Indirect multivariate response linear regression

Aaron J. Molstad  
School of Statistics  
University of Minnesota  
molst029@umn.edu

Adam J. Rothman  
School of Statistics  
University of Minnesota  
arothman@umn.edu

July 16, 2015

Abstract

We propose a new class of estimators of the multivariate response linear regression coefficient matrix that exploits the assumption that the response and predictors have a joint multivariate Normal distribution. This allows us to indirectly estimate the regression coefficient matrix through shrinkage estimation of the parameters of the inverse regression, or the conditional distribution of the predictors given the responses. We establish a convergence rate bound for estimators in our class and we study two examples. The first example estimator exploits an assumption that the inverse regression’s coefficient matrix is sparse. The second example estimator exploits an assumption that the inverse regression’s coefficient matrix is rank deficient. These estimators do not require the popular assumption that the forward regression coefficient matrix is sparse or has small Frobenius norm. Using simulation studies, we show that our example estimators outperform relevant competitors for some data generating models.

1 Introduction

Some statistical applications require the modeling of a multivariate response. Let $y_i \in \mathbb{R}^q$ be the measurement of the $q$-variate response for the $i$th subject and let $x_i \in \mathbb{R}^p$ be the nonrandom values of the $p$ predictors for the $i$th subject ($i = 1, \ldots, n$). The multivariate response linear regression model assumes that $y_i$ is a realization of the random vector

$$Y_i = \mu_i + \beta_i' x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad (1)$$

where $\mu_i \in \mathbb{R}^q$ is the unknown intercept, $\beta_i$ is the unknown $p$ by $q$ regression coefficient matrix, and $\epsilon_1, \ldots, \epsilon_n$ are independent copies of a mean zero random vector with covariance matrix $\Sigma_{i,i}$. The ordinary least squares estimator of $\beta_i$ is

$$\hat{\beta}_{(OLS)} = \arg \min_{\beta \in \mathbb{R}^{p \times q}} \|Y - X\beta\|_F^2, \quad (2)$$

where $\| \cdot \|_F$ is the Frobenius norm, $\mathbb{R}^{p \times q}$ is the set of real valued $p$ by $q$ matrices, $Y$ is the $n$ by $q$ matrix with $i$th row $(Y_i - n^{-1} \sum_{i=1}^n Y_i)'$, and $X$ is the $n$ by $p$ matrix with $i$th row $(x_i - n^{-1} \sum_{i=1}^n x_i)'$ ($i = 1, \ldots, n$). It is well known that $\hat{\beta}_{(OLS)}$ is the maximum likelihood estimator of $\beta_i$ when $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed $N_q(0, \Sigma_{i,i})$ and the corresponding maximum likelihood estimator of $\Sigma_{i,i}$ exists.
Many shrinkage estimators of $\beta_*$ have been proposed by penalizing the optimization in (2). Some of these estimators simultaneously estimate $\beta_*$ and remove irrelevant predictors (Turlach et al., 2005; Obozinski et al., 2010; Peng et al., 2010). Others encourage an estimator of reduced rank (Yuan et al., 2007; Chen and Huang, 2012).

Under the restriction that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed $N_q(0, \Sigma_*E)$, shrinkage estimators of $\beta_*$ that penalize or constrain the minimization of the negative loglikelihood have been proposed. These methods simultaneously estimate $\beta_*$ and $\Sigma_*^{-1}$. Examples include maximum likelihood reduced rank regression (Izenman, 1975; Reinsel and Velu, 1998), envelope models (Cook et al., 2010; Su and Cook, 2011, 2012, 2013), and multivariate regression with covariance estimation (Rothman et al., 2010; Lee and Liu, 2012; Bhadra and Mallick, 2013).

To fit (1) with these shrinkage estimators, one exploits explicit assumptions about $\beta_*$, but these may be unreasonable in some applications. As an alternative, we propose an indirect method to fit (1) without making explicit assumptions about $\beta_*$. We exploit the assumption that response and predictors have a joint multivariate Normal distribution and we employ shrinkage estimators of the parameters of the conditional distribution of the predictors given the response. Our method provides an alternative indirect estimator of $\beta_*$, which may be suitable when the existing shrinkage estimators are inadequate.

2 A new class of indirect estimators of $\beta_*$

2.1 Class definition

We assume that the measured predictor and response pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization of $n$ independent copies of $(X, Y)$, where $(X', Y')' \sim N_{p+q}(\mu_*, \Sigma_*)$. We also assume that $\Sigma_*$ positive definite. Define the marginal parameters through the following partitions:

$$
\mu_* = \begin{pmatrix} \mu_{*X} \\ \mu_{*Y} \end{pmatrix}, \quad \Sigma_* = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY} \end{pmatrix}.
$$

Our goal is to estimate the multivariate regression coefficient matrix $\beta_* = \Sigma_*^{-1}XX\Sigma_*XY$ in the forward regression model

$$(Y|X=x) \sim N_q(\mu_*Y + \beta_*'(x - \mu_*X), \Sigma_*E),$$

without assuming that $\beta_*$ is sparse or that $\|\beta_*\|_2^2$ is small. To do this we will estimate the inverse regression’s coefficient matrix $\eta_* = \Sigma_{YY}^{-1}\Sigma_{XY}$ and the inverse regression’s error precision matrix $\Delta_*^{-1}$ in the inverse regression model

$$(X|Y=y) \sim N_p(\mu_*X + \eta_*'(y - \mu_*Y), \Delta_*).$$

We connect the parameters of the inverse regression model to $\beta_*$ with the following proposition.

**Proposition 1.** If $\Sigma_*$ is positive definite, then

$$
\beta_* = \Delta_*^{-1}\eta_*'(\Sigma_{YY}^{-1} + \eta_*\Delta_*^{-1}\eta_*')^{-1}.
$$

(3)

We prove Proposition 1 in Appendix A.1. This result leads us to propose a class of estimators of $\beta_*$ defined by

$$
\hat{\beta} = \hat{\Delta}^{-1}\hat{\eta}'(\hat{\Sigma}_{YY}^{-1} + \hat{\eta}\hat{\Delta}^{-1}\hat{\eta}')^{-1},
$$

(4)
where \( \hat{\eta}, \hat{\Delta}, \) and \( \hat{\Sigma}_{YY} \) are user-selected estimators of \( \eta_s, \Delta_s, \) and \( \Sigma_{sYY} \). If \( n > \max(p, q) \) and the ordinary sample estimators are used for \( \hat{\eta}, \hat{\Delta}, \) and \( \hat{\Sigma}_{YY} \), then \( \hat{\beta} \) is equivalent to \( \hat{\beta}^{(\text{OLS})} \).

We propose to use shrinkage estimators of \( \eta_s, \Delta_s^{-1}, \) and \( \Sigma_{sYY}^{-1} \) in \( \{1\} \). This gives us the potential to indirectly fit an un parsimonious forward regression model by fitting a parsimonious inverse regression model. For example, suppose that \( \eta_s \) and \( \Delta_s^{-1} \) are sparse, but \( \beta_s \) is dense. To fit the inverse regression model, we could use any of the forward regression shrinkage estimators discussed in Section \( \{1\} \).

### 2.2 Related work

Lee and Liu (2012) proposed an estimator of \( \beta_s \) that also exploits the assumption that \( (X', Y')' \) is multivariate Normal; however, unlike our approach that makes no explicit assumptions about \( \beta_s \), their approach assumes that both \( \Sigma_s^{-1} \) and \( \beta_s \) are sparse.

Modeling the inverse regression is a well-known idea in multivariate analysis. For example, when \( Y \) is categorical, quadratic discriminant analysis models \( (X|Y = y) \) as \( p \)-variate Normal. There are also many examples of modeling the inverse regression in the sufficient dimension reduction literature (Adragni and Cook, 2009).

The most closely related work to ours is that by Cook et al. (2013). They proposed indirect estimators of \( \beta_s \) based on modeling the inverse regression in the special case when the response is univariate, i.e., \( q = 1 \). Under the same multivariate Normal assumption on \( (X', Y')' \) that we make, Cook et al. (2013) showed that

\[
\beta_s = \frac{1}{1 + \Sigma'_{sXY} \Delta_s^{-1} \Sigma_{sXY} / \Sigma_{sYY} - \Delta_s^{-1} \Sigma_{sXY} \Delta_s^{-1}}. \tag{5}
\]

They proposed estimators of \( \beta_s \) by replacing \( \Sigma_{sXY} \) and \( \Sigma_{sYY} \) in the right hand side of (5) with their usual sample estimators, and by replacing \( \Delta_s^{-1} \) with a shrinkage estimator. This class of estimators was designed to exploit an abundant signal rate in the forward univariate response regression when \( p > n \).

### 3 Asymptotic Analysis

We present a convergence rate bound for the indirect estimator of \( \beta_s \) defined by \( \{1\} \). Our bound allows \( p \) and \( q \) to grow with the sample size \( n \). In the following proposition, \( \| \cdot \| \) is the spectral norm and \( \varphi_{\min}(\cdot) \) is the minimum eigenvalue.

**Proposition 2.** Suppose that following conditions are true: (i) \( \Sigma_s \) is positive definite for all \( p + q \); (ii) the estimator \( \hat{\Sigma}_{YY}^{-1} \) is positive definite for all \( q \); (iii) the estimator \( \hat{\Delta}^{-1} \) is positive definite for all \( p \); (iv) there exists a positive constant \( K \) such that \( \varphi_{\min}(\Sigma_{sYY}^{-1}) \geq K \) for all \( q \); and (v) there exist sequences \( \{a_n\}, \{b_n\} \), and \( \{c_n\} \) such that \( \|\hat{\eta} - \eta_s\| = O_P(a_n), \|\hat{\Delta}^{-1} - \Delta_s^{-1}\| = O_P(b_n), \|\hat{\Sigma}_{YY}^{-1} - \Sigma_{sYY}^{-1}\| = O_P(c_n) \), \( a_n \|\eta_s\| \cdot \|\Delta_s^{-1}\| + b_n \|\eta_s\|^2 + c_n \) → 0 as \( n \) → ∞. Then

\[
\|\hat{\beta} - \beta_s\| = O_P\left(a_n \|\eta_s\|^2 \|\Delta_s^{-1}\|^2 + b_n \|\eta_s\|^3 \|\Delta_s^{-1}\|^3 + c_n \|\eta_s\| \cdot \|\Delta_s^{-1}\|^3\right).
\]

We prove Proposition 2 in Appendix \( \{A\} \). We used the spectral norm because it is compatible with the convergence rate bounds established for sparse inverse covariance estimators (Rothman et al., 2008; Lam and Fan, 2009; Ravikumar et al., 2011).
If the inverse regression is parsimonious in the sense that \( \| \eta^* \| \) and \( \| \Delta_*^{-1} \| \) are bounded, then the bound in Proposition 2 simplifies to \( \| \hat{\beta} - \beta_* \| = O_P(a_n + b_n + c_n) \). From an asymptotic perspective, it is not surprising that the indirect estimator of \( \beta_* \) is only as good as its worst plug-in estimator. We explore finite sample performance in Section 5.

4 Example estimators in our class

4.1 Sparse inverse regression

We now describe an estimator of the forward regression coefficient matrix \( \beta_* \) defined by (4) that exploits zeros in the inverse regression’s coefficient matrix \( \eta_* \), zeros in the inverse regression’s error precision matrix \( \Delta_*^{-1} \), and zeros in the precision matrix of the responses \( \Sigma_*^{-1}YY \). We estimate \( \eta_* \) with

\[
\hat{\eta}^{L1} = \text{arg min}_{\eta \in \mathbb{R}^{q \times p}} \left\{ \| X - Y \eta \|_F^2 + \sum_{j=1}^{p} \lambda_j \sum_{m=1}^{q} | \eta_{mj} | \right\},
\]

which separates into \( p \) L1-penalized least-squares regressions (Tibshirani, 1996): the first predictor regressed on the response through the \( p \)th predictor regressed on the response. We select \( \lambda_j \) with 5-fold cross-validation, minimizing squared prediction error totaled over the folds, in the regression of the \( j \)th predictor on the response \( (j = 1, \ldots, p) \). This allows us to estimate the columns of \( \eta_* \) in parallel.

We estimate \( \Delta_*^{-1} \) and \( \Sigma_*^{-1}YY \) with L1-penalized Normal likelihood precision matrix estimation (Yuan and Lin, 2007; Banerjee et al., 2008). Let \( \hat{\Sigma}_{\gamma,S}^{-1} \) be a generic version of this estimator with tuning parameter \( \gamma \) and input \( p \times p \) sample covariance matrix \( S \):

\[
\hat{\Sigma}_{\gamma,S}^{-1} = \text{arg min}_{\Omega \in S^p_+} \left\{ \text{tr}(\Omega S) - \log |\Omega| + \gamma \sum_{j \neq k} | \omega_{jk} | \right\},
\]

where \( S^p_+ \) is the set of symmetric and positive definite \( p \times p \) matrices. There are many algorithms that solve (7). Two good choices are the graphical lasso algorithm (Yuan, 2008; Friedman et al., 2008) and the QUIC algorithm (Hsieh et al., 2011). We select \( \gamma \) with 5-fold cross-validation maximizing a validation likelihood criterion (Huang et al., 2006):

\[
\hat{\gamma} = \text{arg min}_{\gamma \in \mathcal{G}} \sum_{k=1}^{5} \left\{ \text{tr}\left( \hat{\Sigma}_{\gamma,S(-k)}^{-1} S(k) \right) - \log \left| \hat{\Sigma}_{\gamma,S(-k)}^{-1} \right| \right\},
\]

where \( \mathcal{G} \) is a user-selected finite subset of the non-negative real line, \( S(-k) \) is the sample covariance matrix from the observations outside the \( k \)th fold, and \( S(k) \) is the sample covariance matrix from the observations in the \( k \)th fold centered by the sample mean of the observations outside the \( k \)th fold.

We estimate \( \Delta_*^{-1} \) using (7) with its tuning parameter selected by (8) and \( S = (X - Y\hat{\eta}^{L1})'(X - Y\hat{\eta}^{L1})/n \). Similarly, we estimate \( \Sigma_*^{-1}YY \) using (7) with its tuning parameter selected by (8) and \( S = Y'Y/n \).
4.2 Reduced rank inverse regression

We propose indirect estimators of $\beta_*$ that exploit the assumption that the inverse regression’s coefficient matrix $\eta_*$ is rank deficient. We have the following simple proposition that links rank deficiency in $\eta_*$ and its estimator to $\beta_*$ and its indirect estimator.

**Proposition 3.** If $\Sigma_*$ is positive definite, then $\text{rank}(\beta_*) = \text{rank}(\eta_*)$. In addition, if $\hat{\Sigma}_{YY}^{-1}$ and $\hat{\Delta}^{-1}$ are positive definite in the indirect estimator $\hat{\beta}$ defined by (4), then $\text{rank}(\hat{\beta}) = \text{rank}(\hat{\eta})$.

The proof of this proposition is simple so we excluded it to save space.

We propose the following two example reduced rank indirect estimators of $\beta_*$:  

1. Estimate $\Sigma_{YY}$ with $Y'Y/n$ and estimate $(\eta_*, \Delta_*^{-1})$ with Normal likelihood reduced rank inverse regression:

$$
(\hat{\eta}^{(r)}, \hat{\Delta}^{-1}(r)) = \arg \min_{(\eta, \Omega) \in \mathbb{R}^{q \times p} \times S_p^+} \left[ n^{-1} \text{tr} \left\{ (X - Y \hat{\eta})'(X - Y \hat{\eta})\Omega \right\} - \log \det(\Omega) \right] 
$$

subject to rank$(\eta) = r$,

where $r$ is selected from $\{0, \ldots, \min(p, q)\}$. The solution to the optimization in (9) is available in closed form (Reinsel and Velu, 1998).

2. Estimate $\eta_*$ with $\hat{\eta}^{(r)}$ defined in (9), estimate $\Sigma_{YY}^{-1}$ with $(7)$ using $S = Y'Y/n$, and estimate $\Delta_*^{-1}$ with $(7)$ using $S = (X - Y\hat{\eta}^{(r)})'(X - Y\hat{\eta}^{(r)})/n$.

Both example indirect reduced rank estimators of $\beta_*$ are formed by plugging in the estimators of $\eta_*, \Delta_*^{-1}$, and $\Sigma_{YY}$ to (4). The first estimator is likelihood-based and the second estimator exploits sparsity in $\Sigma_{YY}^{-1}$ and $\Delta_*^{-1}$. Neither estimator is defined when $\min(p, q) > n$. In this case, which we do not address, a regularized reduced rank estimator of $\eta_*$ could be used instead of the estimator defined in (9), e.g. the factor estimation and selection estimator (Yuan et al., 2007) or the reduced rank ridge regression estimator (Mukherjee and Zhu, 2011).

5 Simulations

5.1 Sparse inverse regression simulation

We compared the following indirect estimators of $\beta_*$ when the inverse regression’s coefficient matrix $\eta_*$ is sparse:

- $I_{L1}$. This is the indirect estimator proposed in Section 4.1.

- $I_S$. This is an indirect estimator defined by (4) with $\hat{\eta}$ defined by (6), $\hat{\Sigma}_{YY} = Y'Y/n$, and $\hat{\Delta} = (X - Y\hat{\eta}^{L1})'(X - Y\hat{\eta}^{L1})/n$.

- $O_{\Delta}$. This is a part oracle indirect estimator defined by (4) with $\hat{\eta}$ defined by (6), $\hat{\Sigma}_{YY}$ defined by (7), and $\hat{\Delta}^{-1} = \Delta_*^{-1}$.

- $O$. This is a part oracle indirect estimator defined by (4) with $\hat{\eta}$ defined by (6), $\hat{\Sigma}_{YY} = \Sigma_{YY}^{-1}$, and $\hat{\Delta}^{-1} = \Delta_*^{-1}$.
OLS/MP. This is the ordinary least squares estimator defined by \( \arg\min_{\beta} \|Y - X\beta\|_F^2 \). When \( n \leq p \), we use the solution \( \hat{X}^{-1} \), where \( \hat{X} \) is the Moore-Penrose generalized inverse of \( X \).

R. This is the ridge penalized least squares estimator defined by

\[
\arg\min_{\beta \in \mathbb{R}^p \times q} (\|Y - X\beta\|_F^2 + \lambda\|\beta\|_1^2) .
\]

\( \ell_2 \). This is an alternative ridge penalized least squares estimator defined by

\[
\arg\min_{\beta \in \mathbb{R}^p \times q} \left( \|Y - X\beta\|_F^2 + \sum_{m=1}^q \lambda_m \sum_{j=1}^p \beta_{jm}^2 \right) ,
\]

where a separate tuning parameter is used for each response.

We selected the tuning parameters for uses of (6) with 5-fold cross-validation, minimizing validation prediction error on the inverse regression. Tuning parameters for \( \ell_2 \) and R were selected with 5-fold cross-validation, minimizing validation prediction error on the forward regression. We selected tuning parameters for uses of (7) with (8). The candidate set of tuning parameters was \( \{10^{-8}, 10^{-7.5}, \ldots, 10^{-2.5}, 10^8\} \).

For 50 independent replications, we generated a realization of \( n \) independent copies of \((X', Y')'\), where \( Y \sim N_q(0, \Sigma_{sY}) \) and \((X|Y = y) \sim N_p(\eta_{s}', y, \Delta_s)\). The \((i, j)\)th entry of \( \Sigma_{sY} \) was set to \( \rho_Y^{i-j} \) and the \((i, j)\)th entry of \( \Delta_s \) was set to \( \rho_{\Delta}^{i-j} \). We set \( \eta_s = Z \circ A \), where \( \circ \) denotes the element-wise product: \( Z \) had entries independently drawn from \( N(0, 1) \) and \( A \) had entries independently drawn from the Bernoulli distribution with nonzero probability \( s_s \). This model is ideal for \( I_{L1} \) because \( \Delta_{s}^{-1} \) and \( \Sigma_{sY}^{-1} \) are both sparse. Every entry in the corresponding randomly generated \( \beta_s \) is nonzero with high probability, but the magnitudes of these entries are small. This motivated us to compare our indirect estimators of \( \beta_s \) to the ridge-penalized least squares forward regression estimators R and \( \ell_2 \).

We evaluated performance with model error (Breiman and Friedman, 1997; Yuan et al., 2007), which is defined by \( \|\Sigma_{sX}(\beta - \beta_s)\|_F^2 \).

We report the average model errors, based on these 50 replications, in Table 1. When \( s_s = 0.1 \), the indirect estimators defined by (4) performed well for all choices of \( \rho_Y \) and \( \rho_\Delta \). Our proposed estimator \( I_{L1} \) was competitive with other indirect estimators also defined by (4), even those that used some oracle information. As \( s_s \) increased with \( \rho_Y = 0.7 \) and \( \rho_\Delta = 0.9 \) fixed, the forward regression estimators performed nearly as well as \( I_{L1} \).

Similarly, Table 2 shows that when \( s_s = 0.1 \), \( I_{L1} \) outperforms all three forward regression estimators. However, unlike in the lower dimensional setting illustrated in Table 1 when \( \eta_s \) is not sparse, i.e. \( s_s \geq .3 \), \( I_{L1} \) is outperformed by forward regression approaches. The part oracle method \( O_Y \) that used the knowledge of \( \Sigma_{sY}^{-1} \) outperformed the other two part oracle indirect estimators \( O \) and \( O_\Delta \) when \( \rho_\Delta = .9 \). Also, when \( \rho_\Delta = .9 \), \( I_{L1} \) was competitive with the part oracle estimators. Taken together, the results in Tables 1 and 2 suggest that when \( \eta_s \) is very sparse, our proposed indirect estimator \( I_{L1} \) may perform nearly as well as the part oracle indirect estimators and the forward regression estimators.
Table 1: Averages of model error from 50 replications when \( n = 100, p = 20, \) and \( q = 20. \) All standard errors were less than or equal to 0.05.

| \( \rho_Y \) | \( \rho_{\Delta} \) | \( s_* \) | \( I_{L1} \) | \( O \) | \( O_{\Delta} \) | \( O_Y \) | \( I_S \) | \( \ell_2 \) | \( R \) |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.7   | 0.0   | 0.1   | 0.61  | 0.32  | 0.53  | 0.40  | 1.35  | 2.10  | 1.23  | 1.22  |
| 0.7   | 0.5   | 0.1   | 0.72  | 0.39  | 0.59  | 0.51  | 1.30  | 1.91  | 1.29  | 1.30  |
| 0.7   | 0.7   | 0.1   | 0.76  | 0.45  | 0.65  | 0.56  | 1.27  | 1.73  | 1.27  | 1.29  |
| 0.7   | 0.9   | 0.1   | 0.83  | 0.66  | 0.85  | 0.64  | 1.26  | 1.35  | 1.05  | 1.09  |
| 0.0   | 0.9   | 0.1   | 0.81  | 0.87  | 0.87  | 0.79  | 2.04  | 2.34  | 1.26  | 1.87  |
| 0.5   | 0.9   | 0.1   | 0.96  | 0.76  | 0.99  | 0.74  | 1.63  | 1.84  | 1.36  | 1.49  |
| 0.9   | 0.9   | 0.1   | 0.46  | 0.39  | 0.47  | 0.36  | 0.63  | 0.62  | 0.48  | 0.48  |
| 0.7   | 0.9   | 0.3   | 0.60  | 0.53  | 0.65  | 0.46  | 0.83  | 0.67  | 0.64  | 0.63  |
| 0.7   | 0.9   | 0.5   | 0.48  | 0.37  | 0.48  | 0.37  | 0.65  | 0.53  | 0.52  | 0.51  |
| 0.7   | 0.9   | 0.7   | 0.42  | 0.29  | 0.39  | 0.31  | 0.55  | 0.46  | 0.45  | 0.44  |

Table 2: Averages of model error from 50 replications when \( n = 50, p = 60, \) and \( q = 60. \) All standard errors were 0.69 or less, except for MP, which had standard errors between 0.77 and 3.16.

| \( \rho_Y \) | \( \rho_{\Delta} \) | \( s_* \) | \( I_{L1} \) | \( O \) | \( O_{\Delta} \) | \( O_Y \) | \( MP \) | \( \ell_2 \) | \( R \) |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.7   | 0.0   | 0.1   | 8.59  | 4.28  | 5.70  | 7.40  | 78.33 | 13.85 | 12.44 |
| 0.7   | 0.5   | 0.1   | 9.67  | 5.09  | 6.37  | 8.49  | 73.82 | 14.79 | 13.34 |
| 0.7   | 0.7   | 0.1   | 10.01 | 6.37  | 7.44  | 8.75  | 70.30 | 15.56 | 14.40 |
| 0.7   | 0.9   | 0.1   | 9.92  | 10.07 | 11.44 | 8.88  | 61.83 | 16.43 | 15.94 |
| 0.0   | 0.9   | 0.1   | 15.17 | 17.09 | 16.93 | 15.23 | 119.60| 28.63 | 29.41 |
| 0.5   | 0.9   | 0.1   | 14.88 | 13.59 | 16.91 | 12.01 | 86.88 | 23.62 | 22.69 |
| 0.9   | 0.9   | 0.1   | 16.86 | 17.43 | 19.66 | 15.44 | 43.88 | 15.30 | 14.14 |
| 0.7   | 0.9   | 0.3   | 16.86 | 17.43 | 19.66 | 15.44 | 43.88 | 15.30 | 14.14 |
| 0.7   | 0.9   | 0.5   | 26.89 | 26.81 | 29.93 | 24.95 | 36.87 | 14.79 | 13.62 |
| 0.7   | 0.9   | 0.7   | 31.86 | 35.98 | 38.64 | 30.36 | 33.58 | 14.35 | 13.65 |

5.2 Reduced rank inverse regression simulation

We compared the performance of the following indirect reduced rank estimators of \( \beta_*: \)

\( I_{ML}^{(r)} \): This is the likelihood-based indirect example estimator 1 proposed in Section 4.2.

\( I^{(r)} \): This is the indirect example estimator 2 proposed in Section 4.2, which uses sparse estimators of \( \Sigma_{YY}^{-1} \) and \( \Delta_*^{-1} \) in (4).

\( O^{(r)} \): This is a part oracle indirect estimator defined by (4) with \( \hat{\eta} \) defined by (9), \( \hat{\Delta}^{-1} = \Delta_*^{-1} \), and \( \hat{\Sigma}_{YY}^{-1} = \Sigma_{YY}^{-1} \).

\( O_{\Delta}^{(r)} \): This is a part oracle indirect estimator defined by (4) with \( \hat{\eta} \) defined by (9), \( \hat{\Delta}^{-1} \) defined by (7), and \( \hat{\Sigma}_{YY}^{-1} = \Sigma_{YY}^{-1} \).

\( O_Y^{(r)} \): This is a part oracle indirect estimator defined by (4) with \( \hat{\eta} \) defined by (9), \( \hat{\Delta}^{-1} = \Delta_*^{-1} \), \( \hat{\Delta}^{-1} \) defined by (7), and \( \hat{\Sigma}_{YY}^{-1} \) defined by (7).
Table 3: Averages of model error from 50 replications when $n = 100$, $p = 20$, and $q = 20$. All standard errors were less than or equal to 0.05.

| $\rho_Y$ | $\rho_\Delta$ | $r_*$ | $I$ | $O_\Delta$ | $O_Y$ | $I_{ML}$ | OLS | RR |
|----------|----------------|-------|-----|------------|------|----------|-----|----|
| 0.7      | 0.0            | 10    | 0.33| 0.04       | 0.86 | 0.75     | 0.64| 1.38| 0.64|
| 0.7      | 0.5            | 10    | 0.34| 0.04       | 0.86 | 0.74     | 0.60| 1.31| 0.60|
| 0.7      | 0.7            | 10    | 0.31| 0.03       | 0.86 | 0.80     | 0.62| 1.32| 0.61|
| 0.7      | 0.9            | 10    | 0.31| 0.02       | 0.85 | 0.88     | 0.60| 1.30| 0.61|
| 0.0      | 0.9            | 10    | 0.15| 0.03       | 1.00 | 1.77     | 1.22| 2.61| 1.21|
| 0.5      | 0.9            | 10    | 0.42| 0.01       | 1.11 | 1.36     | 0.90| 1.97| 0.89|
| 0.9      | 0.9            | 10    | 0.12| 0.01       | 0.32 | 0.30     | 0.22| 0.46| 0.22|
| 0.7      | 0.9            | 4     | 0.35| 0.02       | 1.73 | 2.61     | 0.49| 3.12| 0.49|
| 0.7      | 0.9            | 8     | 0.35| 0.01       | 1.15 | 1.33     | 0.68| 1.73| 0.65|
| 0.7      | 0.9            | 12    | 0.31| 0.04       | 0.64 | 0.59     | 0.55| 0.96| 0.53|
| 0.7      | 0.9            | 16    | 0.25| 0.08       | 0.30 | 0.20     | 0.44| 0.50| 0.42|

We compared these indirect estimators to the following forward reduced rank regression estimator: RR. This is the likelihood based reduced rank regression (Izenman, 1975; Reinsel and Velu, 1998). The estimator of $\beta_*$ and the estimator of the forward regression’s error precision matrix $\Sigma_{-1}^{-1}$ are defined by

$$\hat{\beta}(r), \hat{\Sigma}_{-1}(r) = \arg\min_{(\beta, \Omega) \in \mathbb{R}^{p \times q}} \left\{ n^{-1} \text{tr} \left\{ (Y - X\beta)'(Y - X\beta)\Omega \right\} - \log \det(\Omega) \right\}$$ subject to rank($\beta$) = $r$.

We selected the rank parameter $r$ for uses of (9) with 5-fold cross-validation, minimizing validation prediction error on the inverse regression. The rank parameter for RR was selected with 5-fold cross-validation, minimizing validation prediction error on the forward regression. We selected tuning parameters for uses of (7) with (8). The candidate set of tuning parameters was \{10^{-8}, 10^{-7.5}, \ldots, 10^{-7.5}, 10^8\}.

For 50 independent replications, we generated a realization of $n$ independent copies of $(X', Y')'$ where $Y \sim N_q(0, \Sigma_{YY})$ and $(X|Y = y) \sim N_p(\eta'_y, \Delta_y)$. The $(i, j)$th entry of $\Sigma_{YY}$ was set to $\rho_{Y|Y-i,j}$ and the $(i, j)$th entry of $\Delta$ was set to $\rho_{\Delta|Y-i,j}$. After specifying $r_* \leq \min(p, q)$, we set $\eta_* = PQ$, where $P \in \mathbb{R}^{q \times r_*}$ and $Q \in \mathbb{R}^{r_\times p}$ had entries independently drawn from $N(0,1)$ so that $r_* = \text{rank}(\eta_*) = \text{rank}(\beta_*)$. As we did in the simulation in Section 5.1, we measured performance with model error.

We report the model errors, averaged over the 50 independent replications, in Table 3. Under every setting, $I^{(r)}$ outperformed all non-oracle competitors. When $r_* \leq 12$, $I^{(r)}$ outperformed both $O^{(r)}_\Delta$ and $O^{(r)}_Y$, which suggests that shrinkage estimation of $\Delta_{-1}^{-1}$ and $\Sigma_{-1}^{-1}$ was helpful. In each setting, $I^{(r)}_{ML}$ performed similarly to RR even though they are estimating parameters of different condition distributions.
5.3 Reduced rank forward regression simulation

Our simulation studies in the previous sections used inverse regression data generating models. In this section, we compare the estimators from Section 5.2 using a forward regression data generating model.

For 50 independent replications, we generated a realization of \( n \) independent copies of \((X', Y')\) where \( X \sim N_p(0, \Sigma_{XX}) \) and \((Y | X = x) \sim N_q(\beta'_x, \Sigma_{E})\). The \((i, j)\)th entry of \( \Sigma_{XX} \) was set to \( \rho_{i-j} \) and the \((i, j)\)th entry of \( \Sigma_{E} \) was set to \( \rho_{i-j} \). After specifying \( r^* \leq \min(p, q) \), we set \( \beta^* = ZQ \) where \( Z \in \mathbb{R}^{p \times r^*} \) had entries independently drawn from \( N(0, 1) \) and \( Q \in \mathbb{R}^{r^* \times q} \) had entries independently drawn from \( \text{Uniform}(-1/4, 1/4) \). In this data generating model, neither \( \Delta^{-1} \) nor \( \Sigma_{YY}^{-1} \) had entries equal to zero.

Table 4: Averages of model error from 50 replications when \( n = 100 \), \( p = 20 \), and \( q = 20 \). All standard errors were less than or equal to 0.21.

| \( r^* \) | \( p_X \) | \( p_E \) | \( \rho_X \) | \( \rho_E \) | \( \eta^{(r)} \) | \( \eta^{(x)} \) | \( \xi^{(r)} \) | \( \xi^{(x)} \) | \( \eta^{(r)}_{ML} \) | \( \eta^{(x)}_{ML} \) | \( \text{OLS} \) | \( \text{RR} \) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.0 | 0.9 | 2.79 | 0.54 | 4.27 | 5.05 | 2.48 | 4.99 | 2.82 |
| 10 | 0.5 | 0.9 | 2.90 | 0.47 | 5.36 | 5.94 | 2.73 | 5.00 | 2.89 |
| 10 | 0.7 | 0.9 | 2.97 | 0.51 | 4.64 | 5.03 | 2.71 | 4.93 | 2.76 |
| 10 | 0.9 | 0.9 | 2.84 | 0.73 | 3.78 | 4.16 | 2.67 | 5.19 | 2.73 |
| 10 | 0.7 | 0.0 | 4.66 | 1.92 | 3.59 | 5.88 | 4.53 | 5.11 | 4.34 |
| 10 | 0.7 | 0.5 | 4.27 | 1.65 | 3.88 | 5.51 | 3.99 | 5.06 | 3.97 |
| 10 | 0.7 | 0.7 | 3.55 | 1.26 | 3.99 | 5.29 | 3.43 | 5.00 | 3.44 |
| 4 | 0.7 | 0.9 | 1.27 | 0.08 | 3.84 | 4.71 | 0.95 | 5.00 | 1.11 |
| 8 | 0.7 | 0.9 | 2.39 | 0.36 | 4.15 | 5.15 | 2.05 | 4.81 | 2.22 |
| 12 | 0.7 | 0.9 | 3.58 | 0.79 | 4.44 | 5.21 | 3.20 | 5.15 | 3.27 |
| 16 | 0.7 | 0.9 | 4.53 | 1.29 | 4.62 | 4.42 | 4.33 | 5.11 | 4.38 |

The model errors, averaged over the 50 replications, are reported in Table 4. Both \( I^{(r)} \) and \( I^{(ML)} \) were competitive with RR in most settings. Although neither \( \Delta^{-1} \) nor \( \Sigma_{YY}^{-1} \) were sparse, we again see that \( I^{(r)} \) generally outperforms \( O^{(r)} \) and \( O^{(x)} \), both of which use some oracle information. These results indicate that shrinkage estimators of \( \Delta^{-1} \) and \( \Sigma_{YY}^{-1} \) in (4) are helpful when neither is sparse.

6 Tobacco chemical composition data example

As an example application, we use the chemical composition of tobacco leaves data from Anderson and Bancroft (1952) and Izenman (2009). These data have \( n = 25 \) cases, \( p = 6 \) predictors, and \( q = 3 \) responses. The names of the predictors, taken from page 183 of Izenman (2009), are percent nitrogen, percent chlorine, percent potassium, percent phosphorus, percent calcium, and percent magnesium. The names of the response variables, also taken from page 183 of Izenman (2009), are rate of cigarette burn in inches per 1,000 seconds, percent sugar in the leaf, and percent nicotine in the leaf. In these data, it may inappropriate to assume that \( \Delta^{-1} \) is sparse. For this reason, we consider another example indirect estimator of \( \beta^* \) called \( I_{L2} \) that estimates \( \eta^* \) with (5), estimates \( \Sigma_{YY}^{-1} \) with (7)
using \( S = \frac{Y'Y}{n} \), and estimates \( \Delta_r^{-1} \) with

\[
\text{arg min}_{\Omega \in \mathcal{S}^p_+} \left\{ \text{tr}(\Omega S) - \log \det(\Omega) + \gamma \sum_{j,k} |\omega_{jk}|^2 \right\},
\]

(10)

where \( S = \frac{(Y - X\hat{\eta}^{L1})(Y - X\hat{\eta}^{L1})}{n} \). We compute (10) with the closed form solution derived by Witten and Tibshirani (2009). As before, we select \( \gamma \) from \( \{10^8, 10^{-7.5}, \ldots, 10^{-7.5}, 10^8\} \) using (8).

We also consider the forward regression estimators RR, \( \ell_2 \), and OLS defined in Section 5.1 and Section 5.2. We introduce another competitor \( \ell_1 \), defined as

\[
\text{arg min}_{\beta \in \mathbb{R}^{p \times q}} \left\{ \|Y - X\beta\|^2_F + \sum_{j=1}^q \lambda_j \sum_{l=1}^p |\beta_{jl}| \right\},
\]

which is equivalent to performing \( q \) separate lasso regressions (Tibshirani, 1996). We randomly split the data into a 40% test set and 60% training set in each of 500 replications and we measured the squared prediction error on the test set. All tuning parameters were chosen from \( \{10^8, 10^{-7.5}, \ldots, 10^{-7.5}, 10^8\} \) by 5-fold cross validation.

Table 5 shows squared prediction errors, averaged over the 10 predictions and the 500 replications. These results indicate that \( I_{L2} \) outperforms all the competitors we considered. Also, \( I_{L1} \) was outperformed by \( \ell_2 \), but was competitive with separate lasso regressions. Reduced rank regression was not competitive with the proposed indirect estimators.

Acknowledgment

We thank Liliana Forzani for an important discussion. This research is supported in part by a grant from the U.S. National Science Foundation.
A Appendix

A.1 Proofs

Proof of Proposition 1. Since $\Sigma_s$ is positive definite, we apply the partitioned inverse formula to obtain that

$$
\Sigma_s^{-1} = \begin{pmatrix} 
\Sigma_s^{XX} & \Sigma_s^{XY} \\
\Sigma_s^{XY} & \Sigma_s^{YY}
\end{pmatrix}^{-1} = \begin{pmatrix} 
\Delta_s^{-1} & -\beta_s\Sigma_s^{-1}E \\
-\eta_s\Delta_s^{-1} & \Sigma_s^{-1}E
\end{pmatrix},
$$

where $\Delta_s = \Sigma_s^{XX} - \Sigma_s^{XY}\Sigma_s^{-1}Y\Sigma_s^{XY}$ and $\Sigma_s = \Sigma_s^{YY} - \Sigma_s^{XY}\Sigma_s^{-1}X\Sigma_s^{XY}$. The symmetry of $\Sigma_s^{-1}$ implies that $\beta_s\Sigma_s^{-1}E = (\eta_s\Delta_s^{-1})' \beta_s = \Delta_s^{-1}\eta_s'\Sigma_sE.$

Using the Woodbury identity,

$$
\Sigma_s^{-1}E = \Sigma_s^{-1}Y + \eta_s\Delta_s^{-1}\eta_s',
$$

Using the inverse of the expression above in (11) establishes the result.

In our proof of Proposition 2, we use the matrix inequality

$$
\|A(1)A(2)A(3) - B(1)B(2)B(3)\| \leq \sum_{j=1}^{3} \|A(j) - B(j)\| \prod_{k\neq j} \|B(k)\|
$$

$$
+ \sum_{j=1}^{3} \|B(j)\| \prod_{k\neq j} \|A(k) - B(k)\| + \sum_{j=1}^{3} \|A(j) - B(j)\|.
$$

Bickel and Levina (2008) used (13) to prove their Theorem 3.

Proof of Proposition 2. From (12) in the proof of Proposition 1, $\Sigma_s^{-1} = \Sigma_s^{-1}Y + \eta_s\Delta_s^{-1}\eta_s'$. Define $\hat{\Sigma}_s^{-1} = \Sigma_s^{-1} + \eta_s\hat{\Delta}_s^{-1}\hat{\eta}_s'$. Applying (13),

$$
\|\hat{\beta} - \beta_s\| = \|\hat{\Delta}_s^{-1}\hat{\eta}_s\hat{\Sigma}_s - \Delta_s^{-1}\eta_s\Sigma_sE\|
$$

$$
\leq \|\hat{\Delta}_s^{-1} - \Delta_s^{-1}\| \cdot \|\eta_s\| \cdot \|\Sigma_sE\| + \|\hat{\eta} - \eta_s\| \cdot \|\Delta_s^{-1}\| \cdot \|\Sigma_sE\| + \|\hat{\Sigma}_s - \Sigma_sE\| \cdot \|\Delta_s^{-1}\| \cdot \|\eta_s\|
$$

$$
+ \|\Delta_s^{-1}\| \cdot \|\hat{\eta} - \eta_s\| \cdot \|\hat{\Sigma}_s - \Sigma_sE\| + \|\eta_s\| \cdot \|\hat{\Delta}_s^{-1} - \Delta_s^{-1}\| \cdot \|\Sigma_sE - \Sigma_s\|
$$

$$
+ \|\Sigma_sE\| \cdot \|\hat{\Delta}_s^{-1} - \Delta_s^{-1}\| \cdot \|\hat{\eta} - \eta_s\| + \|\hat{\eta} - \eta_s\| \cdot \|\hat{\Delta}_s^{-1} - \Delta_s^{-1}\| \cdot \|\Sigma_sE - \Sigma_s\|
$$

We will show that the third term in (14) dominates the others. We continue by deriving its bound. Employing a matrix identity used by Cai et al. (2010), we write $\hat{\Sigma}_s - \Sigma_sE = \Sigma_sE(\Sigma_s^{-1}E - \Sigma_s^{-1})\Sigma_sE$, so

$$
\|\hat{\Sigma}_s - \Sigma_sE\| \leq \|\Sigma_sE\| \cdot \|\Sigma_sE\| \cdot \|\Sigma_s^{-1}E - \Sigma_s^{-1}\|.
$$

(15)
Using the triangle inequality and (13),
\[
\|\hat{\Sigma}_E^{-1} - \Sigma_{sE}^{-1}\| \leq \|\hat{\Sigma}_Y^{-1} - \Sigma_{sY}^{-1}\| + \|\hat{\eta}\hat{\Delta}^{-1}\hat{\eta}' - \eta_s\Delta_s^{-1}\eta_s\|
\]
\[
\leq \|\hat{\Sigma}_Y^{-1} - \Sigma_{sY}^{-1}\| + 2\|\hat{\eta} - \eta_s\| \cdot \|\Delta_s^{-1}\| \cdot \|\eta_s\| + \|\hat{\Delta}^{-1} - \Delta_s^{-1}\| \cdot \|\eta_s\|^2
\]
\[
+ 2\|\eta_s\| \cdot \|\Delta_s^{-1}\| \cdot \|\hat{\eta} - \eta_s\| + \|\hat{\eta} - \eta_s\|^2 + \|\hat{\eta} - \eta_s\|^2 \|\hat{\Delta}^{-1} - \Delta_s^{-1}\|
\]
\[
= O_P \left( c_n + a_n\|\eta_s\| \cdot \|\Delta_s^{-1}\| + b_n\|\eta_s\|^2 \right).
\]
Using (16), (17), and (18), in (15),
\[
\|\hat{\Sigma}_E - \Sigma_{sE}\| = O_P \left( a_n\|\eta_s\| \cdot \|\Delta_s^{-1}\| + b_n\|\eta_s\|^2 + c_n \right).
\]
We then see that the third term in (14) dominates and
\[
\|\hat{\beta} - \beta_s\| = O_P \left( \left\{ a_n\|\eta_s\| \cdot \|\Delta_s^{-1}\| + b_n\|\eta_s\|^2 + c_n \right\} \|\eta_s\| \|\Delta_s^{-1}\| \right)
\]
\[
= O_P \left( a_n\|\eta_s\|^2 \|\Delta_s^{-1}\| + b_n\|\eta_s\|^3 \|\Delta_s^{-1}\| + c_n\|\eta_s\| \cdot \|\Delta_s^{-1}\| \right).
\]
\]

References

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

Anderson, R. L. and Bancroft, T. A. (1952). Statistical theory in research.

Banerjee, O., El Ghaoui, L., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38:2118–2144.
Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Cook, R. D., Forzani, L., and Rothman, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3059–3088.

Cook, R. D., Li, B., and Chiaramonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, 20:927–1010.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. K. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, volume 24, pages 2330–2338. MIT Press, Cambridge, MA.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Izenman, A. J. (2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, 37:4254–4278.

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

Mukherjee, A. and Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4(6):612–622.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

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. *The Annals of Applied Statistics*, 4(1):53.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-rank Regression*. Springer.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.
Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98:133–146.

Su, Z. and Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika*, 99:687–702.

Su, Z. and Cook, R. D. (2013). Scaled envelopes: Scale invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100:921–938.

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

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636.

Yuan, M. (2008). Efficient computation of l1 regularized estimates in gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.