Electronic Journal of Statistics
Vol. 14 (2020) 3821–3844
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1764

Capturing between-tasks covariance and similarities using multivariate linear mixed models

Aviv Navon*

Bar-Ilan University, Israel
e-mail: avivnav@gmail.com

and

Saharon Rosset

Tel Aviv University, Israel
e-mail: saharon@post.tau.ac.il

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 multivariate regression models, that directly capture and model the within-group similarities, by employing a multivariate linear mixed model formulation, with a joint estimation of covariance matrices for coefficients and errors via penalized likelihood. Our approach, which we term MrRCE for Multivariate random Regression with Covariance Estimation, encourages structured similarity in parameters, in which coefficients for the same variable in related tasks share 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 and phrases: Covariance selection, EM algorithm, multivariate regression, penalized likelihood, regularization methods, sparse precision matrix.

Received November 2019.

1. Introduction

In many cases, a common set of predictor variables is used for predicting different but related target variables. For example, we may be interested in modeling and predicting the price of a single product in multiple markets, given a set of product and market characteristics and historical observations. Or we may be interested in modeling daily demand for a similar product offered by multiple

*This work was done while the author was at Tel Aviv University, Israel.

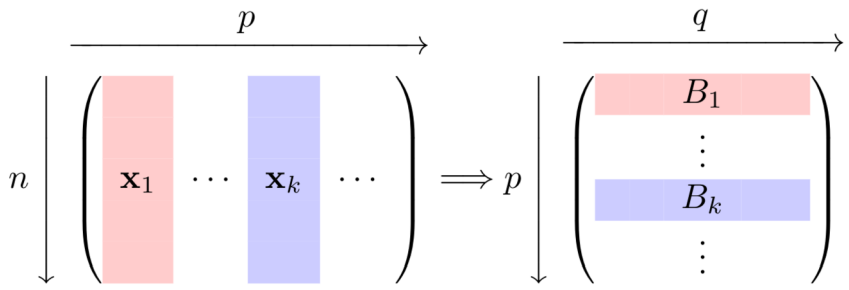


FIG 1. *The multivariate regression framework naturally induces a group structure over the coefficient matrix B, in which every explanatory variable, $\mathbf{x}_i$, corresponds to a group of q coefficients $B_i = (\beta_{i1}, ..., \beta_{iq})^T$.*

companies, given common explanatory variables. We provide real data examples along both these lines in Section 6.

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 multivariate regression, 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 ith sample, respectively. Let $X = (X_1, ..., X_n)^T = (\mathbf{x}_1, ..., \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ denote the predictor matrix and $Y = (Y_1, ..., Y_n)^T = (\mathbf{y}_1, ..., \mathbf{y}_q) \in \mathbb{R}^{n \times q}$ 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 Σ is the among-tasks covariance matrix. The multivariate regression model is given by,

$$Y = XB + E, \tag{1.1}$$

where B is a $p \times q$ regression coefficient matrix. The random matrices in (1.1) are assumed to follow a matrix-variate normal distribution [10, 16], with $E \sim MVN_{n \times q}(0, I_n, \Sigma)$ and $Y \sim MVN_{n \times q}(XB, I_n, \Sigma)$. 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 the full information contained in the data [3, 32, 41].

The multivariate regression framework naturally induces a group structure over the coefficient matrix, B, in which every explanatory variable, $\mathbf{x}_i$ for $i = 1, ..., p$, corresponds to a group of q coefficients, $B_i = (\beta_{i1}, ..., \beta_{iq})$ (see Figure 1). To utilize the relatednesses in coefficients and responses, previous MR methods focus on jointly learning all tasks, while (i) Accounting for the between tasks covariance [3, 32, 44, 27, 41]; and (ii) Encouraging similarities in the regression coefficients for the different tasks [38, 29, 24, 28]. The latter is commonly achieved by enforcing a group structure over the coefficient matrix, and encouraging the sparsity of the entire group.

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 a 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, in addition to a similar magnitude. For instance, consider the demand prediction problem mentioned above, and assume we want to model the daily demand for $q = 2$ suppliers of a similar product. We generally expect that the effect of various explanatory variables would be similar for both models; for example, the effect of a holiday on demand will be similar for both suppliers. However, it is likely to not be identical due to differences in suppliers, client population, etc. The exact level of similarity between effects may not be known in advance.

To address this we propose a general approach for constructing an estimator for 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 random regression model with an underlying group structure over the coefficient matrix, designed to encourage related coefficients to share a common sign and similar magnitude. Specifically, we study a random variant of (1.1), which is a special case of Multivariate Linear Mixed Model (mvLMMs). mvLMMs can be viewed as a generalization of MR, allowing both fixed and random effects. Consider the MR problem (1.1), but with an additional term for the set of random predictors, collected into the matrix $Z = (Z_1, ..., Z_n)^T = (\mathbf{z}_1, ..., \mathbf{z}_r) \in \mathbb{R}^{n \times r}$. The multivariate mixed model is given by,

$$\begin{aligned} &Y = XB + Z\Gamma + E, \\ &E \sim MVN_{n\times q}\left(0, I_n, \Sigma\right), \Gamma \sim MVN_{r\times q}\left(0, R, G\right), \end{aligned} \tag{1.2}$$

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

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

$$Y = Z\Gamma + E. \tag{1.3}$$

Under this formulation we are interested in estimating the covariance components and using the model to predict new observation. This is done by first predicting the random component Γ using the Empirical-Best Linear Unbiased Predictor, E-BLUP [17, 18]. Our method accounts for correlations between responses and similarities among coefficients, captured by estimating a joint equicorrelation covariance matrix for the rows of Γ (see Eq. (2.1) 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 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 $\Omega = \Sigma^{-1}$.

To summarize, we make the following novel contributions: (i) We propose MrRCE, a new approach for multivariate regression that utilizes correlations in coefficients and responses; (ii) We propose an Expectation-Maximization (EM) [11] based computational algorithm for parameter estimation under MrRCE and an E-BLUP approach for prediction, and; (iii) We evaluate MrRCE on a variety of multivariate regression tasks. We make our source code publicly available at https://github.com/AvivNavon/MrRCE.

The remainder of the paper is structured as follows. In Section 2, we describe the MrRCE method and corresponding Expectation-Maximization (EM) based computational algorithm. In Section 3, we cover previous works on multivariate regression, demonstrating their similarities to the MrRCE model and highlighting aspects unique to MrRCE. In Section 4, we establish a connection between the proposed method and the multivariate Ridge estimator. Simulation studies are performed in Section 5 to compare our method with competing estimators, and Section 6 contains two real data applications of MrRCE.

2. The MrRCE method

Consider the random effect regression model of (1.3), with $r = p$ covariates. Assume both the error matrix E and the coefficient matrix Γ follow a matrix variate normal distribution,

$$E \sim MVN_{n\times q}\left(0, I_n, \Sigma\right), \Gamma \sim MVN_{p\times q}\left(0, I_p, \sigma^2 C\right). \tag{2.1}$$

Further assume an equicorrelation structure for the matrix C, controlled by the unknown intra-group correlation coefficient $\rho \in [0, 1)$,

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

The unknown parameter ρ can be thought of as a relative measure of the *within-group similarity* [6]. Large values for ρ correspond to high similarity among coefficients 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 Γ. We refer to the random variable Γ as unobserved data, and to (Y, Γ) as the *full data*. Denote the likelihood function of the full data by $\mathcal{L}(\cdot)$, and the collection of parameters by $\Theta = \left\{\Omega, \sigma^2, \rho\right\}$, we have

$$\begin{aligned}\mathcal{L}(Y, \Gamma; \Theta) &= \mathcal{L}_{Y|\Gamma}(Y \mid \Gamma; \Theta)\, \mathcal{L}_{\Gamma}(\Gamma \mid \Theta) \\ &= \mathcal{L}_{Y|\Gamma}(Y \mid \Gamma; \Omega)\, \mathcal{L}_{\Gamma}\left(\Gamma \mid \sigma^2, \rho\right).\end{aligned}$$

Thus, the negative log-likelihood function of the complete data is given by (up to a constant)

$$\ell(Y, \Gamma; \Theta) = \operatorname{tr}\left[\frac{1}{n}\Omega(Y - Z\Gamma)^T(Y - Z\Gamma)\right] - \log|\Omega| + \operatorname{tr}\left[\frac{1}{p}\Delta\Gamma^T\Gamma\right] - \log|\Delta|,$$

where $\Delta^{-1} = \sigma^2 C$. We construct an estimator of Θ using a penalized negative normal log-likelihood, adding an L_1-penalty over the off-diagonal entries of Ω,

$$\hat{\Theta} = \arg\min_{\Theta} \ell(Y, \Gamma; \Theta) + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}|, \tag{2.2}$$

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

2.1. The algorithm

We propose an iterative algorithm for solving (2.2), which is a variant of the Expectation-Maximization (EM) algorithm [11], for penalized likelihood [15]. We note that unlike the standard EM procedure, we minimize the expression during the M-step, effectively applying an expectation-minimization procedure; equivalently, one can negate (2.2) to retrieve the standard EM formulation. Alg. 1 provides a schematic overview of the MrRCE algorithm.

Using eigendecomposition (similar to Zhou and Stephens [48], Furlotte and Eskin [14]), we write,

$$C = UDU^T \text{ and } ZZ^T = LSL^T, \tag{2.3}$$

where S and $D := D_\rho = diag(d_1(\rho), ..., d_q(\rho))$ are diagonal matrices, and U is independent of ρ. We then multiply (1.3) 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

$$\tilde{\Gamma} = \Gamma U \sim MVN_{p\times q}\left(0, I_p, \sigma^2 U^T C U\right) = MVN_{p\times q}\left(0, I_p, \sigma^2 D_\rho\right),$$
$$\tilde{E} = L^T E U \sim MVN_{n\times q}\left(0, L^T L = I_n, \tilde{\Sigma} := U^T \Sigma U\right) = MVN_{n\times q}\left(0, I_n, \tilde{\Sigma}\right).$$

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

$$\begin{aligned} &Y = Z\Gamma + E, \\ &E \sim MVN_{n\times q}\left(0, I_n, \Sigma := \Omega^{-1}\right), \Gamma \sim MVN_{p\times q}\left(0, I_p, \sigma^2 D_\rho\right), \end{aligned} \tag{2.4}$$

namely

$$Y \sim MVN_{n\times q}\left(0, S, \sigma^2 D_\rho\right) + MVN_{n\times q}\left(0, I_n, \Sigma\right).$$

Next, we describe an EM-based algorithm for solving (2.2) under the assumptions (2.4).

E-step Denote Θ_{t-1} the estimator for Θ at iteration $t-1$. At step t, we wish to evaluate the conditional expectation of the negative log-likelihood, $\mathbb{E}\left[\ell\left(Y, \Gamma; \Theta\right) \middle| Y, \Theta_{t-1}\right]$. Equivalently, we simplify the last expression and evaluate the following,

$$Q_t^1 = \mathbb{E}\left[\left(Y - Z\Gamma\right)^T \left(Y - Z\Gamma\right) \mid Y, \Theta_{t-1}\right], \tag{2.5}$$
$$Q_t^2 = \mathbb{E}\left[\Gamma^T \Gamma \mid Y, \Theta_{t-1}\right]. \tag{2.6}$$

We let $\otimes$ denote the Kronecker product and $\mathrm{vec}\left(\cdot\right)$ the vectorization operator.[1] For a matrix $A \in \mathbb{R}^{k\times p}$, we let $A\Gamma := G = \begin{pmatrix}\mathbf{g}_1 & \cdots & \mathbf{g}_q\end{pmatrix}$, with $\mathbf{g}_j$ the jth column of G. The joint distribution of $\mathbf{g} = \mathrm{vec}\left(G\right)$ and $\mathbf{y} = \mathrm{vec}\left(Y\right)$ is given by

$$\begin{pmatrix}\mathbf{g} \\ \mathbf{y}\end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix}\Delta^{-1}\otimes AA^T & \Delta^{-1}\otimes AZ^T \\ \Delta^{-1}\otimes ZA^T & \Sigma\otimes I_n + \Delta^{-1}\otimes ZZ^T\end{bmatrix} := \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),$$

hence, the conditional distribution of $\mathbf{g} \mid \mathbf{y}$ is given by

$$\mathbf{g} \mid \mathbf{y} \sim N\left(\Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right). \tag{2.7}$$

In order to evaluate (2.5) and (2.6), we calculate the following terms $\mathbb{E}\left[\Gamma \mid Y, \Theta_{t-1}\right]$ and $\mathbb{E}\left[\Gamma^T A^T A\Gamma \mid Y, \Theta_{t-1}\right]$, for $A = I_p, Z$. The former is the E-BLUP [17, 18] (see ***Prediction*** below), whereas the latter can be easily obtained from (2.7) since

$$\mathbb{E}\left[G^T G \mid Y, \Theta_{t-1}\right]_{i,j} = \mathbb{E}\left[\mathbf{g}_i^T \mathbf{g}_j \mid \mathbf{y}, \Theta_{t-1}\right].$$

[1] Let $\mathrm{vec}\left(\cdot\right)$ denote the concatenation of a $k \times l$-dimensional matrix's columns into a kl-dimensional vector.

M-step The minimization of the objective over Θ can be split into two disjoint minimization problems:

$$\arg\min_{\Omega\succeq 0} \operatorname{tr}\left[\frac{1}{n}\Omega Q_t^1\right] - \log|\Omega| + \lambda_\omega \sum_{j\neq j'} |\omega_{jj'}|\,, \tag{2.8}$$

$$\arg\min_{\sigma>0,\rho\in[0,1)} \operatorname{tr}\left[\frac{1}{p}\Delta Q_t^2\right] - \log|\Delta|\,. \tag{2.9}$$

The first minimization problem is exactly the L_1-penalized precision matrix estimation problem considered by Yuan and Lin [46], dÁspremont, Banerjee and El Ghaoui [9], Friedman, Hastie and Tibshirani [13], Rothman, Levina and Zhu [32], Hsieh et al. [21], among others. We solve (2.8) by applying the graphical lasso algorithm of Friedman, Hastie and Tibshirani [13]. The second minimization problem, (2.9), can be easily solved in closed-form by utilizing the diagonal form of Δ.

Prediction Given $\hat{\Theta}$, our estimate for Θ, we apply the model for prediction by computing the E-BLUP [17, 18] for $\boldsymbol{\gamma} = \operatorname{vec}(\Gamma)$. Denote, $\tilde{Z} = I_q \otimes Z$, $L = \hat{\sigma}^2 \hat{D}_\rho \otimes I_p$ and $R = \hat{\Omega}^{-1} \otimes I_n$, the E-BLUP $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$, is given by,

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

Alternatively, as proved by Henderson et al. [19], $\hat{\boldsymbol{\gamma}} = L^T \tilde{Z}^T \Psi^{-1} \mathbf{y}$ where, $\Psi = \tilde{Z} L \tilde{Z}^T + R$. In order to predict Γ, we simply compute $\hat{\Gamma} = \operatorname{unvec}(\hat{\boldsymbol{\gamma}})$, where $\operatorname{unvec}(\cdot)$ represents the reversal of the $\operatorname{vec}(\cdot)$ operation.

Starting Value and Stopping Criteria We initialize $\Omega_0 = I_q$, $\Delta^{-1} = I_q$, and consider two alternatives for the MrRCE algorithm's stopping criterion.

1. Set a tolerance value, $\tau > 0$, and let $\ell_{pen,t}$ denote the penalized negative log-likelihood at iteration t. Iterate until the relative change in the log-likelihood value, $\left|\frac{\ell_{pen,t-1} - \ell_{pen,t}}{\ell_{pen,t-1}}\right|$, is smaller than τ.
2. Set a tolerance value, $\tau > 0$. Iterate until the sum of absolute changes in the values of Θ in two successive iterations is smaller than the tolerance value.

Convergence The MrRCE algorithm is a variant of the EM algorithm for penalized likelihood, hence each step ensures a decrease in the objective [15]. Thus, the sequence $\{\ell_{pen,t}\}_t$ converges to some value ℓ^*_{pen}, provided that the penalized negative log-likelihood is bounded below. We further discuss the convergence of the MrRCE algorithm in Appendix B.

3. Related work

In this section, we review previous works on multivariate regression. Here we focus on the frequentist approach, and refer the readers to Deshpande, Rockova

Algorithm 1 (MrRCE): EM-based optimization procedure (see text for details)

Require: Regularization parameter $\lambda_\omega > 0$.
1: **Initialize:** set $t = 0$ and $\Omega_t = \Delta_t^{-1} = I_q$.
2: **repeat**
$t \leftarrow t + 1$
E-step: calculate $Q_t^1 = \mathbb{E}\left[(Y - Z\Gamma)^T (Y - Z\Gamma) \mid Y, \Theta_{t-1}\right]$
and $Q_t^2 = \mathbb{E}\left[\Gamma^T \Gamma \mid Y, \Theta_{t-1}\right]$
M-step: solve $\Omega_t = \arg\min_{\Omega \succeq 0} \operatorname{tr}\left[\frac{1}{n}\Omega Q_t^1\right] - \log|\Omega| + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}|$
and $(\sigma_t, \rho_t) = \arg\min_{\sigma>0, \rho \in [0,1)} \operatorname{tr}\left[\frac{1}{p}\Delta Q_t^2\right] - \log|\Delta|$
3: **until** stopping criterion is reached.
4: **predict** Γ**:** compute the E-BLUP for Γ, $\hat{\Gamma} = \operatorname{unvec}\left(\mathbb{E}\left[\boldsymbol{\gamma} \mid \mathbf{y}, \Theta_t\right]\right)$.
5: **return** $\left(\hat{\Gamma}, \Theta_t\right)$

and George [12] and references therein, for a comprehensive review of Bayesian methods for estimation and prediction with the MR model.

Reducing the number of parameters. 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 [5] generalized the classical Reduced-Rank Regression (RRR) [1, 22, 40] to high dimensional settings, estimating a low-rank coefficient matrix by penalizing the rank of B. Yuan et al. [47] 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.

Utilizing the group structure. As stated above, the MR framework induces a group structure over the coefficient matrix. 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 [45], 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 [37] and the L_2-penalty used in ridge regression [20], aimed 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 [29]. Chen and Huang [7] utilized the group lasso penalty to achieve sparse reduced-rank regression (SRRR) with variable selction. Peng et al. [31] 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 [38] studied a different constraint function, placing an L_∞-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 [24], Li, Nan and Zhu [28]). 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 [32], Chen and Huang [8], Wilms and Croux [41] have recently shown that accounting for this additional information in MR problems can be beneficial for both coefficients estimation and prediction.

Accounting for the between-tasks covariance. In multivariate normal theory, the entries of Ω 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 [13], in which simultaneous sparsity structure identification and covariance estimation are achieved by minimizing the L_1-regularized negative log-likelihood function of Ω [46, 9, 33]. 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 [42, 32].

Rothman, Levina and Zhu [32] 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 Ω,

$$\arg\min_{B,\Omega} -\log|\Omega| + \operatorname{tr}\left[\frac{1}{n}\Omega\left(Y - XB\right)^T\left(Y - XB\right)\right] + \lambda_1 \|B\|_1 + \lambda_2 \sum_{j \neq j'} |\omega_{jj'}|, \tag{3.1}$$

where $\operatorname{tr}(\cdot)$ denotes the trace, λ_1 and λ_2 are the regularization parameters and $\omega_{jj'}$ is the (j, j') element of Ω. Lee and Liu [27] extended the approach of Rothman, Levina and Zhu [32] to allow for weighted L_1-penalties over the elements of B and Ω. Yin and Li [44] considered a similar objective to the one in (3.1), and proposed an algorithm for the sparse estimation of the coefficient and inverse covariance matrices. However, unlike Rothman, Levina and Zhu [32], their method aimed at improving the estimation of Ω, 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.

Utilizing the group structure and correlated errors. While MRCE accounts for correlated responses through the precision matrix Ω, it does not learn structured sparsity in B, essentially selecting relevant covariates for each response separately. In a recent work, Wilms and Croux [41] proposed an algorithm for the multivariate group lasso with covariance estimation, replacing the lasso penalty in (3.1) with an L_1/L_2-penalty over a pre-specified group structure.

Chen and Huang [8] 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 [35] proposed an approach for group sparse multivariate regression that can jointly learn both the response structure and regression coefficients with structured sparsity. Similarly, in this work, we propose a method that accounts for the error covariance structure and the group structure among covariates. However, unlike previous approaches, our formulation does not produce group sparsity, but instead, it encourages related coefficients to share a sign and similar magnitude.

Multivariate Linear Mixed Models (mvLMMs). mvLMMs [18] 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 [26], Kang et al. [23], Korte et al. [25], Vattikuti, Guo and Chow [39]).

4. Connection to ridge regression

We present a connection between the MrRCE method and the Ridge Regression (RR) estimator [20]. More specifically, we explore a special case in which the BLUP for Γ derived by the MrRCE algorithm is equivalent to the multivariate RR estimator [4].

Consider the model,

$$\begin{aligned}&\mathbf{y} = \tilde{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},\\ &\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \Sigma_0 \otimes I_n := \Sigma\right), \boldsymbol{\gamma} \sim N\left(\mathbf{0}, \Lambda_0 \otimes I_p := \Lambda\right).\end{aligned}$$

The joint distribution of $(\mathbf{y}, \boldsymbol{\gamma})$ is given by

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{y} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \Lambda & \Lambda\tilde{Z}^T \\ \tilde{Z}\Lambda & \tilde{Z}\Lambda\tilde{Z}^T + \Sigma \end{bmatrix}\right),$$

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

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{\text{BLUP}} &= \mathbb{E}\left[\boldsymbol{\gamma} \mid \mathbf{y}\right] \\ &= \Lambda\tilde{Z}^T\left(\tilde{Z}\Lambda\tilde{Z}^T + \Sigma\right)^{-1}\mathbf{y}.\end{aligned}$$

The RR estimator can be extended to the multivariate case as in Brown and Zidek [4]:

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

where $K \succ 0$ is the $pq \times pq$ ridge matrix. We apply the generalized Sherman-Morrison-Woodbury [34, 43] formula to the inverse of $\tilde{Z}^T\tilde{Z} + K$, to obtain

$$\hat{\boldsymbol{\gamma}}_{\text{RR}} = K^{-1}\tilde{Z}^T\left[I - \left(I + \tilde{Z}K^{-1}\tilde{Z}^T\right)^{-1}\tilde{Z}K^{-1}\tilde{Z}^T\right]\mathbf{y}. \tag{4.1}$$

Eq. (4.1) can be simplified as follow,

$$\hat{\boldsymbol{\gamma}}_{\mathrm{RR}} = K^{-1}\tilde{Z}^T\left[\tilde{Z}K^{-1}\tilde{Z}^T + I\right]^{-1}\mathbf{y}.$$

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,

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{\mathrm{RR}} &= \sigma_\epsilon^{-2}\Lambda\tilde{Z}^T\left[\sigma_\epsilon^{-2}\tilde{Z}\Lambda\tilde{Z}^T + I\right]^{-1}\mathbf{y}\\ &= \Lambda\tilde{Z}^T\left[\tilde{Z}\Lambda\tilde{Z}^T + \Sigma\right]^{-1}\mathbf{y}\\ &= \hat{\boldsymbol{\gamma}}_{\mathrm{BLUP}}.\end{aligned}$$

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 prediction for Γ obtained by the MrRCE algorithm is equivalent to the multivariate RR estimator with Ridge matrix $K = \left(\hat{\Sigma}_0 \otimes I_p\right)\hat{\Lambda}^{-1}$.*

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 ρ. 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,

$$a = \frac{1}{1-\rho},\; b = \frac{-\rho}{1-\rho}\left[\frac{1}{1+(q-1)\,\rho}\right].$$

For simplicity, we only examine the penalty structure for $q = 2, p = 1$. Denote the coefficients vector by $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12})^T$. The ridge penalty is given by,

$$\begin{aligned}\eta\left[\boldsymbol{\gamma}^T C^{-1}\boldsymbol{\gamma}\right] &= \eta\left[(a+b)\,\|\boldsymbol{\gamma}\|_2^2 + 2b\gamma_{11}\cdot\gamma_{12}\right] \qquad (4.2)\\ &= \eta\frac{1}{1-\rho^2}\,\|\boldsymbol{\gamma}\|_2^2 + 2\eta b\,(\gamma_{11}\cdot\gamma_{12}).\end{aligned}$$

Note that (4.2) 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 (4.2) kicks-in. We note that $b < 0$ for $\rho \in (0,1)$, meaning that the second penalty term in (4.2) is negative, for same sign coefficients. This simple example illustrates that the MrRCE method favors equal sign coefficients, within groups.

5. 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 MrRCE method significantly outperforms all competitors, in terms of Model Error, for the vast majority of simulated settings.

5.1. 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.

5.2. 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 $(\Sigma_Z)_{ij} = \rho_Z^{|i-j|}$ and $\rho_Z = .7$ (similar to Yuan et al. [47], Peng et al. [31], Rothman, Levina and Zhu [32]). Following Rothman, Levina and Zhu [32], the coefficient matrix Γ is generated as the element-wise product of three matrices: First, we sample a $p \times q$ matrix $W \sim MVN_{p\times q}\left(0, I_p, \sigma^2 C_\rho\right)$, with $C_\rho = I + \rho(J - I)$, where J is a matrix of ones and I is the identity matrix, both of dimensions $q \times q$. The values of ρ range from 0 to 0.8, where $\rho = 0$ corresponds to $i.i.d$ samples, $\gamma_{ij} \sim N\left(0, \sigma^2\right)$. 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 $\text{Ber}(1 - s)$, 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 Γ. Additional experiments using different distribution for W are presented in Appendix A. 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_ρ (see Eq. (2.3)):

1. *Independent Errors.* The errors are drawn i.i.d form $N_q(0, I_q)$.
2. *Autoregressive Error Covariance — $AR(1)$.* We let $\tilde{\Sigma}_{ij} = \rho_E^{|i-j|}$. 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,

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

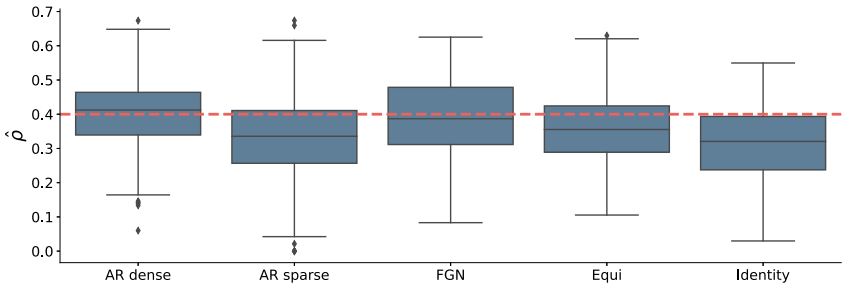


FIG 2. *Estimated similarity level. The estimated similarity level for* $\rho = 0.4$ *(red horizontal dashed line) by simulation settings.*

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.

5.3. Performance measure

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

$$\begin{aligned} ME_m^{(r)}\left(\boldsymbol{\gamma}_j^{(r)}, \hat{\boldsymbol{\gamma}}_j^{(r)}(m)\right) &= \int \left[\left(\boldsymbol{\gamma}_j^{(r)} - \hat{\boldsymbol{\gamma}}_j^{(r)}(m)\right)^T z\right]^2 p(z)\, dz \\ &= \left(\boldsymbol{\gamma}_j^{(r)} - \hat{\boldsymbol{\gamma}}_j^{(r)}(m)\right)^T \Sigma_Z \left(\boldsymbol{\gamma}_j^{(r)} - \hat{\boldsymbol{\gamma}}_j^{(r)}(m)\right), \end{aligned}$$

where $p(z)$ and Σ_Z are the density function and covariance matrix of z, respectively. We evaluate the performance using the model error (ME), following Breiman and Friedman [3], Yuan et al. [47], Rothman, Levina and Zhu [32],

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

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

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

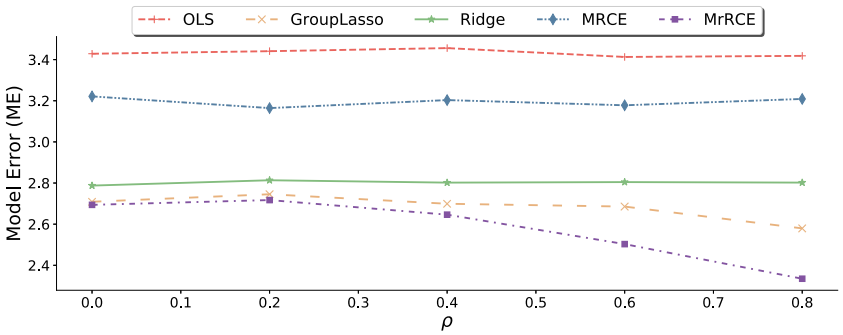


FIG 3. *Independent Errors. Average model error (ME) versus the correlation parameter ρ, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = 0.2, s_g = 0$.*

5.4. Results

We simulate $N = 200$ replications with $n = 50$, $p = 20$ and $q = 5$, for each setting. Additional experiments using larger values for n and p are presented in Appendix A. The correlation parameter ρ ranges from 0 to 0.8, with 0.2 steps. Significance tests were performed using paired t-test. We report the estimates for $\rho = 0.4$ under the different error covariance structures in Figure 2.

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 small values of ρ we do not expect any advantage for our method over the competitors. The average ME is displayed in Figure 3. Indeed, for $\rho = 0, .2$, our method achieves no significant improvement over Group Lasso. For $\rho > .2$, the MrRCE method achieves significant improvement over all competitors (all p-values $< 1\text{e} - 2$).

Autoregressive (AR). Let $\tilde{\Sigma_{ij}} = \rho_E^{|i-j|}$, with $\rho_E = 0.75$. We use two settings for the sparsity levels, $s = s_g = 0$ — dense AR, and $s = s_g = 0.1$ — sparse AR. Although the transformed precision matrix is a sparse, banded matrix, the assumptions of MrRCE only partially hold, as we induce sparsity in Γ as well. The results are displayed in Figure 4. For both settings, the MrRCE method achieves the best ME performance, with a significant improvement over competing methods (all p-values $< 1\text{e} - 3$).

Fractional Gaussian Noise. This covariance structure for the error terms was also considered by Rothman, Levina and Zhu [32]. We construct a dense coefficient matrix, by setting $s = s_g = 0$. The results are presented in Figure 5, showing that our proposed method provides a considerable improvement over competitors (all p-values $< 1\text{e} - 19$). The margin by which MrRCE outperforms the other methods increases with ρ.

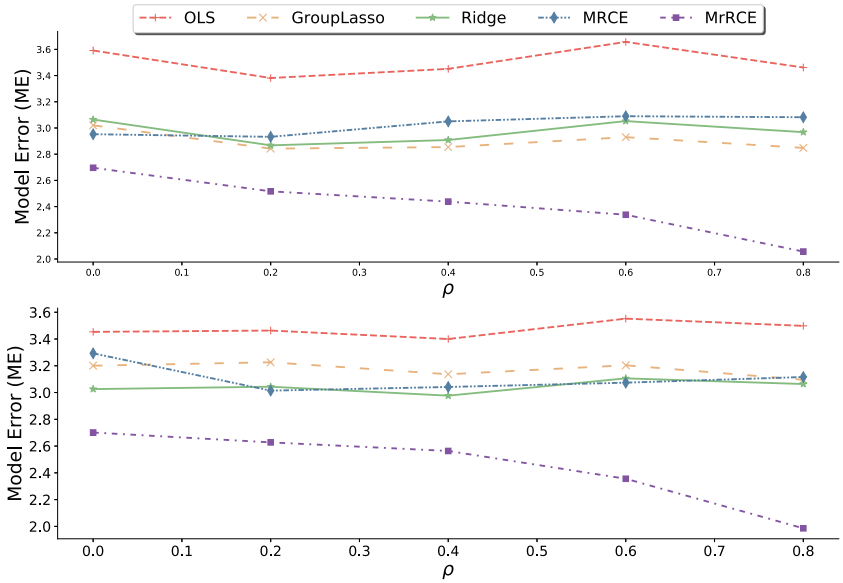


FIG 4. *Autoregressive. Average model error (ME) versus the correlation parameter ρ, based on $N = 200$ replications with $n = 50, p = 20, q = 5$. Top: sparse AR ($s = s_g = 0.1$); Bottom: dense AR ($s = s_g = 0$).*

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 6. The MRCE method exploits the correlated errors, achieving better performance than the Group Lasso, Ridge and OLS methods, and is second only to MrRCE, which significantly outperforms all competitor methods for all values of ρ (all p-values $< 1e-8$).

6. 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.

[2]The data is available at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.
[3]The data is available at https://www.kaggle.com/neuromusic/avocado-prices.

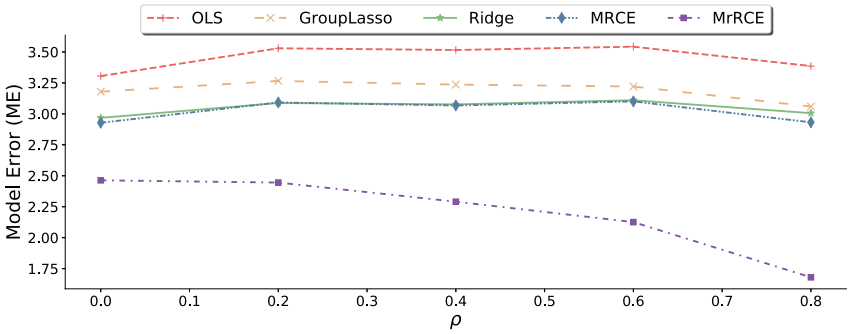


FIG 5. *Fractional Gaussian Noise. Average model error (ME) versus the correlation parameter ρ, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = s_g = 0$.*

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.
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. For a regular period P, we utilize the Fourier series to model the periodic effects [2, 36], by constructing $2 \cdot N_P$ features of the form

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

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

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

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

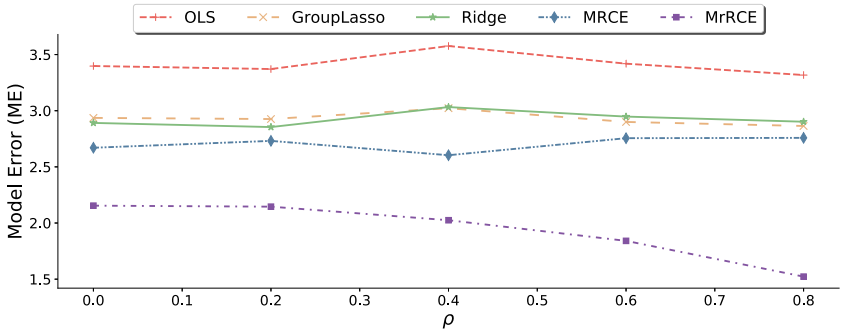


FIG 6. *Equicorrelation. Average model error (ME) versus the correlation parameter ρ, based on $N = 200$ replications with $n = 50, p = 20, q = 5$ and sparsity levels $s = s_g = 0.1$.*

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 jth task at time t, obtained by dividing the original observation by the maximal response value for that given task.

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 [36]. For cutoff $k = 0, ..., K-1$, we use the first $n_{train,k} = 365 + k \cdot 14$ days for training, and the next $n_{test} = 14$ observations as the test set. The performance of method m over the kth "fold" is measured according to the Mean Squared Error (MSE),

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

where T_k are the time indices for the kth test set, and $\hat{y}_{j,t}(m)$ is the forecast for the jth task at time t, produced using method m. Using the above procedure, we obtain $K = 26$ realizations of the MSE, $\{MSE_k^m\}_{k=0}^{K-1}$, 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 $\hat{\rho} = 0.992$.

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

TABLE 1
Real-life applications: Average (±STD) of the MSE, estimated over $K = 26$ cutoffs for the NYC Taxi Rides dataset and $k = 10$ folds for the Avocado Prices dataset.

	NYC Taxi Rides	Avocado Prices
OLS	2.00e-2±1.40e-2	73.1e-2±41.3e-2
Sep. Ridge	4.59e-3±5.34e-3	71.0e-2±38.7e-2
Sep. Lasso	5.75e-3±7.12e-3	72.0e-2±36.0e-2
Ridge	4.59e-3±5.34e-3	1.5e-2±39.8e-2
Group Lasso	5.68e-3±7.72e-3	66.7e-2±29.9e-2
MRCE	4.61e-3±5.12e-3	63.4e-2±29.0e-2
MrRCE (ours)	**3.85e-3±4.57e-3**	**53.9e-2±22.6e-2**

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 1. 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 $\hat{\rho} = 0.689$.

7. 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 approach exploits similarities among random coefficients and accounts for correlated errors. We have proposed an efficient EM-based algorithm for computing MrRCE. 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.

In our presentation, we limited the correlation matrix between coefficients for the same variable across tasks to be an equicorrelation matrix, while the correlation between coefficients for the same task is zero. We view these as natural assumptions in the spirit of other grouped modeling methods like Multivariate Group Lasso [30]. However, other, less restrictive assumptions can also be integrated into our framework and estimation approach, and require only minor changes in the M-step of our algorithm. For example, in some cases, it may be reasonable to assume that the q tasks are divided into two groups with an equal correlation between coefficients within-group, and a different (or zero) correlation between groups. Detailed examination of the potential effect of adding parameters to the model in such a manner is left for future research.

In some applications, it may be appropriate to also include fixed effects in the spirit of Eq. (1.2). The fixed effects may not be shared across the different tasks, or be subject to the regularization that the random effects view offers. This can also be naturally accommodated within our approach and algorithm.

Appendix A: Additional experiments

We further evaluate MrRCE on the synthetic data of Section 5. First, we compare MrRCE to the competing methods while varying the sample size n, and the dimension of predictors p. Next, we evaluate the estimation performance of MrRCE and provide the estimates for ρ and σ. Finally, we evaluate MrRCE on a model setting in which the distributional assumption of MrRCE for Γ is violated.

A.1. Varying the number of samples

We first examine the effect of increasing the number of training samples compared to $n = 50$ as in the main text. We generate the data according to the dense AR setting. Table 2 shows the result of varying the sample size from $n = 100$ to $n = 200$. Our method significantly outperforms all competitors for all sample sizes.

TABLE 2
Varying sample size: Average model error (±STD) over 200 repetitions.

Sample Size	100			150			200		
ρ	0	.4	.8	0	.4	.8	0	.4	.8
OLS	1.25±.3	1.27±.3	1.29±.3	.76±.2	.77±.2	.77±.2	.55±.1	.55±.1	.57±.1
Ridge	1.18±.3	1.21±.3	1.25±.3	.74±.2	.75±.2	.74±.2	.54±.1	.54±.1	.56±.1
Group Lasso	1.21±.3	1.24±.3	1.26±.3	.75±.2	.76±.2	.75±.2	.55±.1	.55±.1	.56±.1
MRCE	1.18±.3	1.22±.3	1.25±.3	.75±.2	.76±.2	.75±.2	.56±.1	.56±.1	.58±.1
MrRCE (ours)	**1.12±.3**	**1.13±.2**	**1.02±.2**	**.74±.2**	**.74±.2**	**.69±.1**	**.54±.1**	**.54±.1**	**.53±.1**

A.2. Varying the number of predictors

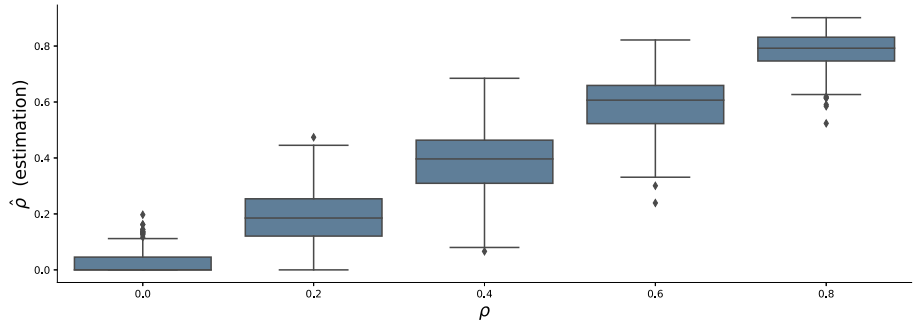
Next, we examine the effect of increasing the number of predictors from $p = 20$ as in the main text to $p = 50$. We set the number of observations to $n = 50$ and generate the data according to the dense AR setting. The results are presented in Table 3. We omit the results for OLS for $p = n = 50$ as it performs poorly. In addition, we omit the MRCE method since it was computationally intractable for these settings.

TABLE 3
Varying predictor dimension: Average model error (±STD) over 200 repetitions.

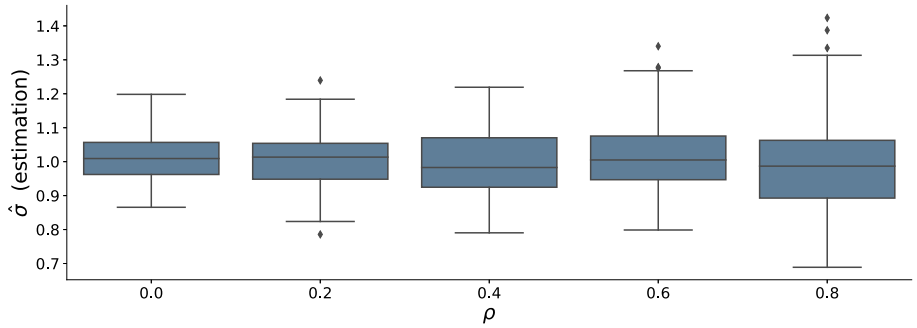
Predictor dim.	30			40			50		
ρ	0	.4	.8	0	.4	.8	0	.4	.8
OLS	7.8±2.4	7.8±2.3	8.1±2.6	23.3±11.1	22.4±9.7	23.3±10.2	–	–	–
Ridge	6.0±1.3	6.0±1.3	6.2±1.7	11.6±2.8	11.3±2.8	11.4±3.7	21.1±5.8	21.3±6.0	20.8±6.0
Group Lasso	6.5±1.4	6.3±1.3	6.3±1.8	12.8±2.8	12.3±3.1	12.1±3.8	23.2±4.8	23.8±6.4	24.6±7.5
MrRCE (ours)	**5.2±1.0**	**4.7±.8**	**3.7±.7**	**9.9±2.0**	**9.2±2.0**	**7.6±1.7**	**19.1±3.2**	**19.3±4.5**	**16.5±4.7**

A.3. Estimating ρ and σ

We now present the estimates for the covariance components ρ and σ. We use the dense AR setting with $n = 100$, $p = 20$ and $q = 5$. Recall that we evaluate each experimental setting using 200 random replications. We set the variance component to $\sigma^2 = 1$ for all experiments, and vary the intra-group correlation coefficient $\rho = 0, .2, .4, .6, .8$. Figure 7 shows a boxplot of the estimates for both parameters against the value of ρ.



(a)



(b)

FIG 7. *Estimating the covariance components: (a) Estimates for ρ; (b) Estimates for σ. True parameter values are marked in red.*

A.4. Violating the distributional assumption

The MrRCE method assumes Γ follows a specific matrix-variate normal distribution, and in Section 5 we generate realization of Γ from that distribution. Here we evaluate MrRCE in setting in which $\gamma_{i,j} \sim \text{Unif}(-1, 1)$, where $\Gamma = (\gamma_{i,j})_{i,j}$. We measure the performance under three setting for the error covariance matrix, as described in Section 5: (i) Fractional Gaussian Noise — FGN; (ii) AR dense; and (iii) Equicorrelation — Equi. We use $n = 50$, $p = 20$ and $q = 5$.

TABLE 4
Uniform distribution for Γ: Average model error (±STD) over 200 repetitions.

	FGN	AR dense	Equi
OLS	3.46±1.30	3.30±1.35	3.58±1.21
Ridge	2.92±0.85	2.77±0.81	2.96±0.82
Group Lasso	2.91±0.78	2.59±0.78	2.89±0.71
MRCE	2.83±0.98	2.51±0.97	3.11±0.94
MrRCE (ours)	**1.80±0.46**	**1.52±0.44**	**2.14±0.44**

Appendix B: Convergence of the MrRCE algorithm

The MrRCE algorithm is a variant of the EM-algorithm for penalized likelihood [15]. Here we discuss the convergence of the EM sequence. We let S denote the parameter space and ℓ_{pen} denote the penalized negative log-likelihood for the full data,

$$\ell_{pen} = \ell\left(Y, \Gamma; \Theta\right) + \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}|.$$

The EM sequence $\{\ell_{pen}(\Theta_t)\}_t := \{\ell_{pen,t}\}_t$ is monotonic non-increasing [15] and thus converge to some ℓ^*_{pen} provided that ℓ_{pen} is bounded bellow on $S_{\Theta_0} := \{\Theta \mid \ell_{pen}(\Theta) \leq \ell_{pen}(\Theta_0)\}$ for finite start value $\ell_{pen}(\Theta_0)$.

Recall that $\ell_{pen} - \lambda_\omega \sum_{j \neq j'} |\omega_{jj'}|$ can be decomposed into two terms of the form

$$f(B) = \mathrm{tr}[BA^TA] - \log|B|,$$

where B is a positive semi-definite matrix. We wish to show that f is bounded from below under mild assumptions: First, note that the trace term is non-negative since

$$\mathrm{tr}[BA^TA] = \mathrm{tr}[(AC)^TAC]$$

where $C = B^{1/2}$. Second, recall that $B = M^{-1}$ is the precision matrix for the covariance matrix Σ or $\sigma^2 D$. Let the eigenvalues of the covariance matrix be $\lambda_1 \geq \ldots \geq \lambda_q$. We have that

$$-\log|B| = \sum_i \log(\lambda_i) \geq q \log(\lambda_q).$$

Thus, assuming λ_q is bounded away from zero, we have that f is bounded from below. Concretely, for $\sigma^2 D$ the assumption implies there exists some $\delta > 0$ such that $\rho \leq 1 - \delta$ and $\sigma^2 \geq \delta$.

Acknowledgement

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

References

[1] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* 327–351. MR0042664

[2] Andrew, H. C. and Neil, S. (1993). Structural time series models. In *Econometrics. Handbook of Statistics* **11** 261–302. Elsevier. MR1247247

[3] 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** 3–54. MR1436554

[4] Brown, P. J. and Zidek, J. V. (1980). Adaptive multivariate ridge regression. *The Annals of Statistics* 64–74. MR0557554

[5] Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics* 1282–1309. MR2816355

[6] Chatfield, C., Zidek, J. and Lindsey, J. (2010). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC. MR0604358

[7] 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** 1533–1545. MR3036414

[8] Chen, L. and Huang, J. Z. (2016). Sparse reduced-rank regression with covariance estimation. *Statistics and Computing* **26** 461–470. MR3439385

[9] dÁspremont, A., Banerjee, O. and El Ghaoui, L. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516. MR2417243

[10] Dawid, P. A. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

[11] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–22. MR0501537

[12] Deshpande, S. K., Rockova, V. and George, E. I. (2017). Simultaneous Variable and Covariance Selection with the Multivariate Spike-and-Slab Lasso. *arXiv preprint arXiv:1708.08911*. MR4045858

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

[14] Furlotte, N. A. and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics* **200** 59–68.

[15] Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* **52** 443–452. MR1086796

[16] Gupta, A. K. and Nagar, D. K. (2018). *Matrix Variate Distributions*. Chapman and Hall/CRC. MR1738933

[17] Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 423–447.

[18] Henderson, C. R. (1984). Applications of linear models in animal breeding: University of Guelph.
[19] Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15** 192–218.
[20] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
[21] Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K. and Sustik, M. A. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems* 2330–2338.
[22] Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* **5** 248–264. MR0373179
[23] Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42** 348.
[24] Kim, S. and Xing, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics* **6** 1095–1117. MR3012522
[25] Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q. and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44** 1066.
[26] Kruuk, L. E. (2004). Estimating genetic parameters in natural populations using the 'animal model'. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **359** 873–890.
[27] 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. MR2944419
[28] Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240
[29] Obozinski, G. R., Wainwright, M. J. and Jordan, M. I. (2009). High-dimensional support union recovery in multivariate regression. In *Advances in Neural Information Processing Systems* 1217–1224. MR2797839
[30] Obozinski, G., Wainwright, M. J., Jordan, M. I. et al. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics* **39** 1–47. MR2797839
[31] 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** 53. MR2758084
[32] Rothman, A. J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19** 947–962. MR2791263

[33] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515. MR2417391

[34] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* **21** 124–127. MR0035118

[35] Sohn, K.-A. and Kim, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Artificial Intelligence and Statistics* 1081–1089.

[36] Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician* **72** 37–45. MR3790566

[37] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288. MR1379242

[38] Turlach, B. A., Venables, W. N. and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics* **47** 349–363. MR2164706

[39] Vattikuti, S., Guo, J. and Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genetics* **8** e1002637.

[40] Velu, R. and Reinsel, G. C. (2013). *Multivariate Reduced-Rank Regression: Theory and Applications* **136**. Springer Science & Business Media. MR1719704

[41] Wilms, I. and Croux, C. (2018). An algorithm for the multivariate group lasso with covariance estimation. *Journal of Applied Statistics* **45** 668–681. MR3750468

[42] 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** 615–636. MR2749910

[43] Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report* **42** 336. MR0038136

[44] Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5** 2630. MR2907129

[45] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

[46] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[47] 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** 329–346. MR2323756

[48] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11** 407.