Capturing Between-Tasks Covariance and Similarities Using Multivariate Linear Mixed Models

Aviv Navon, Saharon Rosset

Abstract

We consider the problem of predicting several response variables using the same set of explanatory variables. This setting naturally induces a group structure over the coefficient matrix, in which every explanatory variable corresponds to a set of related coefficients. Most of the existing methods that utilize this group formation assume that the similarities between related coefficients arise solely through a joint sparsity structure. In this paper, we propose a procedure for constructing an estimator of a multivariate regression coefficient matrix that directly models and captures the within-group similarities, by employing a multivariate linear mixed model formulation, with joint estimation of covariance matrices for coefficients and errors via penalized likelihood. Our approach, which we term Multivariate random Regression with Covariance Estimation (MrRCE) encourages structured similarity in parameters, in which coefficients for the same variable in related tasks sharing the same sign and similar magnitude. We illustrate the benefits of our approach in synthetic and real examples, and show that the proposed method outperforms natural competitors and alternative estimators under several model settings.

Keywords: Multivariate regression; Regularization methods; Sparse precision matrix; Covariance selection; Variance components; Penalized likelihood

1 Introduction

In many cases, a common set of predictor variables is used for predicting different but related target variables. For example, an on-demand transportation company may attempt forecasting demand and supply in different time frames and geographic locations; a real-estate firm may be interested in predicting both the construction costs and the sale prices of residential apartments, given a set of project’s physical and financial covariates, and external economic variables.

The general task of modeling multiple responses using a joint set of covariates can be expressed using multivariate regression (MR), or multiple response regression — a generalization of the classical regression model to regressing \( q > 1 \) responses on \( p \) predictors. In the MR settings, one is presented with \( n \) independent observations, \( \{(X_i, Y_i)\}_{i=1}^n \), where
$X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$ contain the predictors and responses for the $i$th sample, respectively. Let $X = (X_1, ..., X_n)^T = (x_1, ..., x_n) \in \mathbb{R}^{n \times p}$ denote the predictor matrix and $Y = (Y_1, ..., Y_n)^T = (y_1, ..., y_q) \in \mathbb{R}^{n \times q}$ denote the response matrix. For simplicity of notation, assume that the columns of $X$ and $Y$ have been centered so that we need not consider an intercept term. We further assume that the i.i.d $N_q(0, \Sigma)$ error terms are collected into an $n \times q$ error matrix $E$, where $\Sigma$ is the among-tasks covariance matrix. The multivariate regression model is given by,

$$Y = XB + E$$

(1)

where $B$ is a $p \times q$ regression coefficient matrix. The random matrices in (1) are assumed to follow a matrix-variate normal distribution (Dawid, 1981; Gupta and Nagar, 2018), $E \sim \text{MVN}_{n \times q}(0, I_n, \Sigma)$ and $Y \sim \text{MVN}_{n \times q}(XB, I_n, \Sigma)$. For reasons that will later become clear, when considering the noise structure of the MR model, the precision matrix, $\Omega = \Sigma^{-1}$, is commonly the preferred object.

Straightforward prediction and estimation with the MR model can become quite challenging when the number of predictors and responses is large relative to $n$, as it requires one to estimate $pq$ parameters. The univariate regression model ($q = 1$) has been widely studied, and numerous methods have been developed for variable selection (support recovery) and coefficients estimation. A naive approach to the MR problem is to apply one of these methods to each of the $q$ tasks independently. However, in many cases, the different problems are related, and this oversimplified approach fails to utilize all the information contained in the data (see, e.g., Breiman and Friedman (1997) and Rothman, Levina, and Zhu (2010)). For a review of Bayesian approaches for estimation and prediction with the MR model see Deshpande, Rockova, and George (2017) and references therein.

In the MR literature, many approaches seek to reduce the number of parameters to be estimated through a penalized (or constrained) least squares framework. Bunea, She, and Wegkamp (2011) generalized the classical Reduced-Rank Regression (RRR) (Anderson, 1951; Izenman, 1975; Velu and Reinsel, 2013) to high dimensional settings, estimating a low-rank coefficient matrix by penalizing the rank of $B$. Yuan, Ekici, et al. (2007) proposed a method called Factor Estimation and Selection (FES), in which an $L_1$-penalty is applied to the singular values of $B$. FES induces sparsity in the singular values of $B$, conducting dimension reduction and coefficients estimation simultaneously. One major drawback of dimension reduction techniques, is that the interpretation of the model is often limited, in terms of the original data, since the set of predictors is reduced to a few important principal factors.

The multivariate regression framework naturally induces a group structure over the coefficient matrix, $B$, in which every explanatory variable, $x_i$ for $i = 1, ..., p$, corresponds to a group of $q$ coefficients, $B_i = (\beta_{i1}, ..., \beta_{iq})$ (see Figure 1). While many approaches make no assumption over the group structure, others utilize it for learning structured sparsity. In the multi-task learning literature, the $L_1/L_2$-penalty, also known as the group lasso penalty (Yuan and Lin, 2006), has been applied with the rows of $B$ as groups. The $L_1/L_2$-penalty can be viewed as an intermediate between the $L_1$-penalty used in lasso regression (Tibshirani, 1996) and the $L_2$-penalty used in ridge regression (Hoerl and Kennard, 1970), aimed
Figure 1: The multivariate regression framework naturally induces a group structure over the coefficient matrix $B$, in which every explanatory variable, $x_i$, corresponds to a group of $q$ coefficients $B_i = (\beta_{i1}, ..., \beta_{iq})^T$.

at utilizing the relatedness among tasks for identifying the joint support, i.e., the set of predictors with non-zero coefficients across all $q$ responses (Obozinski, Wainwright, and Jordan, 2009). Peng et al. (2010) proposed a mixed constraint function, by applying both the lasso and the group lasso penalties to the elements and rows of $B$, respectively. This approach produces element-wise as well as row-wise sparsity in the coefficient matrix. Turlach, Venables, and Wright (2005) studied a different constraint function, placing an $L_\infty$-penalty over the rows of $B$. As noted by the authors, this method is only suitable for variable selection and not for estimation. Extensions of mixed norm penalties to overlapping groups have been proposed in order to handle more general and complex group structures (see, e.g., Kim and Xing (2012) and Y. Li, Nan, and Zhu (2015)). These methods produce highly interpretable models, however, they are limited to the case $\Omega \propto I_n$, and do not account for correlated errors. Rothman, Levina, and Zhu (2010), Chen and Huang (2016), and Wilms and Croux (2018) have recently shown that accounting for this additional information in MR problems can be beneficial for both coefficients estimation and prediction.

In multivariate normal theory, the entries of $\Omega$ that equal zero correspond to pairs of variables that are conditionally independent, given all of the other variables in the data. The problem of sparse precision matrix estimation has drawn considerable recent attention, and several methods have been proposed for both support recovery and parameter estimation. Perhaps the most widely used approach is the graphical lasso (Friedman, Hastie, and Tibshirani, 2008), in which simultaneous sparsity structure identification and coefficients estimation are achieved by minimizing the $L_1$-regularized negative log-likelihood function of $\Omega$ (Yuan and Lin, 2007; d’Aspremont, Banerjee, and El Ghaoui, 2008; Rothman, Bickel, et al., 2008). Recently, sparse precision matrix estimation has also been considered in regression frameworks, in which the main goal for this explicit estimation is to improve prediction (Witten and Tibshirani, 2009; Rothman, Levina, and Zhu, 2010).

Rothman, Levina, and Zhu (2010) proposed Multivariate Regression with Covariance Estimation (MRCE), a method for sparse multivariate regression that directly accounts for correlated errors. MRCE minimizes the negative log-likelihood function with an $L_1$-penalty
for both $B$ and $\Omega$,

$$\arg \min_{B, \Omega} -n \log |\Omega| + \text{tr} \left( \frac{1}{n} (Y - XB)^T \Omega (Y - XB) \right) + \lambda_1 \|B\|_1 + \lambda_2 \sum_{j \neq j'} |\omega_{jj'}|$$  \hspace{1cm} (2)

where $\text{tr}(\cdot)$ denotes the trace, $\lambda_1$ and $\lambda_2$ are the regularization parameters and $\omega_{jj'}$ is the $(j,j')$ element of $\Omega$. Lee and Liu (2012) extended the approach of Rothman, Levina, and Zhu (2010) to allow for weighted $L_1$-penalties over the elements of $B$ and $\Omega$. Yin and H. Li (2011) considered a similar objective to the one in (2), and proposed an algorithm for the sparse estimation of the coefficient and inverse covariance matrices. However, unlike Rothman, Levina, and Zhu (2010), their method aimed at improving the estimation of $\Omega$, rather than $B$. Our work further leverages correlations between the different problems to improve the accuracy of the estimators and predictions, by not only accounting for the correlation between the error terms but the similarities between the coefficients as well.

While MRCE accounts for correlated responses through the precision matrix $\Omega$, it does not learn structured sparsity in $B$, essentially selecting relevant covariates for each response separately. In a recent work, Wilms and Croux (2018) proposed an algorithm for the multivariate group lasso with covariance estimation, replacing the lasso penalty in (2) with an $L_1/L_2$-penalty over a pre-specified group structure. Chen and Huang (2016) developed a method within the reduced-rank regression framework that simultaneously performs variable selection and sparse precision matrix estimation. These methods for learning group sparsity assume that the sparsity structure is known a-priori. Instead, Sohn and Kim (2012) proposed an approach for group sparse multivariate regression that can jointly learn both the response structure and regression coefficients with structured sparsity.

All the above methods which considered a group structure over the coefficient matrix, essentially assume that the within-group similarities arise solely through a joint sparsity structure. In many applications, these structured (and unstructured) sparsity assumptions are not suitable, for instance, if one expects many covariates of small or medium effect. Furthermore, these sparse estimators encourage within-group coefficients to be of similar absolute magnitude, and do not favor same sign coefficients. However, in various real-life examples it is more natural to encourage coefficients within the same group to also share a sign. To address these issues, we construct an estimator for the multivariate regression by directly modeling and capturing the within-group similarities, while also accounting for the error covariance structures. Our method, titled Multivariate random Regression with Covariance Estimation (MrRCE), involves a multivariate linear mixed model with an underlying group structure over the coefficient matrix, designed to encourage related coefficients to share a common sign and similar magnitude.

Multivariate Linear Mixed Models (mvLMMs) (Henderson, 1984) are MR models that relate a joint set of covariates to multiple correlated responses. mvLMMs are applied in many real-life problems and frequently used in genetics due to their ability to account for relatedness among observations (see, e.g., Kruuk (2004), Kang et al. (2010), Korte et al. (2012), and Vattikuti, Guo, and Chow (2012)). The mvLMMs model can be viewed as a generalization of MR (similar to the way Linear Mixed Models (LMMs) are a generalization...
of linear regression models), allowing both fixed and random effects. Consider the MR problem (1), but with an additional term for the set of random predictors, collected into the matrix $Z = (Z_1, ..., Z_n)^T = (z_1, ..., z_r) \in \mathbb{R}^{n \times r}$. The mvLMM model is given by,

$$\begin{align*}
Y &= XB + Z\Gamma + E \\
E &\sim \text{MVN}_{n \times q} (0, I_n, \Sigma), \quad \Gamma \sim \text{MVN}_{r \times q} (0, R, G)
\end{align*}$$

where $B$ is a $p \times q$ fixed effect coefficient matrix and $\Gamma$ is an $r \times q$ random effect coefficient matrix. Here, $R$ and $G$ are the common covariance matrices of columns and rows of $\Gamma$, respectively.

In this paper we consider the problem of estimation and prediction under the multivariate random effect regression — an mvLMMs model strictly involving random effects,

$$Y = Z\Gamma + E$$

Under the proposed formulation and unlike the standard mvLMM framework, we are interested in estimating not only the covariance components but also in predicting the random component $\Gamma$. Our method accounts for correlations between responses and similarities among coefficients, captured by estimating a joint equicorrelation covariance matrix for the rows of $\Gamma$ (see Eq. 5 for details). Hence, the MrRCE method is an example of what one could call structured similarity learning, in which the different coefficient groups are assumed to be independent, whereas a within-group similarity is encouraged. This covariance structure for the random coefficient matrix reduces the MR problem of estimating $pq$ parameters, into the problem of estimating two covariance components — the coefficients’ common variance, and the intra-group correlation coefficient, or similarity level. The estimation of the covariance structure is achieved through a penalized likelihood, adding an $L_1$-penalty over the off-diagonal entries of $\Omega = \Sigma^{-1}$.

The remainder of the paper is structured as follows. Section 2 describes the MrRCE method and corresponding computational algorithms. Section 3 establishes a connection between the proposed method and the multivariate Ridge estimator. Simulation studies are performed in Section 4 to compare our method with competing estimators, and Section 5 contains two real data applications of MrRCE. Section 6 concludes with a brief discussion.

## 2 The MrRCE Method

Consider the random effect regression model (4) with $r = p$. Assume both the error matrix $E$ and the coefficient matrix $\Gamma$ follow a matrix variate normal distribution,

$$\begin{align*}
E &\sim \text{MVN}_{n \times q} (0, I_n, \Sigma), \quad \Gamma \sim \text{MVN}_{p \times q} (0, I_p, \sigma^2 C)
\end{align*}$$

$$\text{MVN}_{n \times q} (0, I_n, \Sigma)$$

$$\text{MVN}_{p \times q} (0, I_p, \sigma^2 C)$$

$$\text{MVN}_{n \times q} (0, I_n, \Sigma)$$

$$\text{MVN}_{p \times q} (0, I_p, \sigma^2 C)$$
Further assume an equicorrelation structure for the matrix $C$, controlled by the unknown intra-group correlation coefficient $\rho \in [0, 1)$,

$$C = C_\rho = \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \rho \\
\rho & \cdots & \rho & 1 
\end{pmatrix}$$

The unknown parameter $\rho$ can be thought of as a relative measure of the *within-group similarity* (Chatfield, J. Zidek, and Lindsey, 2010). Large values for $\rho$ correspond to high similarity among members of the same group, leading to a similar magnitude and same sign coefficients, whereas $\rho = 0$ corresponds to *i.i.d* draws for the entries of the coefficient matrix $\Gamma$.

Denote the negative log-likelihood function of $Y$ by $h(\cdot)$, and the collection of parameters by $\Theta = \{\Omega, \sigma^2, \rho\}$. We construct an estimator of $\Theta$ using a penalized normal likelihood, adding an $L_1$-penalty over the off-diagonal entries of $\Omega$,

$$\hat{\Theta} = \arg \min_{\Theta} h(Y; \Theta) + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}|$$

(6)

where $\lambda_\omega > 0$ is a regularization parameter.

### 2.1 The Algorithm

We propose an iterative tri-step algorithm for solving (6). On the first step we estimate a sparse precision matrix for the errors, followed by an optimization step over $(\sigma, \rho)$ and an Empirical-Best Linear Unbiased Predictor (E-BLUP) (Henderson, 1975; Henderson, 1984) step for the prediction of $\Gamma$. Alg. 1 provides a schematic overview of the MrRCE algorithm.

Using eigendecomposition (similar to Zhou and Stephens 2014 and Furlotte and Eskin 2015), we write,

$$C = U D U^T \quad \text{and} \quad Z Z^T = L S L^T$$

(7)

where $S$ and $D := D_\rho = \text{diag}(d_1(\rho), \ldots, d_q(\rho))$ are diagonal matrices, and $U$ is independent of $\rho$. We then multiply (4) by the orthogonal matrices $U$ and $L^T$ from the right and left correspondingly, to obtain,

$$\tilde{Y} = \tilde{Z} \tilde{\Gamma} + \tilde{E}$$

where $\tilde{Y} = L^T Y U$, $\tilde{Z} = L^T Z$, and,

$$\tilde{\Gamma} = \Gamma U \sim MVN_{p \times q}(0, \sigma^2 U^T CU) = MVN_{p \times q}(0, \sigma^2 D_\rho)$$

$$\tilde{E} = L^T E U \sim MVN_{n \times q}(0, L^T L = I_n, \tilde{\Sigma} := U^T \Sigma U) = MVN_{n \times q}(0, I_n, \tilde{\Sigma})$$

We lose the $\tilde{\cdot}$ notation and assume (with a slight abuse of notation) that the original data is
of the form,

\[ Y = Z\Gamma + E \]

\[ E \sim \text{MVN}_{n \times q} \left(0, I_n, \Sigma := \Omega^{-1}\right), \Gamma \sim \text{MVN}_{p \times q} \left(0, I_p, \sigma^2 D_\rho\right) \]  

namely,

\[ Y \sim \text{MVN}_{n \times q} \left(0, S, \sigma^2 D_\rho\right) + \text{MVN}_{n \times q} \left(0, I_n, \Sigma\right) \]

Next, we describe the algorithm for solving (6) under the assumptions (8).

**Step 1: Solving for \( \Omega \) condition on conditional on \( \Gamma \).** Starting with a realization \( \Gamma_0 \) for \( \Gamma \), we have,

\[ Y \mid \Gamma_0 \sim \text{MVN}_{q \times n} \left(Z\Gamma_0, I_n, \Sigma\right) \]

The negative log-likelihood of \( Y \mid \Gamma_0 \) can be expressed up to a constant as,

\[ g(\Omega) = \text{tr} \left[ \frac{1}{n} (Y - Z\Gamma_0)^T \Omega (Y - Z\Gamma_0) \right] - \log |\Omega| \]

Thus, given \( \Gamma_0 \) and a fixed \( \lambda_\omega > 0 \), we solve the penalized negative log-likelihood problem for \( \Omega \),

\[ \hat{\Omega} (\Gamma_0) = \arg \min_{\Omega \succeq 0} \left[ g(\Omega) + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}| \right] \]  

(9)

This minimization problem is exactly the L1-penalized precision matrix estimation problem considered by Yuan and Lin (2007), d'Aspremont, Banerjee, and El Ghaoui (2008), Friedman, Hastie, and Tibshirani (2008), Rothman, Levina, and Zhu (2010), and Hsieh et al. (2011), among others. We solve (9) by applying the graphical lasso algorithm of Friedman, Hastie, and Tibshirani (2008) over the residuals \( Y - Z\Gamma \).

**Step 2: Solving for \( \{\sigma, \rho\} \) with fixed \( \Omega \).** Given \( \hat{\Omega} \), we write \( Y = Z\Gamma + \hat{E} \) where \( \hat{E} := E \mid \hat{\Omega} \sim \text{MVN}_{n \times q} \left(0, I_n, \hat{\Omega}^{-1}\right) \). We let \( \otimes \) denote the Kronecker product and apply the vec (\( \cdot \)) operator\(^1\) to obtain,

\[ y = \text{vec} \left( Y \right) = (I_q \otimes Z) \gamma + \hat{\epsilon} \]

where \( \gamma \) and \( \hat{\epsilon} \) denote vec (\( \Gamma \)) and vec (\( \hat{E} \)) correspondingly, and,

\[ y \sim N_{nq} \left(0, \sigma^2 D_\rho \otimes S + \hat{\Omega}^{-1} \otimes I_n \right) \]

Denote the covariance matrix of \( y \) by \( V = V(\sigma, \rho) = \sigma^2 D_\rho \otimes S + \hat{\Omega}^{-1} \otimes I_n \). The negative

\(^1\)Let vec (\( \cdot \)) denote the concatenation of a \( k \times l \)-dimensional matrix’s columns into a \( kl \)-dimensional vector.
log-likelihood function of $y \mid \hat{\Omega}$ is given by (up to a constant),

$$
\tilde{h} \left( y \mid \hat{\Omega}; \sigma, \rho \right) = \log |V| + y^T V^{-1} y
$$

We minimize (10) with respect to $(\sigma, \rho)$, to obtain,

$$(\tilde{\sigma}, \tilde{\rho}) = \arg \min_{\sigma > 0, \rho \in [0, 1)} \tilde{h}$$

**Step 3: Solving for $\Gamma$ given $\Theta$.** Given $\hat{\Theta}$, we compute the E-BLUP (Henderson, 1975; Henderson, 1984) for $\Gamma$. Denote, $\tilde{Z} = I_q \otimes Z$, $L = \tilde{\sigma}^2 \tilde{C} \otimes I_p$ and $R = \hat{\Omega}^{-1} \otimes I_n$, the E-BLUP $\gamma^*$ for $\gamma$, is given by,

$$
\gamma^* = \left( \tilde{Z}^T R^{-1} \tilde{Z} + L^{-1} \right)^{-1} \tilde{Z}^T R^{-1} y
$$

Alternatively, as proved by Henderson et al. (1959), $\gamma^* = L^T \tilde{Z}^T \Lambda^{-1} y$ where, $\Lambda = \tilde{Z} L \tilde{Z}^T + R$. In order to predict $\Gamma^*$, we simply compute $\Gamma^* = \text{unvec}(\gamma^*)$ where unvec$(\cdot)$ represents the reversal of the vec$(\cdot)$ operation.

**Starting value and Stopping Criteria.** We start by taking $\Gamma = 0_{p \times q}$, and note that both steps 1 and 2 ensure a decrease in the objective function value. We consider three alternatives for the MrRCE algorithm’s stopping criteria.

1. Set $0 < K_{max} \in \mathbb{N}$. Stop if the number of iteration $k > K_{max}$.

2. Set a tolerance value, $\tau > 0$. Iterate until the sum of absolute changes in the values of $\Theta$ in two successive iterations is smaller than the tolerance value.

3. Let $\Gamma_{\text{ridge}} = \left( Z^T Z + \lambda I \right)^{-1} Z^T Y$, the ridge regression (RR) estimator for $\Gamma$. We stop at iteration $k$ if,

$$
\sum_{j,\ell} \left| \gamma_{j\ell}^{*(k)} - \gamma_{j\ell}^{*(k-1)} \right| < \xi \sum_{j,\ell} \left| \gamma_{j\ell}^{\text{ridge}} \right|
$$

for some $\xi > 0$. Here, $\gamma_{j\ell}^{*(k)}$ is the $(j, \ell)$ entry of $\Gamma^*$ at iteration $k$. This stopping criterion was proposed by Rothman, Levina, and Zhu (2010), and is well-defined since the RR estimator always exists. The criterion requires tighter convergence for small effects versus large ones.

3 Connection to Ridge Regression

We present a connection between the MrRCE method and the Ridge Regression (RR) estimator (Hoerl and Kennard, 1970). More specifically, we explore a special case in which the BLUP for $\Gamma$ derived at the third step of the MrRCE algorithm, is equivalent to the multivariate RR estimator (Brown and J. V. Zidek, 1980). For this special case, the MrRCE
Algorithm 1 (MrRCE): Iterative tri-step optimization procedure (see text for details)

Require: Regularization parameter $\lambda_\omega > 0$.

1. **Initialize**: set $k = 0$ and $\Gamma_k^* = 0$.
2. repeat
   
   $k \leftarrow k + 1$
   
   Step 1: solve $\Omega_k (\Gamma_{k-1}^*) = \arg\min_{\Omega \succeq 0} \left[ g(\Omega) + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}| \right]$  
   
   Step 2: solve $(\sigma_k, \rho_k) = \arg\min_{\sigma > 0, \rho \in [0,1]} \tilde{h}(y | \Omega_k; \sigma, \rho)$  
   
   Step 3: compute the E-BLUP for $\Gamma, \Gamma_k^*$.

3. until stopping criterion is reached.
4. return $(\Gamma_k^*, \rho_k, \sigma_k, \Omega_k)$

The algorithm can be viewed as an alternating “$L_1$-$L_2$ penalty” procedure: $L_1$-penalty for the estimation of $\Omega$, followed by an “$L_2$-penalty” for the prediction of $\Gamma$.

Consider the model,

$$y = \tilde{Z}\gamma + \epsilon$$

$$\epsilon \sim N(0, \Sigma_0 \otimes I_n := \Sigma), \gamma \sim N(0, \Lambda_0 \otimes I_p := \Lambda)$$

The joint distribution of $(y, \gamma)$ is given by,

$$
\begin{pmatrix}
\gamma \\
y
\end{pmatrix}
\sim N

\begin{bmatrix}
\Lambda & \Lambda \tilde{Z}^T \\
\tilde{Z} \Lambda & \tilde{Z} \Lambda \tilde{Z}^T + \Sigma
\end{bmatrix}
$$

and the BLUP for the random coefficient vector is the expectation of $\gamma$ conditional on $y$,

$$
\hat{\gamma}_{\text{BLUP}} = E[\gamma | y] = \Lambda \tilde{Z}^T \left( \tilde{Z} \Lambda \tilde{Z}^T + \Sigma \right)^{-1} y
$$

The RR estimator can be extended to the multivariate case as in Brown and J. V. Zidek (1980),

$$
\hat{\gamma}_{\text{RR}} = \left( \tilde{Z}^T \tilde{Z} + K \right)^{-1} \tilde{Z}^T y
$$

where $K \succ 0$ is the $pq \times pq$ ridge matrix. We apply the generalized Sherman-Morrison-Woodbury (Sherman and Morrison, 1950; Woodbury, 1950) formula to the inverse of $\tilde{Z}^T \tilde{Z} + K$, to obtain,

$$
\hat{\gamma}_{\text{RR}} = K^{-1} \tilde{Z}^T \left[ I - \left( I + \tilde{Z} K^{-1} \tilde{Z}^T \right)^{-1} \tilde{Z} K^{-1} \tilde{Z}^T \right] y
$$

(11)
Eq. (11) can be simplified as follow,
\[ \hat{\gamma}_{RR} = K^{-1} \hat{Z}^T \left[ \hat{Z} K^{-1} \hat{Z}^T + I \right]^{-1} y \]

Thus, under the i.i.d error model, i.e., \( \Sigma_0 = \sigma_\epsilon^2 I_q \), setting \( K = (\Sigma_0 \otimes I_p) \Lambda^{-1} \) yields,
\[ \hat{\gamma}_{RR} = \sigma_\epsilon^{-2} \Lambda \hat{Z}^T \left[ \sigma_\epsilon^{-2} \Lambda \hat{Z}^T + I \right]^{-1} y \\
= \Lambda \hat{Z}^T \left[ \hat{Z} \Lambda \hat{Z}^T + \Sigma \right]^{-1} y \\
= \hat{\gamma}_{BLUP} \]

This is a well known connection between the RR estimator and BLUP which proves the following result:

**Proposition 1.** Assuming \( \hat{\Sigma}_0 \propto I \), the BLUP estimator of the third step of the MrRCE algorithm is equivalent to the multivariate RR estimator with Ridge matrix \( K = (\hat{\Sigma}_0 \otimes I_p) \hat{\Lambda}^{-1} \).

To better understand this result, consider the case \( \Sigma_0 = \sigma_\epsilon^2 I_q \) and \( \Lambda_0 = \sigma_\gamma^2 C \), where \( C = C_\rho \) is an equicorrelation matrix with parameter \( \rho \). Let \( K = (\Sigma_0 \otimes I_p) \Lambda^{-1} = \eta C^{-1} \otimes I_p \) where \( \eta = (\sigma_\epsilon/\sigma_\gamma)^2 \). It is easy to verify that \( C^{-1} \) is itself an equicorrelation matrix, \( C^{-1} = aI_q + bJ_q \), where,
\[ a = \frac{1}{1-\rho}, \quad b = \frac{-\rho}{1-\rho} \left[ \frac{1}{1+(q-1)\rho} \right] \]

For simplicity, we only examine the penalty structure for \( q = 2, p = 1 \). Denote the coefficients vector by \( \gamma = (\gamma_{11}, \gamma_{12})^T \). The ridge penalty is given by,
\[ \eta \left[ \gamma^T C^{-1} \gamma \right] = \eta \left[ (a + b) \| \gamma \|^2_2 + 2b \gamma_{11} \cdot \gamma_{12} \right] \]
\[ = \eta \frac{1}{1-\rho^2} \| \gamma \|^2_2 + 2\eta b (\gamma_{11} \cdot \gamma_{12}) \]

Note that (12) can be reduced to the univariate ridge penalty by setting \( \rho = 0 \), i.e., by considering i.i.d coefficients. For \( \rho > 0 \), the second term in (12) kicks-in. We note that \( b < 0 \) for \( \rho \in (0,1) \), meaning that the second penalty term in (12) is negative, for same sign coefficients. This simple example illustrates that the MrRCE method favors equal sign coefficients, within groups.

### 4 Simulation Study

In this section, we compare the performance of the MrRCE method to other multivariate regression estimators, over several settings of simulated data sets. We show that the Mr-
RCE method outperforms all competitors, in terms of Model Error, for the vast majority of simulated settings.

4.1 Estimators
We construct estimators using natural competitors of the MrRCE method, and report the results for the following methods:

1. **Ordinary Least Squares (OLS)**: Perform $q$ separate LS regressions.

2. **Group Lasso**: Place an $L_1/L_2$-penalty over the rows of the coefficient matrix, with 3-fold cross-validation (CV) for the selection the tuning parameter.

3. **Ridge Regression**: The tuning parameter is selected via leave-one-out cross-validation (LOO-CV) and is shared across all task.

4. **MRCE**: The tuning parameters are selected using 5-fold CV.

5. **MrRCE**: The $L_1$-regularization parameter (for the graphical lasso algorithm) is selected via 3-fold CV.

4.2 Models
For each settings and every replication, we generate an $n \times p$ predictor matrix $Z$ with rows drawn independently from $N_p(0, \Sigma_Z)$, where $(\Sigma_Z)_{ij} = \rho |i-j|$ and $\rho = .7$ (similar to Yuan, Ekici, et al. (2007), Peng et al. (2010), and Rothman, Levina, and Zhu (2010)). Following Rothman, Levina, and Zhu (2010), the coefficient matrix $\Gamma$ is generated as the element-wise product of three matrices: First, we sample a $p \times q$ matrix $W \sim MVN_{p \times q}(0, I_p, \sigma^2 C_p)$, with $C_p = I + \rho (J - I)$, where $J$ is a matrix of ones and $I$ is the identity matrix, both of dimensions $q \times q$. The values of $\rho$ are ranging from 0 to 0.8, where $\rho = 0$ corresponds to i.i.d samples, $\gamma_{ij} \sim N(0, \sigma^2)$. Next, we set,

$$\Gamma = W \odot K \odot Q$$

where $\odot$ denotes the element-wise product. The entries of the $p \times q$ matrix $K$ are drawn independently from Ber($1 - s$), and the elements in each row of the matrix $Q$ are all equal zero or one, according to $p$ independent Bernoulli draws with success probability $1 - s_g$. Hence, setting $s, s_g > 0$ will induce element-wise and group sparsity in $\Gamma$. The rows of the error matrix $E$ are drawn independently from $N_q(0, \Sigma)$. We consider several structures for the error covariance matrix, specified in the form of the transformed error covariance matrix, $\tilde{\Sigma} := U^T \Sigma U$, where $U$ is the orthogonal matrix obtained via eigendecomposition over the matrix $C_p$ (see Eq. [7]):

1. **Independent Errors.** The errors are drawn i.i.d form $N_q(0, I_q)$. 

   11
2. **Autoregressive Error Covariance — AR** (1). We let $\tilde{\Sigma}_{ij} = \rho_E^{\left| i-j \right|}$. The transformed error covariance matrix is dense, whereas the precision matrix $\tilde{\Omega}$ is a sparse, banded matrix.

3. **Fractional Gaussian Noise (FGN).** The transformed error covariance matrix is given by,

$$
\tilde{\Sigma}_{i,j} = 5 \left[ \left( |i-j| + 1 \right)^{2H} - 2 |i-j|^{2H} + (|i-j| - 1)^{2H} \right]
$$

with $H = .95$. Both the transformed error covariance matrix $\tilde{\Sigma}$ and its inverse have a dense structure.

4. **Equicorrelation Covariance Structure.** We let $\tilde{\Sigma}_{ij} = \rho_E$ for $j \neq i$, and $\tilde{\Sigma}_{ij} = 1$ for $j = i$. Both the transformed error covariance matrix and its inverse have a dense structure.

### 4.3 Performance Measure

For a given realization of the coefficient matrix and method $m$, and for each replication $r$, let $\gamma_j^{(r)}$ denote the true coefficient vector and $\hat{\gamma}_j^{(r)}(m)$ denote the estimated coefficient vector, both for the $j$th response. The mean-squared estimation error is given by,

$$
ME_m^{(r)} \left( \gamma_j^{(r)}, \hat{\gamma}_j^{(r)}(m) \right) = \int \left[ \left( \gamma_j^{(r)} - \hat{\gamma}_j^{(r)}(m) \right)^T z \right] p(z) dz
$$

$$
= \left( \gamma_j^{(r)} - \hat{\gamma}_j^{(r)}(m) \right)^T \Sigma_Z \left( \gamma_j^{(r)} - \hat{\gamma}_j^{(r)}(m) \right)
$$

where $p(z)$ and $\Sigma_Z$ are the density function and covariance matrix of $z$, respectively. We evaluate the performance using the model error (ME), following the approach of Breiman and Friedman (1997), Yuan, Ekici, et al. (2007), and Rothman, Levina, and Zhu (2010),

$$
ME_m^{(r)} \left( \Gamma^{(r)}, \hat{\Gamma}^{(r)}(m) \right) = \text{tr} \left[ \left( \Gamma^{(r)} - \hat{\Gamma}^{(r)}(m) \right)^T \Sigma_Z \left( \Gamma^{(r)} - \hat{\Gamma}^{(r)}(m) \right) \right]
$$

The ME over all $N$ replications is averaged to obtain our performance measure,

$$
ME_m = \frac{1}{N} \sum_{r=1}^{N} ME_m^{(r)}
$$

### 4.4 Results

We simulate $N = 200$ replications with $n = 50$, $p = 20$ and $q = 5$, for each setting. The correlation parameter $\rho$ ranges from 0 to 0.8, with 0.2 steps. We note that the MRCE method often failed to converge for medium-high values of the correlation parameter, leading to inferior performance (using the default number of maximal iterations in the R-package implementation). Significance tests were performed using paired $t$–test.
**Independent Errors.** We first consider an identity error covariance structure, $\tilde{\Sigma} = I_q$, and set the sparsity and group sparsity levels at $s = 0.2, s_g = 0$. Hence, for $\rho = 0$ (*i.i.d* sampled coefficients), we do not expect any advantage for our method, over the competitors. The average ME is displayed in Figure 2. Indeed, for $\rho = 0$, all methods perform nearly the same in terms of ME, aside from the OLS method, which performs significantly worse (none of the methods significantly outperform all others). For $\rho > 0$, the MrRCE method achieves significant improvement over all competitors (all $p$-values $< 10e - 5$).

![Figure 2: Independent Errors. Average model error (ME) versus the correlation parameter $\rho$, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = 0.2, s_g = 0$.](image)

**Autoregressive (AR).** Let $\Sigma_{ij} = \rho_{E} |i - j|$, with $\rho_{E} = 0.75$. We use two settings for the sparsity levels, $s = s_g = 0$, and $s = s_g = 0.1$. Although the transformed precision matrix is a sparse, banded matrix, the assumptions of MrRCE only partially hold, as we induce sparsity in $\Gamma$ as well. The results are displayed in Figure 3. For both settings, the MrRCE method achieves the best ME performance, with a significant improvement over competing methods (all $p$-values $< 10e - 3$).

**Fractional Gaussian Noise.** This covariance structure for the error terms was also considered by Rothman, Levina, and Zhu (2010). We construct a dense coefficient matrix, by setting $s = s_g = 0$. The results are presented in Figure 4, showing that our proposed method provides a considerable improvement over competitors (all $p$-values $< 10e - 16$). The margin by which MrRCE outperforms the other methods increases with $\rho$.

**Equicorrelation.** Finally, we let $\tilde{\Sigma}_{ij} = \rho_{E} = 0.9$ for $i \neq j$, and set $s = s_g = 0.1$. The results are displayed in Figure 5. The MRCE method exploits the correlated errors, achieving better performance than the lasso, ridge and OLS methods, and is second only to MrRCE, which significantly outperforms all competitor methods for all values of $\rho$ (all $p$-values $< 10e - 5$).
Figure 3: *Autoregressive*. Average model error (ME) versus the correlation parameter $\rho$, based on $N = 200$ replications with $n = 50, p = 20, q = 5$. Top: $s = s_g = 0$. Bottom: $s = s_g = 0.1$.

5 Applications

We consider two publicly available real-life datasets:

1. *NYC Taxi Rides*\(^2\) The data consists of the daily number of New-York City (NYC) taxi rides, ranging from January 2016 to December 2017.

2. *Avocado Prices*\(^3\) The data was provided by the Hass Avocado Board website and represents weekly retail scan data for national retail volume (units) and price.

We measure and report the performance of the following methods:

1. *Ordinary Least Squares.*

2. *Group Lasso.* Apply 3-fold CV for the selection of the tuning parameter.

3. *Separate Lasso.* Perform $q$ separate lasso regression models with 3-fold CV for selecting the tuning parameters.

4. *Ridge Regression.* Perform $q$ separate ridge regression models, with shared regularization parameter, selected via LOO-CV (e.g. same ridge penalty for all $pq$ parameters).

5. *Separate Ridge Regression.* Perform $q$ separate ridge regression models with LOO-CV for selecting the tuning parameters.

6. *MRCE.* Apply 5-fold CV for selecting the regularization parameters.

---

\(^2\)Non-aggregated data is available at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

\(^3\)The data is available at https://www.kaggle.com/neuromusic/avocado-prices.
Figure 4: Fractional Gaussian Noise. Average model error (ME) versus the correlation parameter $\rho$, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = s_g = 0$.

7. MrRCE. Apply 3-fold CV for selecting the graphical lasso regularization parameter.

**NYC Taxi Rides.** We consider the problem of forecasting the performance of $q = 2$ taxi vendors in NYC, using historical records of the daily number of rides, spanning from January 2016 to December 2017 ($n = 730$). This multivariate time-series data is generated according to human activities and actions, and as such can be expected to be strongly affected by multiple seasonalities and holidays effects (see Figure 6). For a regular period $P$, we utilize the Fourier series to model the periodic effects (Andrew and Neil, 1993; Taylor and Letham, 2018), by constructing $2 \cdot N_P$ features of the form,

$$Z_P(t) = \left\{ \cos \left( \frac{2\pi nt}{P} \right), \sin \left( \frac{2\pi nt}{P} \right) \right\}_{n=1,\ldots,N_P}$$

We account for the weekly and yearly seasonalities and introduce the corresponding $P$-cyclic covariates. For a holiday $H$, which occurs at times $T(H)$, we use a simple indicator predictors of the form,

$$Z_H(t) = 1_{\{t \in T(H)\}}$$

Lastly, we incorporate covariates for the modeling of a piecewise linear trend. These transformations shift the multivariate time-series problem into a feature space with $p = 68$, where the linear assumption is appropriate. We denote the transformed observations by,

$$\{Z(t), Y(t)\}_{t=1,\ldots,T}$$

where $Z(t) \in \mathbb{R}^p$ contains measurements of the covariates, $Y(t) \in \mathbb{R}^q$ contains the $q$ responses, and $Y_j(t) \in [0, 1]$ represents the scaled response of the $j$th task at time $t$, obtained
Figure 5: *Equicorrelation*. Average model error (ME) versus the correlation parameter $\rho$, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = s_g = 0.1$.

by dividing the original observation by the maximal response value for that given task.

Figure 6: *NYC Taxi Rides*. Daily number of NYC taxi rides, for two vendors, spanning from January 2016 to December 2017.

We evaluate the forecast performance of the different methods using cross-validation like approach, in which we produce $K$ forecasts at multiple cutoff points along the history (Taylor and Letham, 2018). For cutoff $k = 0, \ldots, K - 1$, we use the first $n_{\text{train},k} = 365 + k \cdot 14$ days for training, and the next $n_{\text{test}} = 14$ observations as the test set. The performance of method
Over the $k$-th “fold” is measured according to the Mean Squared Error (MSE),

$$\text{MSE}^m_k = \frac{1}{n_{\text{test}}} \cdot \frac{1}{q} \sum_{t \in T_k} \sum_{j=1}^q (y_{j,t} - \hat{y}_{j,t}^m)^2$$

where $T_k$ are the time indices for the $k$-th test set, and $\hat{y}_{j,t}^m$ is the forecast for the $j$-th task at time $t$, produced using method $m$. Using the above procedure, we obtain $K = 26$ realizations of the MSE, $\{\text{MSE}^m_k\}_{k=0}^{K-1}$, for each method $m$. The mean and standard deviation of the MSE for each of the methods are reported in Table 1. The MrRCE method attains the best forecast performance, with lowest mean MSE and smallest standard deviation, followed by the Ridge and Separate-Ridge methods. A paired $t$-test confirms that the improvement in accuracy achieved by our method is significant (all $p$-values $< 0.05$). We also note that the estimated similarity level for this data is $\hat{\rho} = 0.99$.

| Model   | Mean   | Std    |
|---------|--------|--------|
| MrRCE   | 3.83e-3| 4.50e-3|
| Ridge   | 4.59e-3| 5.34e-3|
| Sep. Ridge | 4.59e-3| 5.34e-3|
| MRCE    | 4.61e-3| 5.12e-3|
| Group Lasso | 5.68e-3| 7.72e-3|
| Sep. Lasso | 5.75e-3| 7.12e-3|
| OLS     | 2.00e-2| 1.40e-2|

Table 1: NYC Taxi Rides. NYC daily taxi rides prediction. Mean and standard deviation of the MSE, estimated over $K = 26$ cutoffs.

**Avocado Prices.** We consider the weekly average avocado prices for $q = 5$ regions in the US, spanning from January 2015 to April 2018 ($n = 169$). We use national volume metrics and one-hot encoding for years ($p = 12$) to predict the average avocado prices for each region. The performance is measured according to the MSE, with 10-fold CV. The mean and standard deviation of the MSE, calculated over all folds, are reported in Table 2. Our proposed method attains the best prediction performance, with lowest mean MSE and smallest standard deviation. A paired $t$-test confirms that the improvement in accuracy is significant (all $p$-values $< 0.05$). We also report the estimated similarity level for this data, at $\hat{\rho} = 0.699$.

## 6 Summary and Discussion

We have presented the MrRCE method to produce an estimator of the covariance components and a predictor of the multivariate regression coefficient matrix. Our method exploits similarities among random coefficients and accounts for correlated errors. We have pro-
| Model       | Mean  | Std   |
|-------------|-------|-------|
| MrRCE       | 52.4e-2 | 21.1e-2 |
| MRCE        | 63.4e-2 | 29.0e-2 |
| Group Lasso | 64.9e-2 | 29.9e-2 |
| Sep. Ridge  | 71.0e-2 | 38.7e-2 |
| Ridge       | 71.5e-2 | 39.8e-2 |
| Sep. Lasso  | 72.0e-2 | 36.0e-2 |
| OLS         | 73.1e-2 | 41.3e-2 |

Table 2: *Avocado Prices*. Avocado price prediction. Mean and standard deviation of the MSE, estimated over $K = 10$ folds.

posed an efficient algorithm for computing MrRCE. By using simulated and real data, we have illustrated that the proposed method can outperform the commonly used methods for multivariate regression, in settings were errors or coefficients are related.

Our method can be extended in several ways. For example, one could consider an arbitrary group structure over the coefficient matrix, model the similarities via different covariance structure, or allow for per-group similarity coefficient. In addition, as pointed out in Rothman, Levina, and Zhu [2010], one could use other penalties over the precision matrix $\Omega$, in particular, ones which will introduce less bias.

7 Acknowledgement

This research was partially supported by Israeli Science Foundation grant 1804/16.

References

Anderson, Theodore Wilbur (1951). “Estimating linear restrictions on regression coefficients for multivariate normal distributions.” In: *The Annals of Mathematical Statistics*, pp. 327–351.

Andrew, Harvey C and Shephard Neil (1993). “Structural time series models.” In: *Econometrics*. Vol. 11. Handbook of Statistics. Elsevier, pp. 261–302.

Breiman, Leo and Jerome H Friedman (1997). “Predicting multivariate responses in multiple linear regression.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.1, pp. 3–54.

Brown, Philip J and James V Zidek (1980). “Adaptive multivariate ridge regression.” In: *The Annals of Statistics*, pp. 64–74.

Bunea, Florentina, Yiyuan She, and Marten H Wegkamp (2011). “Optimal selection of reduced rank estimators of high-dimensional matrices.” In: *The Annals of Statistics*, pp. 1282–1309.
Chatfield, Chris, Jim Zidek, and Jim Lindsey (2010). *An introduction to generalized linear models*. Chapman and Hall/CRC.

Chen, Lisha and Jianhua Z Huang (2016). “Sparse reduced-rank regression with covariance estimation.” In: *Statistics and Computing* 26.1-2, pp. 461–470.

dÁspremont, Alexandre, Onureena Banerjee, and Laurent El Ghaoui (2008). “Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data.” In: *Journal of Machine learning research* 9.Mar, pp. 485–516.

Dawid, Philip A (1981). “Some matrix-variate distribution theory: notational considerations and a Bayesian application.” In: *Biometrika* 68.1, pp. 265–274.

Deshpande, Sameer K, Veronika Rockova, and Edward I George (2017). “Simultaneous Variable and Covariance Selection with the Multivariate Spike-and-Slab Lasso.” In: *arXiv preprint arXiv:1708.08911*.

Friedman, Jerome H, Trevor Hastie, and Robert Tibshirani (2008). “Sparse inverse covariance estimation with the graphical lasso.” In: *Biostatistics* 9.3, pp. 432–441.

Furlotte, Nicholas A and Eleazar Eskin (2015). “Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model.” In: *Genetics* 200.1, pp. 59–68.

Gupta, Arjun K and Daya K Nagar (2018). *Matrix variate distributions*. Chapman and Hall/CRC.

Henderson, Charles R (1975). “Best linear unbiased estimation and prediction under a selection model.” In: *Biometrics*, pp. 423–447.

— (1984). “Applications of linear models in animal breeding: University of Guelph.” In: *Applications of linear models in animal breeding, University of Guelph*.

Henderson, Charles R, Oscar Kempthorne, Shayle R Searle, and CM Von Krosigk (1959). “The estimation of environmental and genetic trends from records subject to culling.” In: *Biometrics* 15.2, pp. 192–218.

Hoerl, Arthur E and Robert W Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems.” In: *Technometrics* 12.1, pp. 55–67.

Hsieh, Cho-Jui, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik (2011). “Sparse inverse covariance matrix estimation using quadratic approximation.” In: *Advances in neural information processing systems*, pp. 2330–2338.

Izenman, Alan Julian (1975). “Reduced-rank regression for the multivariate linear model.” In: *Journal of multivariate analysis* 5.2, pp. 248–264.

Kang, Hyun Min, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. (2010). “Variance component model to account for sample structure in genome-wide association studies.” In: *Nature genetics* 42.4, p. 348.

Kim, Seyoung and Eric P Xing (2012). “Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping.” In: *The Annals of Applied Statistics* 6.3, pp. 1095–1117.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). “A mixed-model approach for genome-wide association studies of correlated traits in structured populations.” In: *Nature genetics* 44.9, p. 1066.
Kruuk, Loeske EB (2004). “Estimating genetic parameters in natural populations using the 'animal model'.” In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 359.1446, pp. 873–890.

Lee, Wonyul and Yufeng Liu (2012). “Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood.” In: *Journal of multivariate analysis* 111, pp. 241–255.

Li, Yanning, Bin Nan, and Ji Zhu (2015). “Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure.” In: *Biometrics* 71.2, pp. 354–363.

Obozinski, Guillaume R, Martin J Wainwright, and Michael I Jordan (2009). “High-dimensional support union recovery in multivariate regression.” In: *Advances in Neural Information Processing Systems*, pp. 1217–1224.

Peng, Jie, Ji Zhu, Anna Bergamaschi, Wonsik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang (2010). “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer.” In: *The annals of applied statistics* 4.1, p. 53.

Rothman, Adam J, Peter J Bickel, Elizaveta Levina, and Ji Zhu (2008). “Sparse permutation invariant covariance estimation.” In: *Electronic Journal of Statistics* 2, pp. 494–515.

Rothman, Adam J, Elizaveta Levina, and Ji Zhu (2010). “Sparse multivariate regression with covariance estimation.” In: *Journal of Computational and Graphical Statistics* 19.4, pp. 947–962.

Sherman, Jack and Winifred J Morrison (1950). “Adjustment of an inverse matrix corresponding to a change in one element of a given matrix.” In: *The Annals of Mathematical Statistics* 21.1, pp. 124–127.

Sohn, Kyung-Ah and Seyoung Kim (2012). “Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization.” In: *Artificial Intelligence and Statistics*, pp. 1081–1089.

Taylor, Sean J and Benjamin Letham (2018). “Forecasting at scale.” In: *The American Statistician* 72.1, pp. 37–45.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso.” In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Turlach, Berwin A, William N Venables, and Stephen J Wright (2005). “Simultaneous variable selection.” In: *Technometrics* 47.3, pp. 349–363.

Vattikuti, Shashaank, Juen Guo, and Carson C Chow (2012). “Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits.” In: *PLoS genetics* 8.3, e1002637.

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

Wilms, Ines and Christophe Croux (2018). “An algorithm for the multivariate group lasso with covariance estimation.” In: *Journal of Applied Statistics* 45.4, pp. 668–681.
Witten, Daniela M and Robert Tibshirani (2009). “Covariance-regularized regression and classification for high dimensional problems.” In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71.3, pp. 615–636.

Woodbury, Max A (1950). “Inverting modified matrices.” In: Memorandum report 42.106, p. 336.

Yin, Jianxin and Hongzhe Li (2011). “A sparse conditional gaussian graphical model for analysis of genetical genomics data.” In: The annals of applied statistics 5.4, p. 2630.

Yuan, Ming, Ali Ekici, Zhaosong Lu, and Renato Monteiro (2007). “Dimension reduction and coefficient estimation in multivariate linear regression.” In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69.3, pp. 329–346.

Yuan, Ming and Yi Lin (2006). “Model selection and estimation in regression with grouped variables.” In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1, pp. 49–67.
— (2007). “Model selection and estimation in the Gaussian graphical model.” In: Biometrika 94.1, pp. 19–35.

Zhou, Xiang and Matthew Stephens (2014). “Efficient multivariate linear mixed model algorithms for genome-wide association studies.” In: Nature methods 11.4, p. 407.