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

Abstract

We develop a new method to fit the multivariate response linear regression model that exploits a parametric link between the regression coefficient matrix and the error covariance matrix. Specifically, we assume that the correlations between entries in the multivariate error random vector are proportional to the cosines of the angles between their corresponding regression coefficient matrix columns, so as the angle between two regression coefficient matrix columns decreases, the correlation between the corresponding errors increases. We highlight two models under which this parameterization arises: a latent variable reduced-rank regression model and the errors-in-variables regression model. We propose a novel non-convex weighted residual sum of squares criterion which exploits this parameterization and admits a new class of penalized estimators. The optimization is solved with an accelerated proximal gradient descent algorithm. Our method is used to study the association between microRNA expression and cancer drug activity measured on the NCI-60 cell lines. An R package implementing our method, `MCMVR`, is available online.

Keywords: covariance matrix estimation, genomics, measurement error, multivariate regression, non-convex optimization, reduced-rank regresssion

1 Introduction

Some regression analyses have more than one response and these responses are typically associated. When these responses are numerical variables, it is common to apply the multivariate response linear regression model. Let $y_i \in \mathbb{R}^q$ be the observed response for the ith subject, and let

*Correspondence: amolstad@ufl.edu

$x_i \in \mathbb{R}^p$ be the observed predictor for the ith subject. In the multivariate response linear regression model, y_i is a realization of the random vector

$$\mathbf{Y}_i = \mu_* + \beta_*' x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu_* \in \mathbb{R}^q$ is the unknown intercept, $\beta_* \in \mathbb{R}^{p \times q}$ is the unknown regression coefficient matrix, and $\epsilon_1, \ldots, \epsilon_n$ are independent copies of a mean zero random vector with covariance matrix Σ_*. Chapter 7 of Pourahmadi (2013) gives a detailed overview of modern shrinkage methods that fit (1). We review a subset of these methods here.

Several shrinkage estimators of β_* have been proposed through penalized least squares. If the penalty separates across the columns of the optimization variable, then the estimate of β_* can be computed with q separate penalized least-squares regressions, e.g. lasso-penalized least squares. Other penalized least-squares methods assume rows of β_* are zero (Obozinski et al., 2011; Peng et al., 2010), assume β_* is low rank (Izenman, 1975; Yuan et al., 2007), or assume both (Chen and Huang, 2012).

Under the additional assumption that the ϵ_i's are multivariate normal, (1) can be fit by minimizing a penalized negative Gaussian log-likelihood. These likelihood-based methods simultaneously estimate Σ_* and β_* (Izenman, 1975; Rothman et al., 2010; Yin and Li, 2011). There also exist methods with two steps: they first estimate Σ_*^{-1} and then plug this estimate into a penalized normal negative log-likelihood to estimate β_* (Perrot-Dockès et al., 2018). There are also methods that add an assumption that the predictor and response are $(p+q)$-variate normal and develop estimators based on the inverse regression (Molstad and Rothman, 2016) or based on estimating the joint covariance matrix (Lee and Liu, 2012).

We focus on methods that fit (1) by assuming that the error covariance matrix Σ_* and the regression coefficient matrix β_* are parametrically connected. One example is the envelope model, which assumes that the columns of β_* are in a subspace spanned by eigenvectors of Σ_* with small

corresponding eigenvalues (Cook and Zhang, 2015). Focusing on precision matrix estimation, Pourahmadi (1999) proposed a joint mean-covariance model based on an autoregressive interpretation of the Cholesky factor. In this manuscript, we consider a more explicit parametric connection between Σ_* and β_*: we propose to fit (1) under the assumption that

$$\Sigma_* \propto \beta_*'\beta_* + \sigma_{*\epsilon}^2 I_q, \tag{2}$$

where $\sigma_{*\epsilon}^2 \in (0, \infty)$ is unknown. This parametrization links the angle between the jth and kth columns of β_* and the correlation between the jth and kth responses; the motivation is that in some applications, two responses that depend on predictors in a similar way will also depend on unmeasured factors in similar ways. One setting in which this assumption may hold is the multivariate regression of cancer drug activity on microRNA expression profiles measured on the National Cancer Institute (NCI)-60 cell lines. In particular, because variation in cancer drug activity has been shown to be partly explained by -omic factors other than microRNA expression (Chen and Sun, 2017), it may be reasonable to assume that two drugs which depend on microRNA expression in a similar way also depend on unmeasured -omic factors (e.g., somatic mutations) in similar ways. In Section 7, we show that assuming (2) in this application leads to improved prediction accuracy and fitted models which may provide new biological insights about cancer drug activity.

Formally, (2) implies that for each $(j, k) \in \{1, \ldots, q\}^2$, the cosine of the angle between the jth and kth column of β_* is proportional to the (j, k)th entry in Σ_*, so as the angle between the jth and kth column of β_* decreases, the correlation between the corresponding errors increases. For example, if $\beta_{*j}'\beta_{*k} = 0$ (e.g., the jth and kth responses depend on different predictors), then it may be natural to assume that the jth and kth errors are uncorrelated since the jth and kth responses relate to the p predictors in distinct ways. If $\beta_{*j}'\beta_{*k}$ were instead relatively large, then it may be natural to assume that the jth and kth errors are positively associated since the jth and kth responses relate to the p predictors in similar ways.

2 A new class of regression coefficient matrix estimators

Let $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ and $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. Define $Y \in \mathbb{R}^{n\times q}$ to have ith row $(y_i - \bar{y})'$ and define $X \in \mathbb{R}^{n\times p}$ to have ith row $(x_i - \bar{x})'$. Suppose that the ϵ_i's are q-variate normal and $\Sigma_* = \sigma_{*1}^2\beta_*'\beta_* + \sigma_{*2}^2 I_q$, where σ_{*1}^2 and σ_{*2}^2 are positive constants that represent the proportionality assumption in (2). Then two times the negative log likelihood (up to constants) evaluated at $(\beta, \sigma_1^2, \sigma_2^2)$ is

$$\operatorname{tr}\left\{n^{-1}(Y - X\beta)(\sigma_1^2\beta'\beta + \sigma_2^2 I_q)^{-1}(Y - X\beta)'\right\} + \log\det\left(\sigma_1^2\beta'\beta + \sigma_2^2 I_q\right), \tag{3}$$

where tr and $\det$ are the trace and determinant. A likelihood-based estimator of β_* would require minimization over three optimization variables: β, σ_1^2, and σ_2^2. The scaling factor optimization variable σ_1^2 makes the function defined by (3) difficult to minimize because it scales $\beta'\beta$ in both the trace and determinant terms. This optimization is made more difficult when one penalizes the entries of β. Moreover, since our goal is to estimate β_*, ideally, estimation of nuisance parameters σ_{*1}^2 and σ_{*2}^2 could be avoided entirely. Additional details about maximum likelihood estimation using (3) are given in Section 4 of the Supplementary Material.

Instead, to estimate β_* with the assumption in (2), we propose the class of estimators

$$\widehat{\beta}_{\tau,\lambda} = \underset{\beta\in\mathbb{R}^{p\times q}}{\arg\min}\left\{\mathcal{F}_\tau(\beta) + \frac{\lambda}{\tau}\mathrm{Pen}(\beta)\right\}, \tag{4}$$

where $\mathrm{Pen}(\cdot)$ is a user-specified penalty function; τ and λ are positive tuning parameters; and

$$\mathcal{F}_\tau(\beta) = \operatorname{tr}\left\{n^{-1}(Y - X\beta)(\beta'\beta + \tau I_q)^{-1}(Y - X\beta)'\right\}. \tag{5}$$

The function defined in (5) is similar to (3), except that the scaling factor σ_1^2 and the log determinant term are removed. This function generalizes the function proposed in Gleser and Watson (1973), which we discuss further in Section 3.

We do not require a particular form for $\mathrm{Pen}(\cdot)$, but the algorithm we propose in Section 4 to solve (4) will be most effective when the proximal operator of $\mathrm{Pen}(\cdot)$ can be computed efficiently. In addition, a global minimizer for (4) is only guaranteed to exist when $\mathrm{Pen}(\cdot)$ is coercive (see Remark 1 in Section 4.1).

The function $\mathcal{F}_\tau$ is especially flexible due to the tuning parameter τ. In particular, when $\tau \to \infty$, the matrix $\beta'\beta + \tau I_q$ becomes diagonally dominant and so $\mathcal{F}_\tau$ tends to the unweighted residual sum of squares. This has two main benefits: statistically, the theoretical properties of any penalized least squares estimator apply to (4) by allowing $\tau \to \infty$ at a sufficient rate; practically, this gives practitioners the ability to determine whether, and to what extent, (2) holds in a data-driven fashion. In particular, if (2) does not hold, then our tuning parameter selection criterion, described in Section 1 of the Supplementary Material, should select τ sufficiently large so that $\widehat{\beta}_{\tau,\lambda}$ is effectively the same as the penalized least squares estimator. Viewed in this way, the tuning parameter τ is best thought of as analogous to the tuning parameter needed to specify the Huber loss function (Huber, 1964).

We do not intend that τ be interpreted as a ratio of unknown error variances. Were inference about β_* or variance parameters in (3) the goal of the practitioner, our method may not be appropriate since we treat the variances as nuisance parameters.

3 Models connected to the parametric link

3.1 Errors-in-variables models

The parameterization in (2) holds under a multivariate response "errors-in-variables" linear regression model. Consider the special case of (1) where for a latent (non-random) predictor $z_i \in \mathbb{R}^p$ for the ith subject, y_i is assumed to be a realization of

$$\mathbf{Y}_i = \mu_* + \beta_*' z_i + \widetilde{\epsilon}_i,$$

where the $\widetilde{\epsilon}_i$'s are independent and identically distributed with $\mathrm{E}(\widetilde{\epsilon}_i) = 0$, $\mathrm{Cov}(\widetilde{\epsilon}_i) = \gamma_*^2 I_q$, for $i = 1, \ldots, n$. Suppose we cannot measure z_i exactly for $i = 1, \ldots, n$. Instead, we observe a realization of $\mathbf{X}_i = z_i + \mathbf{U}_i$ where the $\mathbf{U}_i$'s are independent and identically distributed with $\mathrm{E}(\mathbf{U}_i) = 0$ and $\mathrm{Cov}(\mathbf{U}_i) = \sigma_{*u}^2 I_p$ for $i = 1, \ldots, n$; and $\mathbf{U}_i$ is independent of $\widetilde{\epsilon}_i$. It follows that

$$\mathbf{Y}_i = \mu_* + \beta_*' z_i + \widetilde{\epsilon}_i = \mu_* + \beta_*' \mathbf{X}_i - \beta_*' \mathbf{U}_i + \widetilde{\epsilon}_i.$$

Because we do not observe the realization of $\mathbf{U}_i$,

$$\mathrm{Cov}(\mathbf{Y}_i \mid \mathbf{X}_i = x_i) = \beta_*' \mathrm{Cov}(\mathbf{U}_i) \beta_* + \mathrm{Cov}(\widetilde{\epsilon}) \propto \beta_*' \beta_* + \sigma_{*\epsilon}^2 I_q,$$

where $\sigma_{*\epsilon}^2 = \gamma_*^2 / \sigma_{*u}^2$.

Fitting errors-in-variables models is a classical problem in low-dimensional multivariate statistics. If one were willing to make distributional assumptions (e.g., normality) about the $\widetilde{\epsilon}_i$ and $\mathbf{U}_i$, then one could obtain maximum likelihood estimators by maximizing the joint likelihood for $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ and treating the z_i's as unknown parameters. Alternatively, one could fit the model that is conditional on the observed values of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Gleser and Watson (1973) established an interesting connection between these two approaches in the low-dimensional case. In particular, Gleser and Watson (1973) showed that in the special case where $\sigma_{*u}^2 = \gamma_*^2$ and $q = p$, the estimator obtained by maximizing the log-likelihood for the joint distribution of the predictor and response was equivalent to the estimate obtained by maximizing the weighted residual sum of squares:

$$\mathrm{tr}\left\{ n^{-1} (Y - X\beta)(\beta'\beta + I_q)^{-1} (Y - X\beta)' \right\}, \tag{6}$$

which is similar to the negative log-likelihood when $\mathbf{Y}_i \mid \mathbf{X}_i = x_i$ is multivariate normal. However, when σ_{*u}^2 and γ_*^2 are unequal, (6) may perform poorly.

Like (6), our proposed weighted residual sum of squares criterion defined in (5) is similar to

the negative-log likelihood when one assumes ϵ_i's are multivariate normal. Unlike the function in (6), our proposed criterion defined in (5) replaces $\beta'\beta + I_q$ with $\beta'\beta + \tau I_q$. The introduction of the tuning parameter τ allows practitioners to account for the relationship between γ_*^2 and σ_{*u}^2 using cross-validation.

Although the proposed criterion $\mathcal{F}_\tau$ is similar to the normal negative log-likelihood, it is not a valid likelihood function. However, we can justify its use based on the following result:

Theorem 1. *Suppose* (Y, X) *are generated from the multivariate response errors-in-variables model. If* $\tau = \gamma_*^2/\sigma_{*u}^2$*, then* $\mathrm{E}\{\nabla \mathcal{F}_\tau(\beta_*)\} = 0$*, that is, our estimator is Fisher consistent.*

We prove Theorem 1 in the Section 2 of the Supplementary Material. Note that we assume no particular distribution for the $\mathbf{U}_i$ or $\widetilde{\epsilon}_i$.

In Section 4 of the Supplementary Material, we compare estimates based on $\mathcal{F}_\tau$ to the maximum likelihood estimator (MLE) (based on (3)) in low-dimensional settings. We show that with the tuning parameter τ chosen by cross-validation, the unpenalized version of (4) performs similarly to the MLE under various data generating models. This provides some evidence that little efficiency is lost using $\mathcal{F}_\tau$ as an estimation criterion relative to the negative log-likelihood.

3.2 Latent variable reduced-rank regression model

Our parameterization also arises from a particular latent variable reduced-rank regression model (Velu and Reinsel, 2013). This model assumes that the measured response for the ith subject is a realization of

$$\mathbf{Y}_i = \mu_* + \mathcal{A}_* \mathbf{Z}_i + \widetilde{\epsilon}_i, \quad (i = 1, \ldots, n), \tag{7}$$

where $\mathcal{A}_* \in \mathbb{R}^{q \times r}$ with $r \leq \min(p, q)$, and $\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_n$ are independent and identically distributed with mean zero and covariance $\gamma_*^2 I_q$. In addition,

$$\mathbf{Z}_i = \mathcal{B}_* x_i + \mathbf{U}_i, \quad (i = 1, \ldots, n), \tag{8}$$

where $\mathcal{B}_* \in \mathbb{R}^{r\times p}$ is a semiorthogonal matrix, the $x_i \in \mathbb{R}^p$ are the nonrandom values of the predictor for the ith subject, and $\mathbf{U}_1, \ldots, \mathbf{U}_n$ are independent and identically distributed with mean zero and covariance $\sigma^2_{*u} I_r$. It follows that

$$\mathrm{E}(\mathbf{Y}_i) = \mu_* + \mathcal{A}_*\mathcal{B}_* x_i, \quad \mathrm{Cov}(\mathbf{Y}_i) = \sigma^2_{*u}\mathcal{A}_*\mathcal{A}_*' + \gamma^2_* I_q,$$

so that the regression coefficient matrix is $\beta_* = \mathcal{B}_*'\mathcal{A}_*' \in \mathbb{R}^{p\times q}$ with $\mathrm{rank}(\beta_*) = r$. It is straightforward to verify that together, (7) and (8) imply the mean-covariance parameterization in (2).

4 Computation

Although $\mathcal{F}_\tau$ is not convex, it is differentiable and has Lipschitz continuous gradient over bounded sets. We formalize these properties in the following proposition.

Proposition 1. *When $\tau > 0$,*

$$\nabla\mathcal{F}_\tau(\beta) = -2n^{-1}\beta\Omega_\beta^{-1}(Y - X\beta)'(Y - X\beta)\Omega_\beta^{-1} - 2n^{-1}X'Y\Omega_\beta^{-1} + 2n^{-1}X'X\beta\Omega_\beta^{-1},$$

where $\Omega_\beta = \beta'\beta + \tau I_q$. Moreover, $\nabla\mathcal{F}_\tau$ is Lipschitz over the set $\mathcal{D}_\kappa = \{\beta : \beta \in \mathbb{R}^{p\times q}, \|\beta\|_F \le \kappa\}$ where $0 \le \kappa < \infty$, where $\|\cdot\|_F$ is the Frobenius norm.

We prove both parts of Proposition 1 in Section 2 of the Supplementary Material.

Remark 1. *Because* (5) *is bounded below (since the trace of the product of two non-negative definite matrices is non-negative), as long as the penalty function is coercive, i.e.,* $\mathrm{Pen}(\beta) \to \infty$ *as* $\|\beta\|_F \to \infty$, *a global minimizer of* (4) *over* $\mathbb{R}^{p\times q}$ *exists and is in* $\mathcal{D}_\kappa$ *for some finite* κ.

Given the properties established in Proposition 1 and Remark 1, we can use a proximal gradient descent algorithm to obtain a critical point of (4) (Li and Lin, 2015). Since $\mathcal{F}_\tau$ has a Lipschitz

continuous gradient over the bounded set $\mathcal{D}_\kappa$, there exists a positive constant L such that

$$\mathcal{F}_\tau(\beta) \leq \mathcal{F}_\tau(\widetilde{\beta}) + \operatorname{tr}\left\{\nabla\mathcal{F}_\tau(\widetilde{\beta})'(\beta - \widetilde{\beta})\right\} + \frac{L}{2}\|\beta - \widetilde{\beta}\|_F^2, \tag{9}$$

for all $\beta \in \mathcal{D}_\kappa$ and $\widetilde{\beta} \in \mathcal{D}_\kappa$. Thus, the right hand side of (9) is a *majorizing function* of $\mathcal{F}_\tau$ at $\widetilde{\beta}$ (i.e., the right hand side of (9) is greater than or equal to $\mathcal{F}_\tau$ for all $\beta \in \mathcal{D}_\kappa$ with equality when $\beta = \widetilde{\beta}$). Hence, applying the majorize-minimize principle (Lange, 2016), we use an algorithm whose iterates minimize the majorizing function at the previous iterate:

$$\beta^{(k+1)} = \underset{\beta\in\mathbb{R}^{p\times q}}{\arg\min}\left\{\mathcal{F}_\tau(\beta^{(k)}) + \operatorname{tr}\left\{\nabla\mathcal{F}_\tau(\beta^{(k)})'(\beta - \beta^{(k)})\right\} + \frac{t_k}{2}\|\beta - \beta^{(k)}\|_F^2 + \frac{\lambda}{\tau}\mathrm{Pen}(\beta)\right\}, \tag{10}$$

where t_k is a positive step-size parameter; and $\beta^{(k+1)}$ and $\beta^{(k)}$ are the $(k+1)$th and kth iterates of the optimization variable corresponding to β, respectively. This way, for sufficiently large t_k, we are guaranteed that $\mathcal{F}_\tau(\beta^{(k+1)}) + \frac{\lambda}{\tau}\mathrm{Pen}(\beta^{(k+1)}) \leq \mathcal{F}_\tau(\beta^{(k)}) + \frac{\lambda}{\tau}\mathrm{Pen}(\beta^{(k)})$ for all k.

The iterate in (10) can be written in the more familiar notation:

$$\beta^{(k+1)} = \mathrm{Prox}_{t_k^{-1}\frac{\lambda}{\tau}\mathrm{Pen}}\left\{\beta^{(k)} - t_k^{-1}\nabla\mathcal{F}_\tau(\beta^{(k)})\right\},$$

where, using the notation from Parikh et al. (2014), Prox_f denotes the proximal operator of the function f:

$$\mathrm{Prox}_f(y) = \underset{x}{\arg\min}\left\{\frac{1}{2}\|x - y\|_F^2 + f(x)\right\}.$$

The proximal operator can be computed efficiently for a broad class of convex and non-convex penalty functions. In Table 1 of the Supplementary Material, we provide closed form solutions of four proximal operators corresponding to convex penalties used in multivariate response linear regression. For example, if one used the L_1 norm as a penalty, the proximal operator is simply the soft-thresholding operator.

With an appropriate choice of step size parameter t_k for each k, iterates generated from (10) are guaranteed to monotonically decrease the objective function value. However, this is not sufficient to ensure that the iterates converge to a critical point. In our implementation, we use an accelerated variation of the proximal gradient descent algorithm proposed by Li and Lin (2015) specifically designed for solving non-convex optimization problems which ensures that iterates converge to a critical point of (4).

The complete algorithm is sketched in Algorithm 1. We implement this algorithm, along with a number of auxiliary functions, in the the R package `MCMVR`, which is available for download at github.com/ajmolstad/MCMVR.

Algorithm 1: Initialize $\beta^{(0)} = \beta^{(-1)} = \widetilde{\beta}^{(0)} = \overline{\beta}^{(0)}$, $\alpha^{(0)} = \alpha^{(-1)} = 1$, and set $k = 0$.

1. Compute $\widetilde{\beta}^{(k)} \leftarrow \beta^{(k)} + \frac{\alpha^{(k-1)}}{\alpha^{(k)}}\left(\overline{\beta}^{(k)} - \beta^{(k)}\right) + \left(\frac{\alpha^{(k-1)}-1}{\alpha^{(k)}}\right)\left(\beta^{(k)} - \beta^{(k-1)}\right)$.
2. Compute $\overline{\beta}^{(k+1)} \leftarrow \text{Prox}_{\bar{t}_k^{-1}\frac{\lambda}{\tau}\text{Pen}}\left\{\widetilde{\beta}^{(k)} - \bar{t}_k^{-1}\nabla_\beta \mathcal{F}_\tau(\widetilde{\beta}^{(k)})\right\}$.
3. Compute $\Gamma^{(k+1)} \leftarrow \text{Prox}_{t_k^{-1}\frac{\lambda}{\tau}\text{Pen}}\left\{\beta^{(k)} - t_k^{-1}\nabla_\beta \mathcal{F}_\tau(\beta^{(k)})\right\}$.
4. Set $\alpha^{(k+1)} \leftarrow (1 + \sqrt{1 + 4\alpha^{2(k)}})/2$.
5. Set $\beta^{(k+1)} \leftarrow \begin{cases} \overline{\beta}^{(k+1)} : & \mathcal{F}_\tau(\overline{\beta}^{(k+1)}) + \frac{\lambda}{\tau}\text{Pen}(\overline{\beta}^{(k+1)}) < \mathcal{F}_\tau(\Gamma^{(k+1)}) + \frac{\lambda}{\tau}\text{Pen}(\Gamma^{(k+1)}) \\ \Gamma^{(k+1)} : & \text{otherwise} \end{cases}$
6. Set $k \leftarrow k + 1$, and return to Step 1.

Step sizes t_k and $\bar{t}_k$ from Algorithm 1 are chosen using backtracking line searches for both Step 2 and Step 3 of Algorithm 1. See Algorithm 2 of the Supplementary Material to Li and Lin (2015) for the exact version of the algorithm we implement. An application of Theorem 1 of Li and Lin (2015) ensures that the iterates generated by our algorithm are bounded and that the sequence of iterates converge to a critical point of (4). In Section B of the Supplementary Material, we describe how to select tuning parameters by cross-validation, and how to determine a reasonable set of candidate tuning parameter values.

5 Generalizations

In Section 3 of the Supplementary Material, we discuss an extension of our method to settings where only a subset of the response vectors are observed. In the remainder of this section, we describe how to modify our method to deal with covariates unrelated to Σ_* and how to standardize predictors for model fitting.

5.1 Covariates unrelated to Σ_*

An important generalization of our estimator includes a set of measured covariates $v_i \in \mathbb{R}^k$ such that

$$\mathbf{Y}_i = \mu_* + \beta_*' x_i + \eta_*' v_i + \epsilon_i, \quad (i = 1, \ldots, n),$$

and the unknown coefficient matrix $\eta_* \in \mathbb{R}^{k \times q}$ is not parametrically related to $\Sigma_* \propto \beta_*'\beta_* + \sigma_{*\epsilon}^2 I_q$. When our method is motivated through the errors-in-variables model, this may occur when some covariates or confounders are measured without error, e.g., x_i is some -omic profile measured with error and v_i are clinical/demographic variables.

Let $\bar{v} = n^{-1}\sum_{i=1}^n v_i$. Define $V \in \mathbb{R}^{n \times k}$ to have ith row $(v_i - \bar{v})'$ and suppose V has full rank. For this scenario, we propose the class of penalized estimators:

$$(\tilde{\beta}_{\tau,\lambda}, \tilde{\eta}_{\tau,\lambda}) = \underset{\beta \in \mathbb{R}^{p\times q}, \eta \in \mathbb{R}^{k\times q}}{\arg\min} \left[\operatorname{tr}\left\{ n^{-1}(Y - V\eta - X\beta)'(Y - V\eta - X\beta)[\beta'\beta + \tau I_q]^{-1} \right\} + \frac{\lambda}{\tau}\mathrm{Pen}(\beta) \right].$$

Using the first order conditions for η and letting $\mathrm{P}_V = I_n - V(V'V)^{-1}V'$, we replace Y with $\tilde{Y} = \mathrm{P}_V Y$, X with $\tilde{X} = \mathrm{P}_V X$, and solve a modified version of our estimator:

$$\tilde{\beta}_{\tau,\lambda} = \underset{\beta \in \mathbb{R}^{p\times q}}{\arg\min} \left[\operatorname{tr}\left\{ n^{-1}(\tilde{Y} - \tilde{X}\beta)'(\tilde{Y} - \tilde{X}\beta)[\beta'\beta + \tau I_q]^{-1} \right\} + \frac{\lambda}{\tau}\mathrm{Pen}(\beta) \right],$$

so that $\tilde{\eta}_{\tau,\lambda} = (V'V)^{-1}V'(Y - X\tilde{\beta}_{\tau,\lambda})$. Thus computing our estimator with the additional covariates

can be done immediately from Algorithm 1 and $(\tilde{\beta}_{\tau,\lambda}, \tilde{\eta}_{\tau,\lambda})$ will satisfy the first order conditions for $(\tilde{\beta}_{\tau,\lambda}, \tilde{\eta}_{\tau,\lambda})$.

5.2 Predictor standardization and dependent measurement errors

An important property in regression coefficient matrix estimation is invariance under changes in scale of the predictor. Of course, our method is not invariant as the scale of the predictors affects the magnitude of entries in β_*, which affects the weight in the weighted residual sum of squares criterion we propose. However, we can easily generalize our estimator to allow for standardization, and in the context of the errors-in-variables model, allow for dependent measurement errors (assuming their covariance were known).

The generalized version of $\mathcal{F}_\tau$ we propose is

$$\mathcal{H}_{\tau,\Phi}(\beta) = \left\{ n^{-1}(Y - X\beta)'(Y - X\beta) \left[\beta'\Phi\beta + \tau I_q\right]^{-1} \right\},$$

where $\Phi \in \mathbb{R}^{p\times p}$ is some user-specified, symmetric and nonnegative definite weight matrix. Notice, were one to standardize predictors so that columns of $\tilde{X}$ had columnwise average zero and unit standard deviation, we could write $X\beta = \tilde{X}\tilde{\beta}$ where $\tilde{\beta} = S^{-1}\beta$ and $S \in \mathbb{R}^{p\times p}$ has the inverse standard deviations of the predictors on its diagonal and zeros elsewhere. Thus, if $\Phi = S'S$, it follows that

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

where $\beta = S\tilde{\beta}$.

If the model from Section 3.1 were assumed to hold, then this generalization could also be used in the case that measurement errors (i.e., the $\mathbf{U}_i$ from Section 3.2) have covariance Σ_{*u}, which is known or can be estimated reliably from external data. In this case, it would follow that $\operatorname{Cov}(\mathbf{Y}_i \mid \mathbf{X}_i = x_i) = \beta_*'\Sigma_{*u}\beta_* + \gamma_*^2 I_q$, so that a more appropriate weighted residual sum of

squares would be $\mathcal{H}_{\tau,\Sigma_{*u}}$.

The same computational approach developed in Section 4 of the main manuscript can be used since $\mathcal{H}_{\tau,\Phi}$ is differentiable and has Lipschitz continuous gradient over $\mathcal{D}_\kappa$.

6 Simulation studies

6.1 Data generating models and performance metrics

We compare the performance of our method to relevant competitors under three distinct data generating models. Under the first two models, when conditioning on the observed predictors, the mean and covariance of the response are the same in both models. However, the two models differ in a fundamental way: in the first model, we observe a corrupted version of the "true" predictor so that conditioning on the observed predictor, the model in (1) and (2) holds. In the second model, we observe the "true" predictor and the covariance has the parameterization in (2). In the third model we consider, there are "errors-in-variables" but the covariance parameterization (2) does not hold.

In the following, for one hundred independent replications with $p = 200$ and $q = 50$, we generate $n = 100$ independent copies of $(\mathbf{Y}, \mathbf{X})$.

– **Model 1:** We first generate n independent copies of $\mathbf{Z} \sim \mathrm{N}_p(0, \Sigma_{*Z})$ where the (j,k)th entry of Σ_{*Z} equals $0.5^{|j-k|}$. Then, conditional on $\mathbf{Z} = z$, we generate a realization of $\mathbf{X}$ and $\mathbf{Y}$,

$$\mathbf{Y} = \beta_*' z + \epsilon, \quad \mathbf{X} = z + \mathbf{U},$$

where $\mathbf{U} \sim \mathrm{N}_p(0, \sigma_{*u}^2 I_p)$ and $\epsilon \sim \mathrm{N}_p(0, \gamma_*^2 I_q)$ so that $\mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x) = \beta_*' x$ and $\mathrm{Cov}(\mathbf{Y} \mid \mathbf{X} = x) = \sigma_{*u}^2 \beta_*' \beta_* + \gamma_*^2 I_q$, with $\gamma_*^2 = 3$ and σ_{*u}^2 varying across settings.

– **Model 2:** 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

$$\mathbf{Y} = \beta_*' x + \epsilon, \tag{11}$$

where $\epsilon \sim \mathrm{N}_q(0, \sigma_{*u}^2 \beta_*' \beta_* + \gamma_*^2 I_q)$ with $\gamma_*^2 = 3$ and σ_{*u}^2 varying. Note that Model 2 differs from Model 1 as under Model 1, the covariance of the measured predictors $\mathbf{X}$ is $\Sigma_{*X} = \Sigma_{*Z} + \sigma_{*u}^2 I_p$.

– **Model 3:** We first generate data in the same manner as Model 1 except where $\mathbf{U} \sim \mathrm{N}_p(0, \sigma_{*u}^2 I_p)$ and $\epsilon \sim \mathrm{N}_p(0, \gamma_*^2 \Sigma_{*E})$ where $[\Sigma_{*E}]_{j,k} = 0.7 \cdot \mathbf{1}(j \neq k) + \mathbf{1}(j = k)$, where $\sigma_{*u}^2 = 0.50$, and where γ_*^2 is varying across settings.

To generate β_*, we randomly construct three active sets of three variables each: let $a_k = \{a_{k,1}, a_{k,2}, a_{k,3}\} \subset \{1, \ldots, p\}$ for $k = 1, 2, 3$ with $\cap_{k=1}^3 a_k$ being empty. Then, for $l = 1, \ldots, q$, we randomly choose $k \in \{1, 2, 3\}$ with probability $1/3$ each, and set either $[\beta_*]_{(a_{k,j}),l} = -2$ or $[\beta_*]_{(a_{k,j}),l} = 2$, with equal probability for $j = 1, \ldots, 3$. We also select three additional elements of $[\beta_*]_{\cdot,l}$ to be -1 or 1. That is, each column of β_* has six nonzero entries: three entries have magnitudes 2 and three entries have magnitudes 1. Under this construction, $\beta_*' \beta_*$ is approximately block diagonal with three blocks of similar size.

We consider multiple performance metrics. The first we consider is *model error* (Breiman and Friedman, 1997) for the observed predictor: $\|\Sigma_{*X}^{1/2}(\widehat{\beta} - \beta_*)\|_F^2$. Following Datta and Zou (2017), when data are generated from Model 2, we also measure the *latent model error*, i.e, model error under the unobserved predictor Z: $\|\Sigma_{*Z}^{1/2}(\widehat{\beta} - \beta_*)\|_F^2$. Latent model error would be relevant if the true predictor may be observed in future studies. In addition, for both models we also measure (squared) Frobenius norm error: $\|\widehat{\beta} - \beta_*\|_F^2$, and out-of-sample prediction error: $\|Y_T - X_T \widehat{\beta}\|_F^2 / q n_T$. To compute the out-of-sample prediction error, in each replication we generate an independent test set of size $n_T = 1000$, where $Y_T \in \mathbb{R}^{n_T \times q}$ and $X_T \in \mathbb{R}^{n_T \times p}$, using the same data generating model as in the training data. It is important to note that out-of-sample prediction error and model error are distinct metrics. Model error measures how well an estimator predicts the mean function, whereas

prediction error measures sum of squared residuals on a testing set. We also measure true and false positive identification of nonzero entries in β_* to assess variable selection accuracy of the methods.

6.2 Competing methods

We compare our method to two versions of the method proposed by Datta and Zou (2017), two versions of the L_1-penalized least squares estimator, and a two-step convex approximation to (4). Throughout, let $|A|_1 = \sum_{j,k} |A_{j,k}|$ for a matrix or vector A, and let $A_{\cdot,j}$ denote the jth column of A.

For the case that $q = 1$ and data are generated from an errors-in-variables model (e.g., Model 2), Datta and Zou (2017) proposed the convex-conditioned lasso estimator. Their estimator can be naturally extended to the multivariate setting: they replace X and Y in the least squares criterion with versions adjusted to account for the measurement error. Assuming that σ^2_{*u} were known, the estimator of Datta and Zou (2017) modifies the unbiased sample covariance matrix:

$$\widetilde{\Sigma} = \underset{S_1 \geq 0}{\arg\min} \|\widetilde{S} - S_1\|_{\max}, \quad \text{where } \widetilde{S} = n^{-1}X'X - \sigma^2_{*u} I_p. \tag{12}$$

Then the multivariate response generalization of their estimator is

$$\underset{\beta \in \mathbb{R}^{p \times q}}{\arg\min} \left\{ \mathrm{tr}\left(\beta' \widetilde{\Sigma} \beta - 2\beta' \rho \right) + \lambda \mathrm{Pen}(\beta) \right\}, \tag{13}$$

where $\rho = n^{-1}X'Y$, which can be solved using penalized least squares. When $\lambda \mathrm{Pen}(\beta)$ is replaced by $\sum_{j=1}^{q} \lambda_j \mathrm{Pen}(\beta_{\cdot,j})$, the estimator in (13) is equivalent to performing q separate convex-conditioned lasso estimation problems. The estimator in (13) would not be equivalent to q separate estimators if only one tuning parameter were used for all q regressions, or any of the penalties from Table 1 of the Supplementary Material other than the L_1 norm was used. We now formally state the competitors we consider:

– `CoCo-1`: The estimator defined in (13) with $\lambda \mathrm{Pen}(\beta) = \lambda|\beta|_1$ and λ chosen by five-fold cross-validation, using the modified cross-validation procedure (averaged over the q responses) proposed in Datta and Zou (2017). We treat the value of σ^2_{*u} as known.

– `CoCo-q`: The estimator defined in (13) with $\lambda \mathrm{Pen}(\beta)$ replaced by $\sum_{j=1}^{q} \lambda_j |\beta_{\cdot,j}|_1$ and the λ_j each chosen by five-fold cross-validation, using the modified cross-validation procedure proposed in Datta and Zou (2017) for $j = 1, \ldots, q$. We treat the value of σ^2_{*u} as known.

– `CV-CoCo-q`, `CV-CoCo-1`: The same estimators as `CoCo-q` and `CoCo-1`, except σ^2_{*u} is unknown and treated as a tuning parameter. We select both λ and the σ^2_{*u} value by five-fold cross-validation, using the modified cross-validation procedure proposed in Datta and Zou (2017).

– `Lasso-q`: The L_1-penalized least squares estimator

$$\underset{\beta \in \mathbb{R}^{p \times q}}{\arg\min} \left\{ \frac{1}{n} \|Y - X\beta\|_F^2 + \sum_{j=1}^{q} \lambda_j |\beta_{\cdot,j}| \right\} \tag{14}$$

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

– `Lasso-1`: The estimator defined in (14) 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.

– `MC`: The version of our estimator (4) with $\mathrm{Pen}(\beta) = |\beta|_1$, with tuning parameters λ and τ chosen using the five fold cross-validation procedure described in Section 1 of the Supplementary Material.

The sixth competitor we consider, `CA`, is a two-step convex approximation to (4). Given a initial estimator $\widetilde{\beta}$, we re-estimate β_* using

$$\overline{\beta} = \underset{\beta \in \mathbb{R}^{p \times q}}{\arg\min} \left[\mathrm{tr} \left\{ n^{-1}(Y - X\beta)(\widetilde{\beta}'\widetilde{\beta} + \tau I_q)^{-1}(Y - X\beta)' \right\} + \frac{\lambda}{\tau}|\beta|_1 \right]. \tag{15}$$

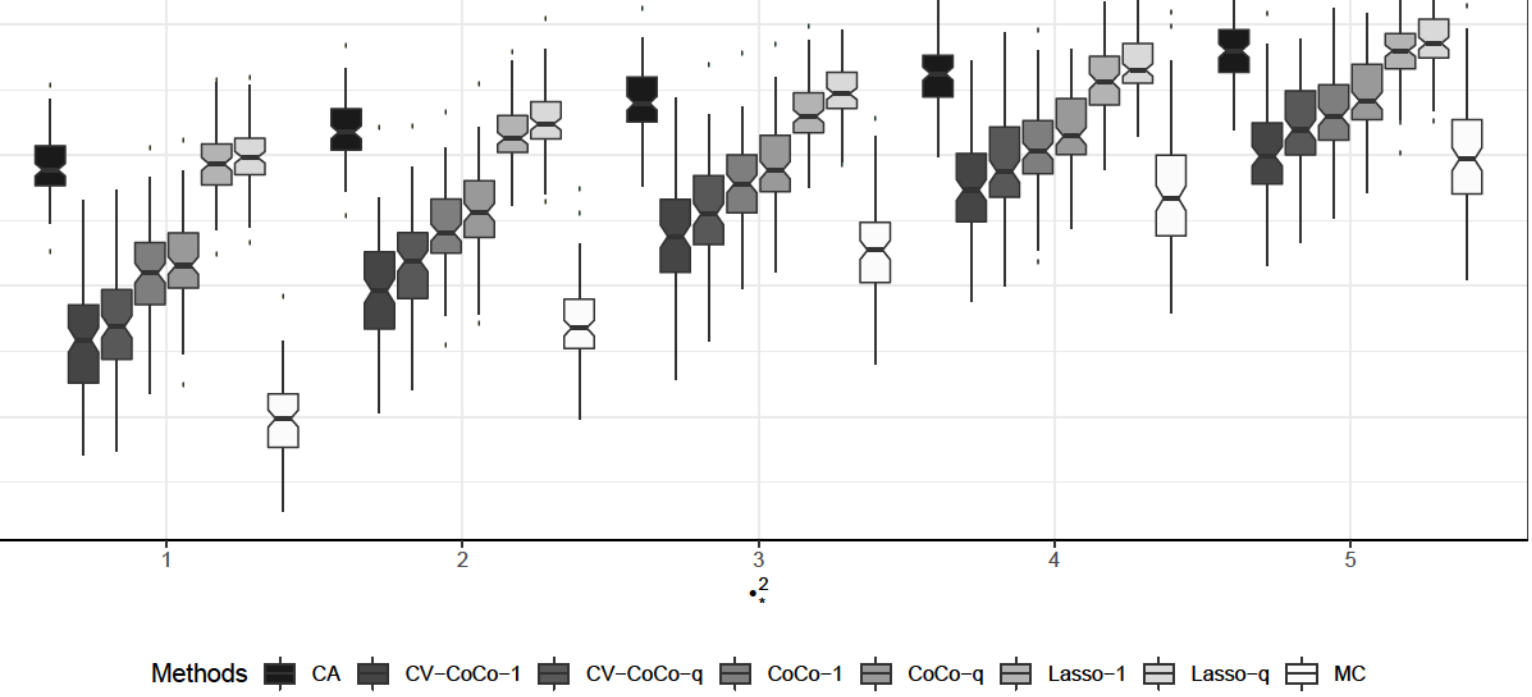

Figure 1: Log model error, log latent model error, and log prediction error for the eight candidate methods over one hundred independent replications under Model 1.

The estimator defined in (15) is computed using the coordinate descent algorithm of Rothman et al. (2010).

In our implementation of the estimator `CA`, we obtain $\widetilde{\beta}$ using `Lasso-q` and select the tuning parameters τ and λ to minimize prediction error in five-fold cross-validation. The estimator `CA` is included to help illustrate the importance of simultaneous estimation of the covariance matrix and regression coefficients under (2).

6.3 Results

Results for Models 1-3 are displayed in Figures 1-3, respectively. We first discuss results under Model 2, which are displayed in Figure 2. As σ_{*u}^2 increases for Model 2 we see that our proposed method, `MC`, outperforms all competitors in terms of model error, Frobenius norm error, and prediction error. Amongst the competitors, `CoCo-1` is best when $\sigma_{*u}^2 < 1$. When $\sigma_{*u}^2 = 1$, `CoCo-1` performs similarly to `Lasso-1` and the convex approximation of our method `CA`. Interestingly, we see both `Lasso-1` and `CoCo-1` outperform their counterparts which select tuning parameters separately for each response. A similar result was observed in the simulations of Molstad and

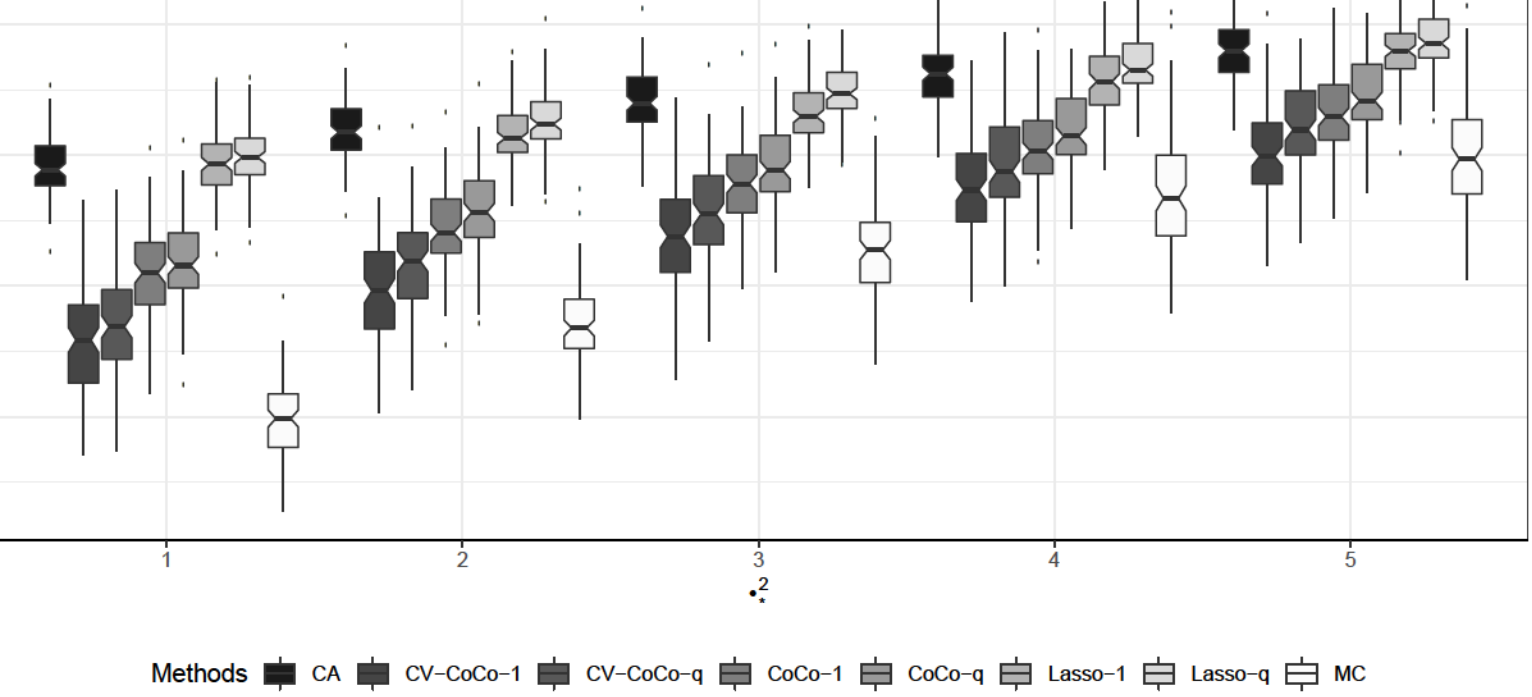

Figure 2: Log model error, log mean squared error, and log prediction error for the eight candidate methods over one hundred independent replications under Model 2.

Rothman (2016).

In Figure 1, we display results for Model 1. Unlike in Model 2, however, we see that `CV−CoCo−1` performs better than all other competitors even as σ_{*u}^2 becomes large. Our method again outperforms all competitors when σ_{*u}^2 is greater than 0.25. This is particularly notable for the latent model error, where the `CoCo` variants outperform both `Lasso` variants.

For Models 1 and 2, although not displayed, we also considered the case that $\sigma_{*u}^2 = 0$, i.e., (2) does not hold. When $\sigma_{*u}^2 = 0$ our method performed similarly to `Lasso−1`, as did the method of Datta and Zou (2017). This result illustrates the property of (5) highlighted in Section 3: when (2) does not hold, cross-validation should select a τ large enough so that (5) is effectively least squares.

In Figure 3, we display model error, latent model error, and prediction error results under Model 3. Recall that under Model 3, (2) does not hold: error correlations are not entirely determined by β_*. Nevertheless, we see that for all the values of γ_*^2 which we consider, `MC` performs best amongst the competing methods. As γ_*^2 becomes larger, the distinction between methods becomes smaller. This is intuitive given that a large γ_*^2 corresponds to a smaller signal to noise ratio.

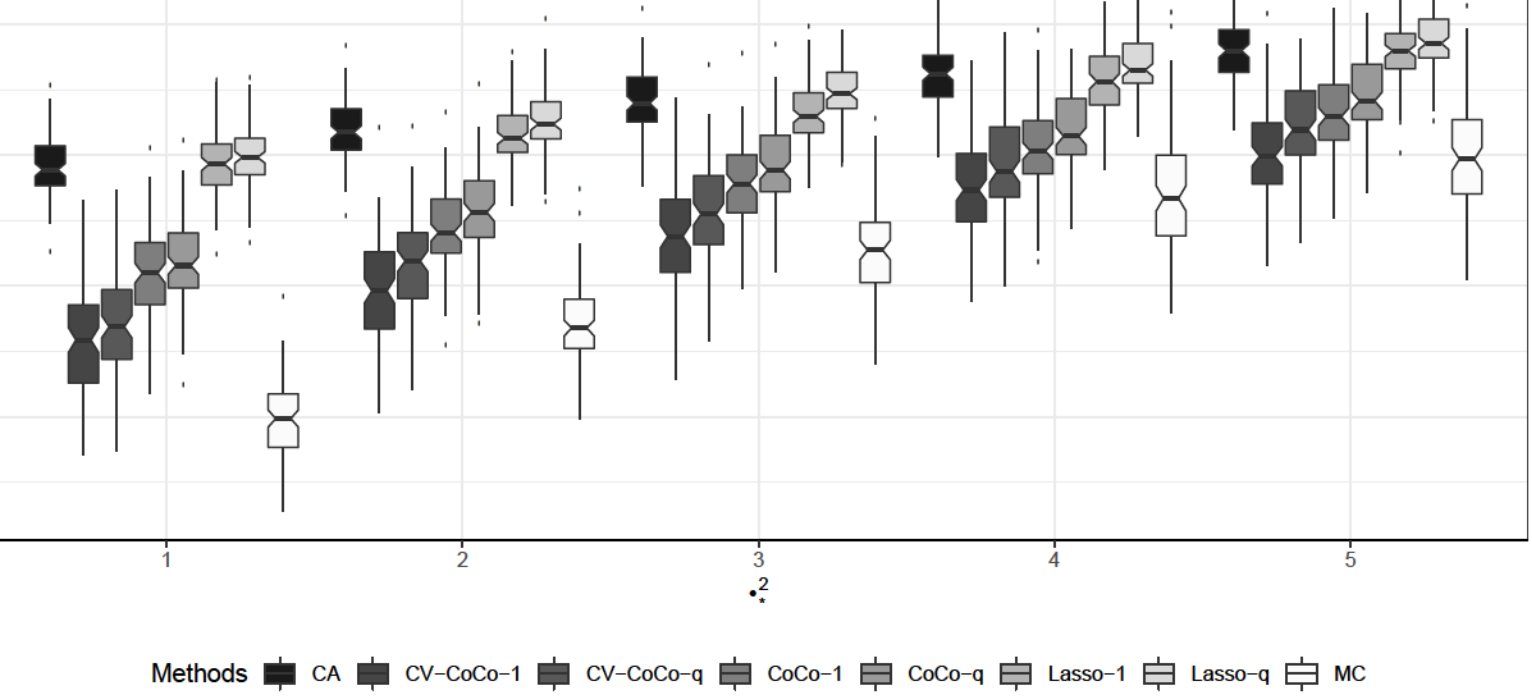

Figure 3: Log model error, log latent model error, and log prediction error for the eight candidate methods over one hundred independent replications under Model 3.

Model selection results are displayed in Table 2 and 3 of the Supplementary Material. We see that as σ_{*u}^2 increases under Model 1 and 2, the true positive rate of our method, `MC`, tends to be significantly higher than any of the competing methods. Interestingly, both `CoCo-1` and `CoCo-q` have smaller false positive rates than `MC` when $\sigma_{*u}^2 \geq 0.50$, but both have significantly smaller true positive rates than `MC`. This may partly explain the difference in performance between the `CoCo` methods and `MC`. Results are similar under Model 3.

In Section 5 of the Supplementary Material, we present results from an additional simulation study under the latent reduced-rank regression model discussed in Section 3.2. We compare two ridge regression variants, nuclear norm-penalized least squares, a nuclear norm-penalized variation of `CA`, and our proposed estimator with $\text{Pen}(\beta)$ equal to the nuclear norm of β. Results are similar to those under Model 1 and Model 2: our proposed estimator (4) outperforms the penalized least squares variants in almost every replication when $\sigma_{*u}^2 > 0$. In Section 7 of the Supplementary Material, we also present simulation study results under a version of Model 3 with $[\Sigma_{*E}]_{j,k} = 0.9^{|j-k|}$.

In Section 6 of the Supplementary Material, we consider an additional competitor, `MC-Or`,

which is the estimator (4) with $\tau = \gamma_*^2/\sigma_{*u}^2$ and λ selected by cross-validation. As was observed with the `CV-CoCo` variants, this version of our method perform substantially worse that that which treats τ as a tuning parameter.

7 Cancer drug activity data analysis

In this section, we use our method to analyze a dataset consisting of microRNA expression profiles and cancer drug activity measurements on the NCI-60 cell lines (Shoemaker, 2006). The NCI-60 cell lines are a panel of 60 human tumor cancer cell lines representing leukemia, melanoma, and numerous cancers coming from distinct tissue types: breast, central nervous system, colon, lung, prostate, ovary, and kidney.

Modelling the relationship between -omic profiles and cancer drug activity is a topic of recent interest: see Chen and Sun (2017) and references therein. The interaction between microRNA expression (predictors) and cancer drug activity (response) is especially relevant given that microRNAs are believed to play a key role in the development of many cancers as they can act as tumor suppressors or oncogenes (Peng and Croce, 2016). Previous studies have found significant correlations between microRNA expression and activity in certain cancer drugs in these particular cell lines (Liu et al., 2010).

The particular dataset we analyze is publicly available through the `FRCC` R package on CRAN (Cruz-Cano and Lee, 2014). These particular data were originally obtained from the CellMiner Database via http://discover.nci.nih.gov/cellminer). Following the analysis of Cruz-Cano and Lee (2014), we restrict our attention to $q = 15$ drugs (specifically, Topoisomerase II Inhibitors) belonging to the A118 drug dataset (http://dtp.cancer.gov). See Table 3 of Cruz-Cano and Lee (2014) for more information about the particular drugs in the A118 dataset. The microRNA expression profiles were measured on a Agilent Human microRNA Microarray and consist of $p = 365$ microRNAs which had sufficient expression in at least 10% of cell lines (Liu et al., 2010). According

to the CellMiner Database, drug activity levels are defined as 50% growth inhibition (molar concentration), and microRNA expression levels are on the log-base-2 scale.

In the analysis of Cruz-Cano and Lee (2014), the authors used regularized canonical correlation analysis (CCA) on this particular dataset to examine low dimensional linear combinations of both microRNA expression and drug activity levels. Since CCA and reduced-rank regression are closely related, we analyzed this dataset using our method with a nuclear norm penalty (Yuan et al., 2007): see the bottommost row of Table 1 in the Supplementary Material for computational details and references.

First, we performed five-fold cross-validation using the entire dataset ($n = 60$) to select tuning parameters. In Figure 4(b), we display a heatmap of the squared prediction error averaged over all 15 drug responses and five folds. Notably, the τ selected by cross-validation is relatively small ($\tau = 0.0231$). For reference, when using nuclear norm-penalized least squares, the cross-validation squared prediction error is effectively equivalent to our method when $\tau = 10^4$ (i.e., the bottom row of Figure 4(b)). In Figure 4(a), we display the nuclear norm of the estimated regression coefficient matrix as a function the tuning parameter λ with $\tau = 0.0231$ fixed. Dashed vertical lines indicate tuning parameter values where the estimated rank increases. The solid vertical line indicates the tuning parameter value which minimized squared prediction error averaged over the five folds and 15 responses. Let (τ^*, λ^*) denote the tuning parameter pair which minimized cross-validation squared prediction error. From this plot, we see that the estimated rank of the regression coefficient matrix was four in these data. When fitting the model using nuclear norm-penalized least squares, the estimated rank is only three.

To interpret our estimated regression coefficient matrix $\widehat{\beta}_{\tau^*,\lambda^*}$, we examine the response factor loadings for the four factors. That is, letting $\widehat{\beta}_{\tau^*,\lambda^*} = UDV'$ be the singular value decomposition of $\widehat{\beta}_{\tau^*,\lambda^*}$, we plot the columns of V in Figure 4(c). These can be interpreted as response factor loadings since $XUD \in \mathbb{R}^{n\times 4}$ can be interpreted as the low-dimensional microRNA expression factors, so that $V \in \mathbb{R}^{q\times 4}$ can be interpreted as their loadings (i.e., V' is the regression coefficient

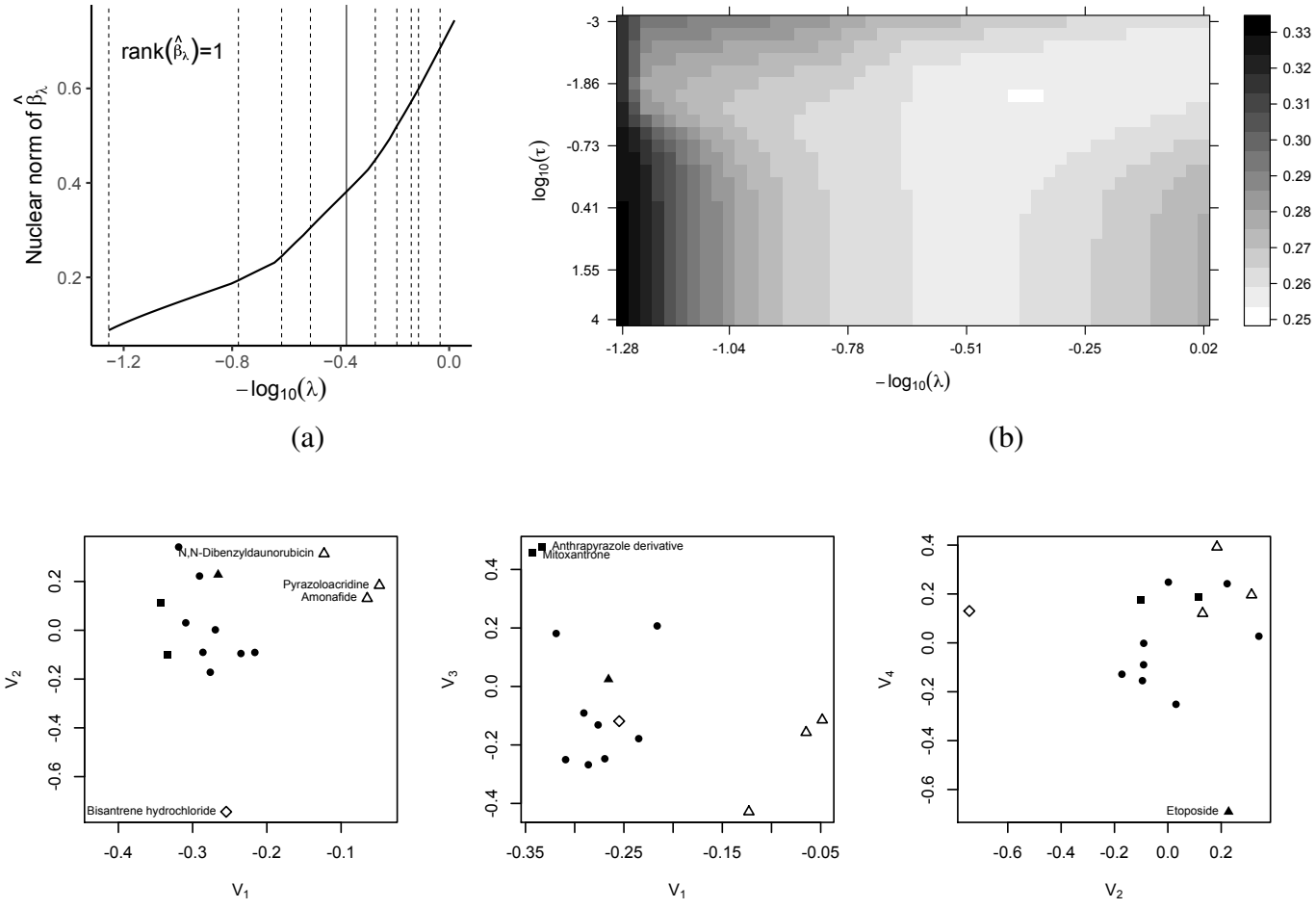

Figure 4: (a) A trace plot displaying the nuclear norm of $\widehat{\beta}_{\tau,\lambda}$ with $\tau = 0.0231$. Dashed vertical lines denote a change in rank, i.e., the dashed vertical line with $-\log_{10}(\lambda) \approx -1.5$ denotes the change from rank zero to rank one. The vertical dashed solid line denotes the value of λ which minimizes the cross-validation squared prediction error. (b) A heatmap displaying the squared prediction error averaged across five folds and all 15 response for 25 candidate values of τ and 50 candidate values of λ. (c) Plots of the responses' factor loading values from the estimated regression coefficient matrix, i.e., columns of V from the singular value decomposition $\widehat{\beta}_{\tau^*,\lambda^*} = UDV'$. Solid points denote drugs which are effective for treating leukemia; transparent triangles denote those effective for brain and spinal cancer and leukemia in mice; and the transparent diamonds denote those for effective for breast cancer. Note that symbols for each drug are the same across the three plots.

matrix for the predictors XUD). In Figure 4(c), we display three two-dimensional pairs of the response factor loadings $V_k \in \mathbb{R}^{15}, (k = 1, \ldots, 4)$. In these plots, responses represented by solid points denote drugs which are effective for treating leukemia in humans, whereas transparent points are effective for other types of cancer or leukemia in mice.

Focusing on the leftmost panel of Figure 4(c), we see that the first factor loading (V_1) separates three responses (N,N-Dibenzyldaunorubicin, Pyrazoloacridine, and Amonafide) from the rest: these three drugs are effective for treating leukemia in mice and brain cancer (see Table 3 of Cruz-Cano and Lee (2014)). In this same plot, we see that loading two (V_2) separates the drug effective for breast cancer (Bisantrene hydrochloride) from the rest. Based on the two rightmost plots, it seems factors three and four separate two (Anthrapyrazole derivative and Mitoxantrone) and one (Etoposide), respectively, of the leukemia drugs from all other drugs. These findings are mostly consistent with those in Cruz-Cano and Lee (2014) who performed CCA on these data. In Figure 3 of the Supplementary Material, we display two-dimensional plots of the XU, which show that cell lines cluster according to their cancer type.

To verify that our estimator also provides an improvement in prediction accuracy, we performed additional cross-validation. For five hundred independent replications, we randomly selected five cell lines to be testing cell lines and fit the model using nuclear norm-penalized least squares, (4) with a nuclear norm penalty, and separate ridge regressions. Tuning parameters for all methods were selected by five-fold cross-validation on the training data in each split. In Table 1, we display the average and median mean squared prediction error (over the five testing cell lines) of our method compared to nuclear norm penalized least squares, separate ridge regressions, and the null model. We observe that in all but one drug, our method provides a substantial improvement in prediction accuracy over the null model and separate ridge regressions. In the majority of drugs, our method outperforms the nuclear norm penalized least squares estimator. Notably, the estimates from our method had average rank 4.79, whereas the penalized least squares estimator had average rank 2.90.

Drug	Average MSPE				Median MSPE			
	NN−MC	NN−LS	Ridge	Null	NN−MC	NN−LS	Ridge	Null
Doxorubicin	27.60	27.59	33.35	34.07	16.42	17.41	24.41	21.24
Amonafide	4.29	4.45	5.11	4.60	3.54	3.51	3.79	3.42
M-AMSA	34.93	35.70	42.23	51.61	33.98	34.89	38.16	50.62
Anthrapyrazole derivative	34.06	34.79	40.81	47.81	29.68	30.49	36.69	42.51
Pyrazoloacridine	7.90	8.25	9.74	8.40	7.21	7.48	9.11	7.45
Bisantrene hydrochloride	41.69	41.07	46.72	42.99	17.51	17.64	21.64	19.98
Daunorubicin	26.37	26.38	34.00	35.11	20.13	20.98	29.09	31.25
Deoxydoxorubicin	27.36	27.37	33.53	33.21	20.68	20.80	26.70	20.90
Mitoxantrone	38.44	39.36	48.61	52.60	30.63	32.72	45.52	49.26
Menogaril	33.06	33.47	39.24	41.31	29.92	30.64	34.76	34.59
N,N-Dibenzyldaunorubicin	19.18	20.04	21.30	22.87	18.27	18.62	18.54	18.91
Oxanthrazole	14.94	15.05	17.92	20.29	12.84	12.89	15.83	17.78
Rubidazone	20.47	20.51	24.21	25.06	14.68	14.68	16.65	17.55
Teniposide	34.52	34.91	44.15	46.56	27.01	27.06	34.42	32.55
Etoposide	32.30	31.57	38.02	44.76	31.80	31.23	35.25	41.81

Table 1: Average and median (over 500 independent training testing splits) mean squared prediction errors ($\times 100$) for each of the fifteen drugs in the NCI-60 dataset we analyzed. Methods considered were the nuclear norm least squares estimator (`NN−LS`), the nuclear norm penalized version of (4) (`NN−MC`), fifteen separate ridge regressions (`Ridge`), and the null model, i.e., the model assuming microRNA expression does not affect mean drug activity. Cells highlighted in gray are those with the lowest MSPE amongst the considered methods.

8 Discussion

We have proposed and studied a particular parametric link between the mean and error covariance in the multivariate response linear regression model. There are multiple important directions for future research.

(a) There are many further extensions to the regression model we consider. For instance, as a referee pointed out, the decomposition of β_* into a low-rank matrix plus a sparse matrix is ubiquitous, and thus, it would be useful to be able to apply our method in such settings.

(b) We have proposed a particular parametric link between the regression function and the error covariance matrix. An alternative type of link is assumed in envelope modelling (Cook and Zhang, 2015). But there may be other link(s) that prove useful, and this could be a fruitful direction for future research to explore.

(c) Our method is based on a sum-of-squares (Frobenius) criterion for estimating the regression coefficients. However, heavy-tailed and contaminated data are often encountered in multivariate response regression applications. In such cases our criterion will not be effective. It would be useful to have alternatives for the Frobenius criterion, such as a Huberized-loss function, or an L_1 criterion function (which would connect the problem to median/quantile regression). One difficulty in those settings is that the loss function, even without the added penalty, will be either nonconvex or nondifferentiable, so a new algorithm for computing the minimizer will need to be developed.

Acknowledgments

A. J. Rothman's research was supported in part by the National Science Foundation DMS-1452068. C. R. Doss's research was supported in part by the National Science Foundation grants DMS-1712664 and DMS-1712706. The authors thank Dr. Raul Cruz-Cano for providing information about the NCI-60 dataset.

Supplementary Material

Additional information, tables, and figures referenced in Sections 2-6 are available in the Supplementary Material online. An R package implementing our method is available for download at github.com/ajmolstad/MCMVR.

References

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. Roy. Statist. Soc. Ser. B*, 59(1):3–54. With discussion and a reply by the authors.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.*, 107(500):1533–1545.

Chen, T.-H. and Sun, W. (2017). Prediction of cancer drug sensitivity using high-dimensional omic features. *Biostatistics*, 18(1):1–14.

Cook, R. D. and Zhang, X. (2015). Foundations for envelope models and methods. *J. Amer. Statist. Assoc.*, 110(510):599–611.

Cruz-Cano, R. and Lee, M.-L. T. (2014). Fast regularized canonical correlation analysis. *Computational Statistics & Data Analysis*, 70:88–100.

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

Gleser, L. J. and Watson, G. S. (1973). Estimation of a linear transformation. *Biometrika*, 60:525–534.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.*, 5:248–264.

Lange, K. (2016). *MM Optimization Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate Anal.*, 111:241–255.

Li, H. and Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 379–387. Curran Associates, Inc.

Liu, H., D'Andrade, P., Fulmer-Smentek, S., Lorenzi, P., Kohn, K. W., Weinstein, J. N., Pommier, Y., and Reinhold, W. C. (2010). mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Molecular Cancer Therapeutics*, 9(5):1080–1091.

Molstad, A. J. and Rothman, A. J. (2016). Indirect multivariate response linear regression. *Biometrika*, 103(3):595–607.

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

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

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.

Peng, Y. and Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1:15004.

Perrot-Dockès, M., Lévy-Leduc, C., Sansonnet, L., and Chiquet, J. (2018). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *J. Multivariate Anal.*, 166:78–97.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. John Wiley & Sons.

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

Shoemaker, R. H. (2006). The NCI-60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813.

Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.*, 5(4):2630–2650.

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.