k & : j \neq k \\ \tau^{-1} - \tau^{-2}\beta_k'[I_p + \tau^{-1}\beta\beta']^{-1}\beta_k & : j = k \end{cases}.$$

As with $\mathcal{F}_\tau$, the weighted residual sum of squares criterion in (39) can be used to define a new class of penalized estimators:

$$\underset{\beta \in \mathbb{R}^{p\times q}}{\arg\min} \left\{ \mathcal{G}_{\tau,\mathcal{O}}(\beta) + \frac{\lambda}{\tau}\mathrm{Pen}(\beta) \right\}. \tag{40}$$

The function $\mathcal{G}_{\tau,\mathcal{O}}$ is differentiable with respect to β, and thus, we can apply the accelerated proximal gradient descent scheme from Algorithm 1 to compute estimators from this class.

As with our proposed estimator, as $\tau \to \infty$, the estimator (40) tends towards the observed data penalized least squares criterion. That is, if the penalty is separable in the components of β (e.g., $\mathrm{Pen}(\beta) = |\beta|_1$), then (40) would tend towards q separate penalized regressions with the jth regression consisting of n_j samples for $j = 1, \ldots, q$.

4 Maximum likelihood estimation

4.1 Blockwise coordinate descent algorithm

In this section, we derive an algorithm to compute the maximum likelihood estimator described in (3). To solve this problem, we propose to use a blockwise coordinate descent scheme. First, we fix the variance parameters (σ_1^2, σ_2^2) and minimize with respect to β. Then, we fix β and minimize with respect to the variance parameters. This is repeated iteratively until the objective function converges.

To solve for β, we use gradient descent. Although this problem is non-convex, first order algorithms can perform well in such settings. First, we derive the following:

$$\nabla_\beta \mathrm{tr}\{n^{-1}(Y - X\beta)(\sigma_1^2\beta'\beta + \sigma_2^2 I_q)^{-1}(Y - X\beta)'\} + \nabla_\beta \log\det(\sigma_1^2\beta'\beta + \sigma_2^2 I_q).$$

By the logic applied in the proof of Proposition 1,

$$\begin{aligned} &\nabla_\beta \mathrm{tr}\{n^{-1}(Y - X\beta)(\sigma_1^2\beta'\beta + \sigma_2^2 I_q)^{-1}(Y - X\beta)'\} \\ &\quad = -\frac{2}{\sigma_1^2 n}\beta\Omega_\beta^{-1}(Y - X\beta)'(Y - X\beta)\Omega_\beta^{-1} + \frac{2}{\sigma_1^2 n}X'Y\Omega_\beta^{-1} + \frac{2}{\sigma_1^2 n}X'X\beta\Omega_\beta^{-1} \end{aligned}$$

where

$$\Omega_\beta = \left(\beta'\beta + \sigma_2^2\sigma_1^{-2}I_q\right).$$

Then, we must compute $\nabla_\beta \log\det(\sigma_1^2\beta'\beta + \sigma_2^2 I_q)$. Using the same approach as in the proof of

Proposition 1, with differential rules coming from Magnus and Neudecker (1988),

$$\begin{aligned}
\mathrm{d}\log\det(\sigma_1^2\beta'\beta+\sigma_2^2 I_q) &= \mathrm{tr}\left\{\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)^{-1}\mathrm{d}\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)\right\} \\
&= \mathrm{tr}\left\{\sigma_1^2\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)^{-1}\mathrm{d}(\beta'\beta)\right\} \\
&= \mathrm{tr}\left\{2\sigma_1^2\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)^{-1}\beta'\mathrm{d}\beta\right\} \\
&= 2\sigma_1^2\beta\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)^{-1}
\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}
&\nabla_\beta \mathrm{tr}\{n^{-1}(Y-X\beta)(\sigma_1^2\beta'\beta+\sigma_2^2 I_q)^{-1}(Y-X\beta)'\}+\nabla_\beta\log\det(\sigma_1^2\beta'\beta+\sigma_2^2 I_q) \\
&\quad= -\frac{2}{\sigma_1^2 n}\beta\Omega_\beta^{-1}(Y-X\beta)'(Y-X\beta)\Omega_\beta^{-1}+\frac{2}{\sigma_1^2 n}X'Y\Omega_\beta^{-1}+\frac{2}{\sigma_1^2 n}X'X\beta\Omega_\beta^{-1}+2\sigma_1^2\left(\sigma_1^2\beta'\beta+\sigma_2^2 I_q\right)^{-1} \\
&\quad= -\frac{2}{\sigma_1^2 n}\beta\Omega_\beta^{-1}(Y-X\beta)'(Y-X\beta)\Omega_\beta^{-1}+\frac{2}{\sigma_1^2 n}X'Y\Omega_\beta^{-1}+\frac{2}{\sigma_1^2 n}X'X\beta\Omega_\beta^{-1}+2\beta\Omega_\beta^{-1} \\
&\quad\equiv \nabla\mathcal{G}(\beta;\sigma_1^2,\sigma_2^2)
\end{aligned}$$

Then, we iteratively update β using $\beta^{(t+1)}=\beta^{(t)}-\frac{1}{\rho}\nabla\mathcal{G}(\beta^{(t)};\sigma_1^2,\sigma_2^2)$ where ρ is a step size for $t=1,2,\ldots$ until the objective function value converges.

Solving for σ_1^2 and σ_2^2 with β fixed is only slightly less challenging as we can use two dimensional solvers. Notice

$$\begin{aligned}
&\mathrm{tr}\{n^{-1}(Y-X\beta)(\sigma_1^2\beta'\beta+\sigma_2^2 I_q)^{-1}(Y-X\beta)'\}+\log\det(\sigma_1^2\beta'\beta+\sigma_2^2 I_q) \\
&\quad= \mathrm{tr}\{S(\sigma_1^2\beta'\beta+\sigma_2^2 I_q)^{-1}\}+\log\det(\sigma_1^2\beta'\beta+\sigma_2^2 I_q)
\end{aligned}$$

where $S=n^{-1}(Y-X\beta)'(Y-X\beta)$. Then, letting UDU' be the eigendecomposition of $\beta'\beta$, we have

$$\begin{aligned}
&= \mathrm{tr}\left\{SU(\sigma_1^2 D+\sigma_2^2 I_q)^{-1}U'\right\}+\sum_{j=1}^{q}\log(\sigma_1^2 D_{jj}+\sigma_2^2) \\
&= \mathrm{tr}\left\{U'SU(\sigma_1^2 D+\sigma_2^2 I_q)^{-1}\right\}+\sum_{j=1}^{q}\log(\sigma_1^2 D_{jj}+\sigma_2^2)
\end{aligned}$$

so that with $W=U'SU$ and W_{jj} being the jth diagonal of W,

$$\begin{aligned}
&= \mathrm{tr}\left\{W(\sigma_1^2 D+\sigma_2^2 I_q)^{-1}\right\}+\sum_{j=1}^{q}\log(\sigma_1^2 D_{jj}+\sigma_2^2) \\
&= \sum_{j=1}^{q}\left(\frac{W_{jj}}{\sigma_1^2 D_{jj}+\sigma_2^2}\right)+\sum_{j=1}^{q}\left\{\log(\sigma_1^2 D_{jj}+\sigma_2^2)\right\}.
\end{aligned}$$

Hence, in our implementation, we use two-dimensional solver `optim` in `R` to solve

$$\underset{\sigma_1^2>0,\sigma_2^2>0}{\arg\min}\left[\sum_{j=1}^{q}\left(\frac{W_{jj}}{\sigma_1^2 D_{jj}+\sigma_2^2}\right)+\sum_{j=1}^{q}\left\{\log(\sigma_1^2 D_{jj}+\sigma_2^2)\right\}\right].$$

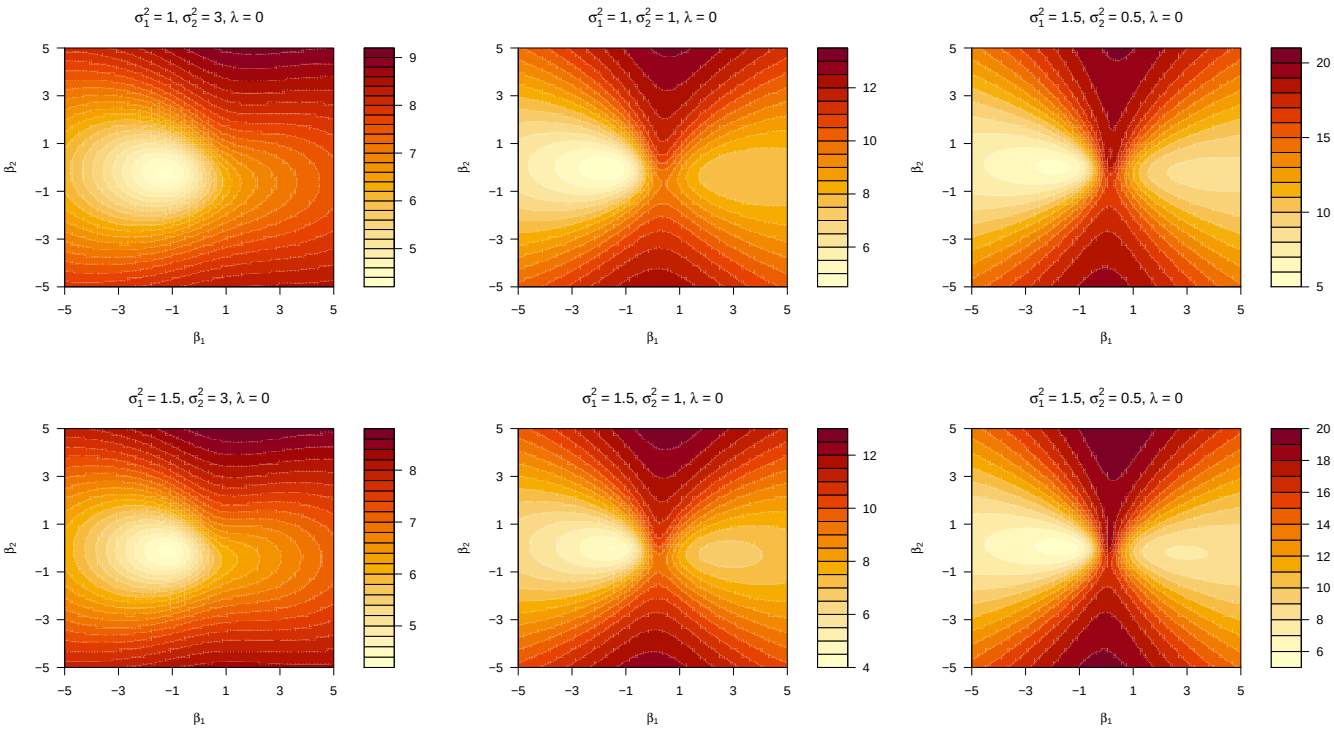

Figure 1: Contour plots of the negative log-likelihood for β with $p = 1$, $q = 2$, $n = 10$, and varying values of σ_1^2 and σ_2^2.

4.2 Challenges of maximum likelihood estimation

In this section, we discuss why maximum likelihood estimation is challenging and discuss heuristically how our approach avoids some of these issues.

One of the first challenges with estimating β_* is that the optimization problems, both ours and maximum likelihood, are non-convex. To show that local minima may indeed be a serious problem for the maximum likelihood estimator, in Figure 1, we display contour plots of the L_1-penalized negative log-likelihood in a simple example with $n = 10$, $q = 2$, and $p = 1$. The data were generated from Model 1 of the main manuscript, with σ_u^2 (here, σ_1^2) and γ_*^2 (here, σ_2^2) equal to one and three respectively; and $\beta = (-.962, -.292)$ generated randomly.

As we see in the contour plots in Figure 1, where no penalty is applied (i.e., $\lambda = 0$), if the values of σ_1^2 and σ_2^2 are poorly initialized, one may end up in a spurious local minimizer. It is important to recall that when performing maximum likelihood estimation, one must iterate between updating β and (σ_1^2, σ_2^2), so if one update of (σ_1^2, σ_2^2) or β leads to a local (rather than global) minimizer, it may be difficult to recover the global joint minimizer. When penalties are applied, e.g., $\lambda \sum_{j,k} |\beta_{j,k}|$, the problem remains: see the contour plots in Figure 2.

To understand how we may avoid this issue with our approach, we display contour plots of our unpenalized weighted residual sum of squares criterion in Figure 3. We see that when τ is large, the criterion is effectively convex: this agrees with our intuition as when $\tau \to \infty$, our

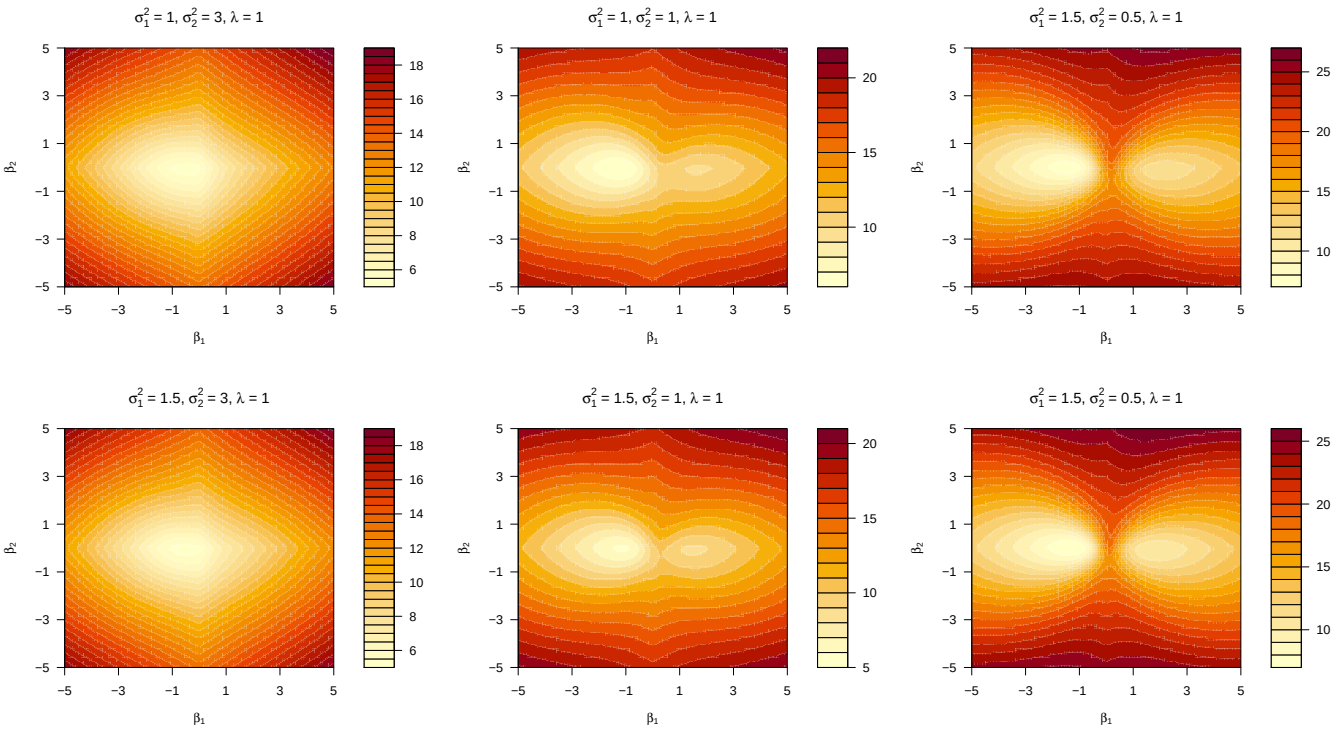

Figure 2: Contour plots of the penalized negative log-likelihood for β with $p = 1$, $q = 2$, $n = 10$, and varying values of σ_1^2 and σ_2^2.

criterion tends towards least squares. As $\tau \to 0$, we notice that there begin to appear local minima. However, recall that when computing the solution path of our estimator, we compute a sequence of β with τ decreasing. First, we compute $\hat{\beta}_1$ with the largest candidate value of τ, τ_1. Then we compute $\hat{\beta}_2$ with the next largest candidate value of τ, τ_2 after initializing the algorithm at $\hat{\beta}_1$. Continuing this procedure, we see that, at least in the examples displayed in Figure 3, this would lead to finding the global minimizer, rather than getting stuck in local minimia. Moreover, unlike maximum likelihood, we need only solve an optimization problem involving β once for a single value of τ, rather than requiring some iterative procedure as is needed in maximum likelihood estimation.

4.3 Comparison to maximum likelihood in low-dimensional settings

In this section, we compare our estimator with τ chosen by cross-validation and $\lambda = 0$ to the maximum likelihood estimator described above. We also include the ordinary least squares estimator to serve as a benchmark.

We generate data in exactly the same fashion as Model 1 of the main manuscript, except we fix $n = 50$, $q = 10$ and let $p \in \{10, 15, 20, 25, 30\}$. We initialize the MLE using $\sigma_1^2 = 0.01$ and σ_2^2 equal to the average variance of the q response variables in the training set. This choice was based on the contour plots in Figure 1, where we saw that small σ_1^2 (relative to σ_2^2) led to

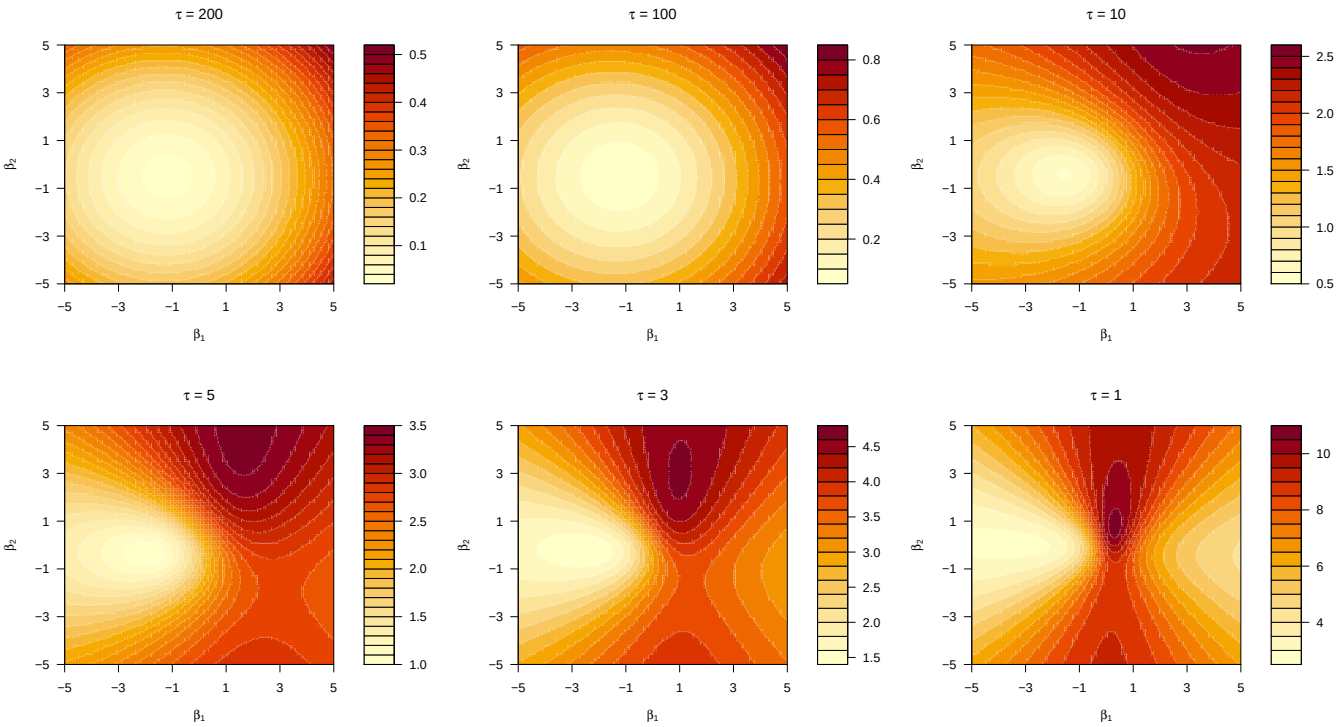

Figure 3: Contour plots of the unpenalized weighted residual sum of squares we propose with $p = 1$, $q = 2$, $n = 10$, and varying values of τ.

fewer local minima. We performed one hundred independent replications, measuring model error, latent model, and test set prediction error for each of the three methods. Results are displayed in Figure 4. We see that in as p approaches n, our method outperforms the MLE, both of which outperform the ordinary least squares estimator (as would be expected). Part of the mechanism of this improvement may be that by treating τ as a tuning parameter, there is a small degree of implicit shrinkage of $\hat{\beta}$, which leads to improved performance over the MLE.

5 Reduced rank regression simulations

In this section, we present results from an additional simulation study under the reduced rank regression data generating model. For one hundred independent replications, we generate data from Model 4, defined below.

– *Model 4*. We first generate n independent copies of $\mathbf{X} \sim \mathrm{N}_p(0, \Sigma_{*X})$ where the (j, k)th entry of Σ_{*X} equals $0.7^{|j-k|}$. Then, conditional on $\mathbf{X} = x$, we generate a realization of $\mathbf{Z}$,

$$\mathbf{Z} = \mathcal{B}_* x + \mathbf{U},$$

and conditioning on $\mathbf{Z} = z$, we generate

$$\mathbf{Y} = \mathcal{A}_* z + \epsilon,$$

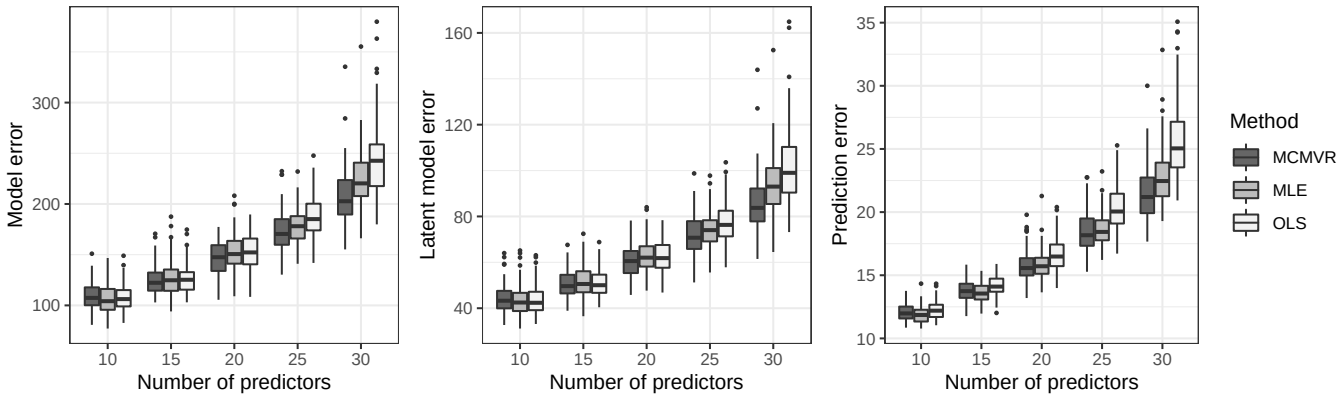

Figure 4: Results comparing the MLE to our method and ordinary least squares under the data generating model described in Section 4.3.

where $\mathcal{B}_* \in \mathbb{R}^{5\times p}$ is a randomly generated semiorthogonal matrix (i.e., $\mathcal{B}_*\mathcal{B}_*' = I_5$); $\mathcal{A}_* \in \mathbb{R}^{q\times 5}$ (details below); $\mathbf{U} \sim \mathrm{N}_p(0, \sigma_{*u}^2 I_p)$, and $\epsilon \sim \mathrm{N}_p(0, \gamma_*^2 I_q)$. Thus, with $\beta_* = \mathcal{B}_*'\mathcal{A}_*'$ and $\mathrm{rank}(\beta_*) = 5$,

$$\mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x) = \beta_*' x, \quad \mathrm{Cov}(\mathbf{Y} \mid \mathbf{X} = x) = \sigma_{*u}^2 \beta_*'\beta_* + \gamma_*^2 I_q,$$

with $\gamma_*^2 = 3$, σ_{*u}^2 varying across settings.

To generate $\mathcal{A}_*$, we randomly assign two elements of each row to be nonzero. These entries are drawn drawn from a uniform distribution on $(0, 10)$. We then standardize each row to have euclidean norm equal to $\sqrt{15}$ and randomly assign the normalized nonzero elements a sign. Under this construction, $\beta_*'\beta_*$ has many entries which are zero, and has diagonals equal to 15, as in the simulation settings in the main manuscript.

We consider a number of competing methods:

– `Ridge-q`. The L_2 penalized least squares estimator

$$\underset{\beta \in \mathbb{R}^{p\times q}}{\arg\min} \left\{ \frac{1}{n}\|Y - X\beta\|_F^2 + \sum_{j=1}^{q} \lambda_j \|\beta_{\cdot,j}\|_2^2 \right\} \tag{41}$$

within tuning parameters λ_j chosen to minimize prediction error in five-fold cross-validation for $j = 1, \ldots, q$ separately.

– `Ridge-1`. The estimator defined in (42) except the tuning parameter $\lambda_j = \lambda$ for $j = 1, \ldots, q$ with λ chosen to minimize prediction error averaged over the q responses in five-fold cross-validation.

– `NN-MC`. The version of our proposed weighted residual sum of squares estimator with $\mathrm{Pen}(\beta) = \|\beta\|_*$, (i.e., the nuclear norm – the norm which sums of the singular values of its matrix argument) with tuning parameters λ and τ chosen using the five fold cross-validation procedure defined Section 1.

$\text{Pen}(\beta)$	$\theta = \text{Prox}_{\tau \text{Pen}}(X)$	References
$\lvert\beta\rvert_1 = \sum_{i,j} \lvert\beta_{i,j}\rvert$	$\theta_{i,j} = \texttt{max}(\lvert X_{i,j}\rvert - \tau, 0)\texttt{sign}(X_{i,j})$	Tibshirani (1996); Rothman et al. (2010)
$\lVert\beta\rVert_{1,2} = \sum_i (\sum_j \beta_{i,j}^2)^{1/2}$	$\theta_{i,\cdot} = \texttt{max}\left(1 - \frac{\tau}{\lVert X_{i,\cdot}\rVert_2}, 0\right) X_{i,\cdot}$	Obozinski et al. (2011); Li et al. (2015)
$\frac{\gamma_1}{\tau}\lvert\beta\rvert_1 + \frac{\gamma_2}{\tau}\lVert\beta\rVert_{1,2}$	(i) $A_{i,j} = \texttt{max}(\lvert X_{i,j}\rvert - \gamma_1, 0)\texttt{sign}(X_{i,j})$ (ii) $\theta_{i,\cdot} = \texttt{max}\left(1 - \frac{\gamma_2}{\lVert A_{i,\cdot}\rVert_2}, 0\right) A_{i,\cdot}$	Peng et al. (2010); Jenatton et al. (2010)
$\lVert\beta\rVert_* = \sum_j \varphi_j(\beta)$	(i) $\texttt{svd}(X) = U\varphi(X)V'$, (ii) $\theta = U\left[\texttt{max}\left\{\varphi(X) - \tau I, 0\right\}\right]V'$	Yuan et al. (2007); Chen et al. (2013)

Table 1: Closed form solutions for the proximal operators of convex penalties used in multivariate response linear regression.

– NN−LS. The nuclear-norm penalized least squares estimator, e.g., Yuan et al. (2007),

$$\underset{\beta \in \mathbb{R}^{p \times q}}{\arg\min} \left\{ \frac{1}{n}\lVert Y - X\beta\rVert_F^2 + \lambda\lVert\beta\rVert_* \right\} \tag{42}$$

within tuning parameter λ chosen to minimize prediction error in five-fold cross-validation averaged across the q responses.

– NN−CA. The nuclear-norm penalize variation of CA, the convex approximation to our estimator described in Section 5.

Computation using the nuclear norm penalty is straightforward from the algorithm proposed in the main manuscript. Specifically, we need only replace the soft-thresholding operator in Algorithm 1 with the proximal operator of the nuclear norm (see Table 1).

For each replication, we record model error, prediction error, and Frobenius norm error as described in Section 5. We display the results for each of the five methods in Figure 5. We see that in each setting, NN−MC performs best, with NN−LS performing similarly. Note that although the differences appear relatively small between methods, this is in part due to the data generating mechanism: in almost every replication, the performance of NN−MC was better than that of NN−LS. Unsurpisingly, the ridge variants and convex approximation to our estimator all perform relatively poorly.

6 Simulations with $\tau = \gamma_*/\sigma_{*u}^2$ fixed

We also considered including an additional competitor in our simulation studies. This competitor was introduced to illustrate the usefulness of treating τ as a tuning parameter as opposed to treating it as the ratio of unknown variance parameters. For this purpose, we reran the simulations from Section 5 of the main manuscript but included one additional competitor: MC−Or, i.e., the "Oracle"

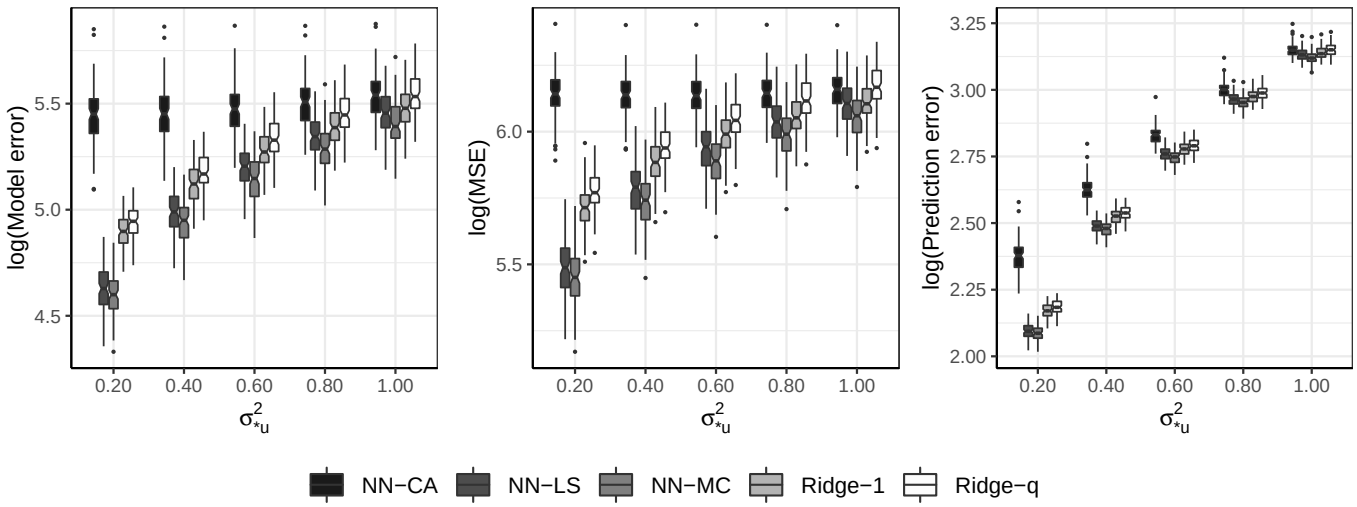

Figure 5: Log model error, log Frobienus norm error, and log prediction error for the five candidate methods over one hundred independent replications under Model 4.

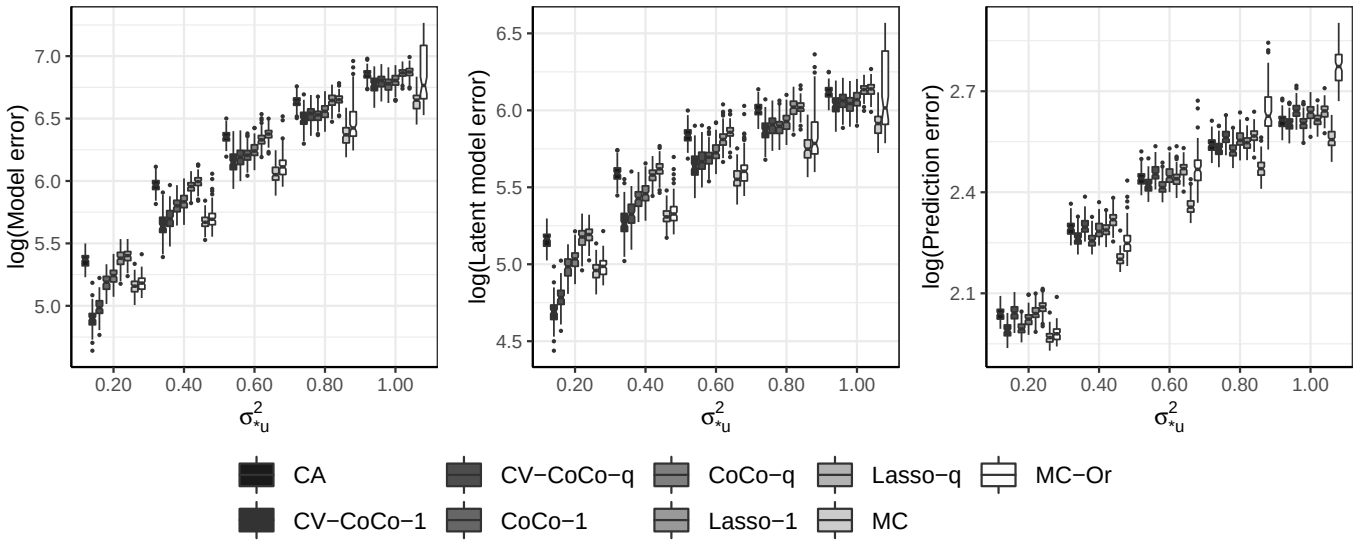

Figure 6: Log model error, log latent model error, and log prediction error for the eight candidate methods over one hundred independent replications under Model 1.

version of our estimator. This version of our estimator is exactly the same as `MC`, except $\tau = \gamma_*/\sigma^2_{*u}$ is fixed throughout. Of course, when $\sigma_{*u} = 0$, this estimator is equivalent to the estimator `Lasso-1`, defined in Section 5 of the main manuscript. In all other situations we considered, this estimator is distinct from `MC` (although, given the appropriate candidate tuning parameter grid, `MC` has `MC-Or` as a special case). Note that to fit `MC-Or`, one must still select the tuning parameter λ, which is done by cross-validation, minimizing validation set prediction error in five-fold cross validation.

Results are displayed in Figure 6. As can be seen in the top row, under Model 1, the "Oracle" version performs similarly to MC when σ^2_{*u}, the measurement error variance, is relatively small. As this variance becomes larger, while MC−Or can perform well, in many replication is performs very poorly. In the bottom row we display results under Model 2, where, we see that in almost every replication, when σ^2_{*u} is larger than 0.25, MC−Or does substantially worse than all competitors. Both of these results can be partly explained by the lack of variable selection accuracy which is highlighted in Table 3. Specifically, when σ^2_{*u} is small, the variable selection accuracy of MC−Or is only slightly worse than the variable selection accuracy of MC. However, as σ^2_{*u} gets larger, e.g., when $\sigma^2_{*u} = 1$ under Model 1, we see that MC−Or has both the lowest TPR and highest FPR of all methods, meaning it excludes the most important variables and includes the most irrelevant variables. Together, these results suggest that the additional flexibility which comes from treating τ as a tuning parameter, even in the case that it has a parameteric interpretation and these parameters are known, lead to improved shrinkage estimation. This results corraborates another observation: that when using the CoCo−lasso estimator of Datta and Zou (2017), treating the measurement error variance as a tuning parameter often leads to better prediction accuracy than using the true value (which is often unknown in practice).

7 Additional simulation settings

We consider another model similar to Model 3 in the main manuscript, which we call Model 5. Specifically, in Model 5 we generate data in exactly the same manner as under Model 3, except we use $[\Sigma_{*E}]_{j,k} = 0.9^{|j-k|}$ for $(j,k) \in \{1,\ldots,q\} \times \{1,\ldots,q\}$. Just as in Model 3, under Model 4, the mean-covariance parametric link is violated as the error correlations are induced from both measurement error and correlation of the errors in the "clean" model.

Results under Model 5 are displayed in Figure 7. Just as under Model 3, we see that as γ_* increases, the performance of our method, MC, approaches the performance of all other methods. Nonetheless, in every setting we considered here, our method performed as well or better than all competing methods. These results further suggest that our method is reasonably robust against violations of the parametric assumption that $[\Sigma_*]_{j,k} \propto \beta'_{*j}\beta_{*k}$ for $j \neq k$.

8 Additional Tables and Figures

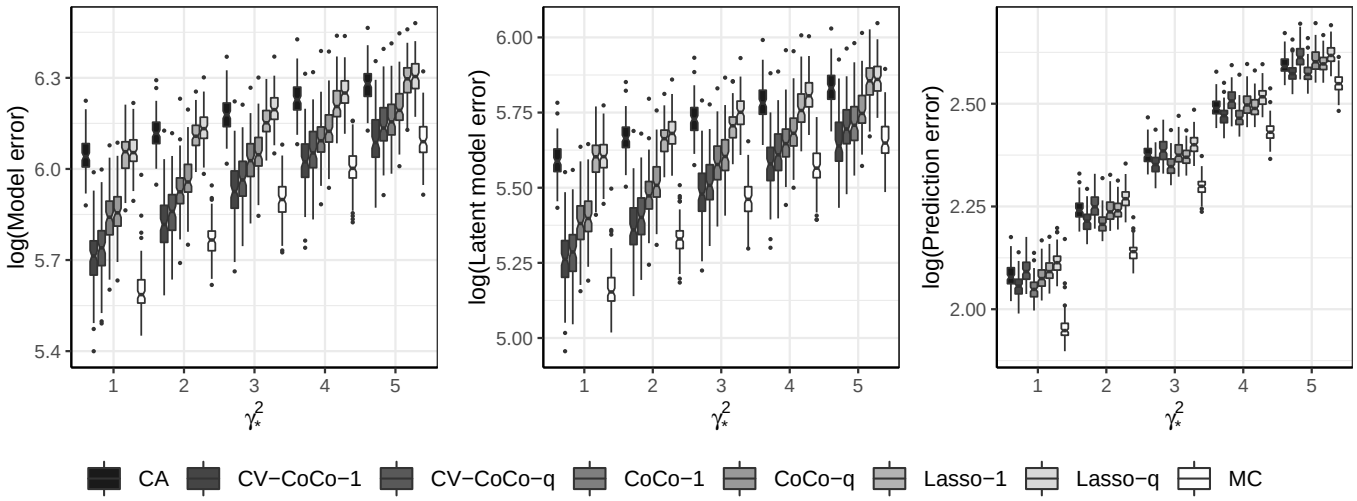

Figure 7: Log model error, log latent model error, and log prediction error for the eight candidate methods over one hundred independent replications under Model 5.

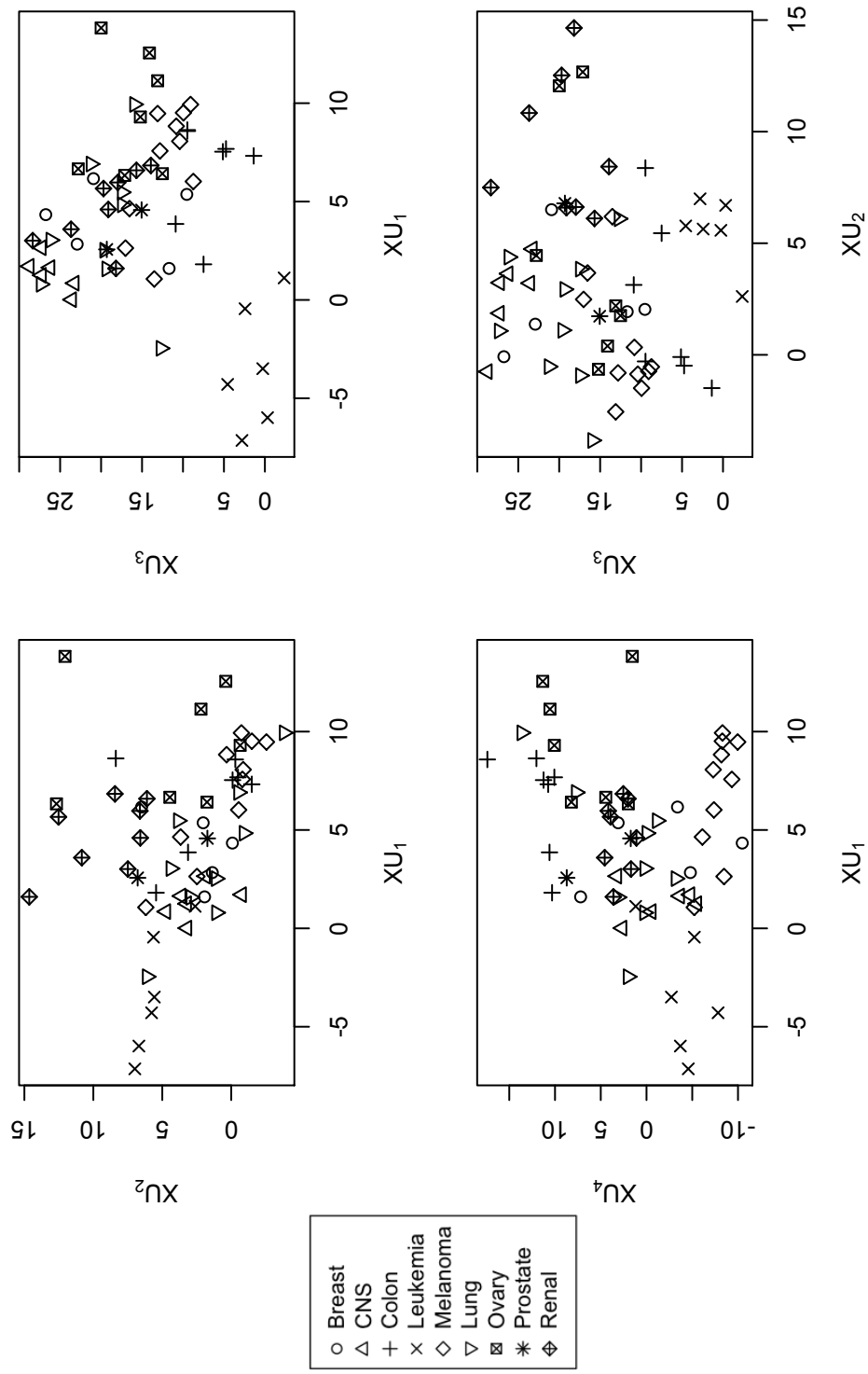

Figure 8: Low dimensional representation of the microRNAs for the NCI-60 cell lines. That is, we plot columns of XU where $\hat{\beta} = UDV'$ is the singular value decomposition of $\hat{\beta}$.

	σ^2_{*u}	CA	CV-CoCo-1	CV-CoCo-q	CoCo-1	CoCo-q	Lasso-1	Lasso-q	MC	MC-Or
1	0.20	0.955 / 0.106	0.917 / 0.040	0.887 / 0.048	0.951 / 0.091	0.942 / 0.096	0.954 / 0.097	0.949 / 0.113	0.973 / 0.078	0.977 / 0.082
	0.40	0.897 / 0.103	0.835 / 0.044	0.819 / 0.059	0.875 / 0.075	0.863 / 0.083	0.897 / 0.099	0.881 / 0.109	0.933 / 0.070	0.946 / 0.111
	0.60	0.843 / 0.098	0.772 / 0.047	0.763 / 0.066	0.807 / 0.068	0.788 / 0.073	0.841 / 0.103	0.820 / 0.104	0.879 / 0.064	0.897 / 0.140
	0.80	0.792 / 0.092	0.722 / 0.051	0.716 / 0.071	0.741 / 0.057	0.713 / 0.062	0.772 / 0.081	0.765 / 0.100	0.819 / 0.059	0.822 / 0.151
	1.00	0.743 / 0.085	0.687 / 0.055	0.668 / 0.074	0.678 / 0.050	0.646 / 0.055	0.722 / 0.070	0.711 / 0.094	0.758 / 0.054	0.602 / 0.107
2	0.20	0.942 / 0.097	0.867 / 0.030	0.851 / 0.039	0.928 / 0.072	0.922 / 0.078	0.936 / 0.084	0.935 / 0.102	0.961 / 0.071	0.951 / 0.079
	0.40	0.890 / 0.091	0.790 / 0.031	0.780 / 0.043	0.837 / 0.051	0.830 / 0.058	0.888 / 0.096	0.870 / 0.094	0.926 / 0.062	0.884 / 0.130
	0.60	0.839 / 0.085	0.728 / 0.033	0.724 / 0.047	0.762 / 0.044	0.746 / 0.045	0.811 / 0.070	0.812 / 0.089	0.878 / 0.057	0.811 / 0.150
	0.80	0.792 / 0.081	0.681 / 0.035	0.675 / 0.049	0.688 / 0.036	0.667 / 0.036	0.779 / 0.080	0.757 / 0.083	0.816 / 0.052	0.673 / 0.128
	1.00	0.746 / 0.075	0.637 / 0.036	0.630 / 0.051	0.613 / 0.028	0.589 / 0.029	0.723 / 0.073	0.708 / 0.079	0.756 / 0.047	0.299 / 0.036

Table 2: True positive and false positive percentages (TPR / FPR) averaged over one hundred independent replications under Models 1 and 2.

	γ^2_*	CA	CV-CoCo-1	CV-CoCo-q	CoCo-1	CoCo-q	Lasso-1	Lasso-q	MC	MC-Or
3	1.00	0.954 / 0.123	0.864 / 0.046	0.853 / 0.064	0.900 / 0.078	0.890 / 0.085	0.914 / 0.099	0.911 / 0.120	0.983 / 0.090	0.987 / 0.196
	2.00	0.917 / 0.113	0.831 / 0.046	0.821 / 0.063	0.872 / 0.077	0.856 / 0.082	0.891 / 0.098	0.881 / 0.114	0.958 / 0.083	0.982 / 0.161
	3.00	0.888 / 0.107	0.800 / 0.046	0.791 / 0.065	0.840 / 0.072	0.822 / 0.079	0.874 / 0.107	0.851 / 0.109	0.923 / 0.077	0.965 / 0.132
	4.00	0.861 / 0.101	0.770 / 0.045	0.765 / 0.066	0.812 / 0.070	0.794 / 0.076	0.841 / 0.098	0.824 / 0.104	0.893 / 0.076	0.941 / 0.114
	5.00	0.840 / 0.097	0.743 / 0.045	0.739 / 0.066	0.785 / 0.067	0.767 / 0.073	0.808 / 0.085	0.796 / 0.100	0.868 / 0.072	0.927 / 0.106

Table 3: True positive and false positive percentages (TPR / FPR) averaged over one hundred independent replications under Model 3.

References

Chen, K., Dong, H., and Chan, K.-s. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920.

Datta, A. and Zou, H. (2017). CoCoLasso for high-dimensional error-in-variables regression. *Ann. Statist.*, 45(6):2400–2426.

Gu, Y., Fan, J., Kong, L., Ma, S., and Zou, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics*, 60(3):1–13.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. R. (2010). Proximal methods for sparse hierarchical dictionary learning. In *ICML*, pages 487–494.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

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

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4(1):53–77.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.*, 19(4):947–962.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(3):329–346.