entropy

MDPI

Article

Multi-Platform Multivariate Regression with Group Sparsity for High-Dimensional Data Integration

Shanshan Qin [1], Guanlin Zhang [2], Xin Gao [2] and Yuehua Wu [2,*]

[1] School of Statistics, Tianjin University of Finance and Economics, Tianjin 300222, China; qinss@tjufe.edu.cn
[2] Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada; xingao@yorku.ca (X.G.)
* Correspondence: wuyh@yorku.ca

Abstract

High-dimensional regression with multivariate responses poses significant challenges when data are collected across multiple platforms, each with potentially correlated outcomes. In this paper, we introduce a multi-platform multivariate high-dimensional linear regression (MM-HLR) model for simultaneously modeling within-platform correlation and cross-platform information fusion. Our approach incorporates a mixture of Lasso and group Lasso penalties to promote both individual predictor sparsity and cross-platform group sparsity, thereby enhancing interpretability and estimation stability. We develop an efficient computational algorithm based on iteratively reweighted least squares and block coordinate descent to solve the resulting regularized optimization problem. We establish theoretical guarantees for our estimator, including oracle bounds on prediction error, estimation accuracy, and support recovery under mild conditions. Our simulation studies confirm the method's strong empirical performance, demonstrating low bias, small variance, and robustness across various dimensions. The analysis of real financial data further validates the performance gains achieved by incorporating multivariate responses and integrating data across multiple platforms.

Keywords: multi-platform; multivariate regression; group sparsity; high-dimensional; data integration

1. Introduction

As data complexity continues to expand in the big-data era, modern datasets increasingly display heterogeneity, high dimensionality, and multiple sources. A common example arises when experiments with the same scientific objective are conducted across different platforms or environments [1–4]. Modeling each dataset individually may result in failing to capture their intrinsic connections, which stem from their shared goal. This motivates the development of methods capable of integrating multi-platform data for unified, simultaneous analysis.

Data integration provides a timely solution by leveraging multiple data sources to enable more robust and efficient statistical inference than relying on any single source alone [5]. In a systematic review, Ref. [5] surveyed integration methods for combining probability samples with non-probability samples and big data sources (see also the references therein). Within the regression framework, a series of studies have advanced data integration methodologies. For instance, Ref. [1] introduced a pseudolikelihood information criterion for high-dimensional multi-experiment data with mixed response

Check for updates

Academic Editor: Leandro Pardo

Received: 22 December 2025
Revised: 15 January 2026
Accepted: 19 January 2026
Published: 23 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

types and varying predictor measurements, establishing selection consistency even with unbounded model size and demonstrating through simulations that data integration substantially outperforms single-source analysis. Building on this work, Ref. [6] implemented the `FusionLearn` R package, which provides a fusion learning algorithm for cross-platform data analysis. Ref. [7] extended the framework to multi-platform data with sub-Gaussian or sub-exponential errors, developing a consistent model selection criterion based on composite likelihood and Bayesian posterior probabilities to recover the union support of predictors under diverging model dimensions. Meanwhile, Ref. [4] addressed multi-task feature learning with mixed continuous and discrete responses using a mixed $\ell_{2,1}$-regularized composite quasi-likelihood function. In a related vein, Ref. [8] proposed a quantile regression approach for high-dimensional multi-source data exhibiting heterogeneity and heavy-tailed error distributions, providing both theoretical guarantees and practical advantages in model recovery.

However, much of the existing literature focuses primarily on univariate response modeling within a single laboratory or experiment. In many modern applications, response variables are multivariate and correlated, yet share the same set of high-dimensional covariates or predictors [9]. For instance, in the UK Biobank population-based cohort study, researchers face large-scale, ultrahigh-dimensional features alongside a wide array of correlated phenotypic outcomes, including lifestyle measures, biomarkers, and disease diagnoses [10]. Such data structures have motivated extensive methodological developments in high-dimensional multivariate regression. Seminal work includes remMap by [11], designed for multivariate response regression in high-dimension–low-sample-size settings. Ref. [9] proposed a blockwise descent algorithm for group-penalized multi-response regression. Further advancing the field, Ref. [12] introduced a regularization method that enhances variable selection by efficiently eliminating irrelevant blocks of regression coefficients. More recently, Ref. [10] developed a scalable sparse reduced-rank regression method for high-dimensional multi-task learning with correlated outcomes. Ref. [13] proposed a novel framework suited for settings with large numbers of responses, response categories, and predictors. Additionally, several quantile regression approaches have been proposed to handle multiple responses under non-Gaussian or heterogeneous error settings, like [14,15], among others.

Nevertheless, relatively few works have simultaneously addressed the complexities of multivariate responses across multiple data sources. We thus propose a multi-platform multivariate high-dimensional linear regression (MM-HLR) method, designed to jointly model within-platform correlation while promoting cross-platform group sparsity. Structured or group sparsity has been well studied recently. Zhang et al. [16] presented a probabilistic framework for subset selection under partition constraints, which aligns with the group-sparsity and cross-platform fusion objectives of our study. In addition, Li et al. [12] proposed a penalized multivariate multiple linear regression model with an arbitrary group structure for the regression coefficient matrix. Besides, the mixed $\ell_{2,1}$ penalty has similarly been employed for multitask feature learning and selection [4,17]. Kawano et al. [18] proposed a multivariate regression modeling in integrative analysis that performed group selection by group lasso estimation. In this article, we extended the mixture of ℓ_1 and ℓ_2 penalty formulation from [12] to simultaneously enforce individual sparsity and cross-platform group sparsity in our multi-platform setting. An efficient optimization framework, combining block coordinate descent with a proxy approximation strategy, is introduced to solve the resulting non-convex regularized problem. Theoretically, we establish non-asymptotic bounds on the prediction error, estimation accuracy, and support recovery of the MM-HLR estimator. Empirically, we evaluate the proposed method under varying sample sizes and predictor dimensionalities, benchmarking it against two alternative approaches:

(i) FusionLearn [6], which is designed for multi-platform integration but treats responses as univariate, thereby ignoring within-platform correlation structures, and (ii) multivariate Group Lasso [19], which is designed for multivariate regression within a single platform and thus overlooks cross-platform grouping. Comprehensive simulation studies are conducted to demonstrate the method's performance across diverse data-generating scenarios. Finally, we validate the proposed method through a real-world financial data analysis to assess the gains of jointly modeling multivariate responses and integrating data across multiple platforms.

The remainder of this article is organized as follows: Section 2 introduces the model framework and parameter estimation. Section 3 establishes the theoretical guarantees of the proposed estimator. Section 4 presents simulation studies to evaluate the performance of the method. We provide a real data analysis in Section 5. Finally, Section 6 concludes this article.

2. Methodology

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices. $\boldsymbol{A} \preceq \boldsymbol{B}$ denotes that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm of $\boldsymbol{A}$. In addition, $\|\boldsymbol{A}\|_1$ and $||\boldsymbol{A}||_\infty$ denote the sum and the maximum of the absolute values of all entries of $\boldsymbol{A}$. Let $\boldsymbol{a}$ be a vector. We denote its ℓ_2 and ℓ_1 norms by $||\boldsymbol{a}||_2$ and $||\boldsymbol{a}||_1$, respectively.

2.1. Model Setup

Assume we have n objects and aim to model the linear relationship between m responses and p predictors for each object. To enhance measurement accuracy, each object is sent to k different platforms, each of which returns two matrices, a response matrix $\boldsymbol{Y}_i \in \mathbb{R}^{n\times m}$ and a design matrix $\boldsymbol{X}_i \in \mathbb{R}^{n\times p}$, $i = 1, \ldots, k$. Due to variations in equipment and measurement protocols across platforms, the scale and continuity of the results may differ [1,6]. Our goal is to identify the common set of influential predictors that affect the responses across all platforms. Moreover, because multivariate responses are measured on the same objects, there exist unknown correlations between responses within the platform and across platforms. Key terminology is summarized in Table 1.

Table 1. Data structure of response matrix $\boldsymbol{Y}_i \in \mathbb{R}^{n\times m}$, design matrix $\boldsymbol{X}_i \in \mathbb{R}^{n\times p}$ and parameter matrix $\boldsymbol{B}_i \in \mathbb{R}^{p\times m}$ for platform i, where y_{ijs} and X_{ijr} denote the j-th observed value of response s and predictor r on platform i, respectively. The corresponding regression coefficient of predictor r on response s is β_{irs}, $i = 1, \ldots, k; j = 1, \ldots, n; s = 1, \ldots, m; r = 1, \ldots, p$.

	Platform 1	$\cdots$	**Platform k**
Parameter matrix	$\boldsymbol{B}_1 = \begin{pmatrix} \theta_{111} & \ldots & \theta_{11m} \\ \vdots & \ddots & \vdots \\ \theta_{1p1} & \ldots & \theta_{1pm} \end{pmatrix}$	$\cdots$	$\boldsymbol{B}_k = \begin{pmatrix} \theta_{k11} & \ldots & \theta_{k1m} \\ \vdots & \ddots & \vdots \\ \theta_{kp1} & \ldots & \theta_{kpm} \end{pmatrix}$
Response matrix	$\boldsymbol{Y}_1 = \begin{pmatrix} Y_{111} & \ldots & Y_{11m} \\ \vdots & \ddots & \vdots \\ Y_{1n1} & \ldots & Y_{1nm} \end{pmatrix}$	$\cdots$	$\boldsymbol{Y}_k = \begin{pmatrix} Y_{k11} & \ldots & Y_{k1m} \\ \vdots & \ddots & \vdots \\ Y_{kn1} & \ldots & Y_{knm} \end{pmatrix}$
Design matrix	$\boldsymbol{X}_1 = \begin{pmatrix} X_{111} & \ldots & X_{11p} \\ \vdots & \ddots & \vdots \\ X_{1n1} & \ldots & X_{1np} \end{pmatrix}$	$\cdots$	$\boldsymbol{X}_k = \begin{pmatrix} X_{k11} & \ldots & X_{k1p} \\ \vdots & \ddots & \vdots \\ X_{kn1} & \ldots & X_{knp} \end{pmatrix}$

From the linearity relationship between responses and predictors, we have the following:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{B}_i + \boldsymbol{E}_i, i = 1, \ldots, k \tag{1}$$

where $\boldsymbol{Y}_i \in \mathbb{R}^{n\times m}$, $\boldsymbol{X}_i \in \mathbb{R}^{n\times p}$, $\boldsymbol{B}_i \in \mathbb{R}^{p\times m}$ are respectively the response matrix, design matrix and regression parameter matrix from ith platform. $\boldsymbol{E}_i \in \mathbb{R}^{n\times m}$ is the error matrix, where the rows of $\boldsymbol{E}_i$ are identically independent from $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\Sigma}_i \in \mathbb{R}^{m\times m}$ being the within-platform error covariance matrix.

To simplify the modeling complexity, we treat each platform independently within the likelihood function. Accordingly, we adopt a marginal composite likelihood approach to integrate information across platforms [7,20,21], i.e., the following is calculated:

$$L(\boldsymbol{B}) = \prod_{i=1}^{k} f(\boldsymbol{Y}_i; \boldsymbol{B}_i), \tag{2}$$

where $\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_k) \in \mathbb{R}^{p\times km}$, and $f(\boldsymbol{Y}_i; \boldsymbol{B}_i)$ denotes the multivariate normal density function of $\boldsymbol{Y}_i$, that is,

$$f(\boldsymbol{Y}_i; \boldsymbol{B}_i) \propto |\boldsymbol{\Sigma}_i|^{-\frac{n}{2}} \exp\{-\frac{1}{2}\mathrm{tr}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)^\top]\}. \tag{3}$$

The log-likelihood function can be given by the following (excluding the constant):

$$\ell(\boldsymbol{B}) = -\frac{n}{2}\sum_{i=1}^{k}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)^\top]. \tag{4}$$

Following [1], the objective of this study is to recover a union subset of predictors, where each selected predictor is associated with at least one outcome across the platforms. To identify these commonly influential predictors, we enforce the constraint that the set of nonzero rows of $\boldsymbol{B}_k$ is identical for all platforms $k = 1, \ldots, K$. Stacking the r-th row vector of all $\boldsymbol{B}_i$s, we obtain the regression coefficient matrix for predictor r across all m responses and k platforms, i.e., written as follows:

$$\boldsymbol{\Theta}^{(r)} = (\boldsymbol{\theta}_1^{(r)}, \ldots, \boldsymbol{\theta}_k^{(r)})^\top \in \mathbb{R}^{k\times m},$$

where $\boldsymbol{\theta}_i^{(r)} \in \mathbb{R}^{m\times 1}$ represents the r-th row of $\boldsymbol{B}_i$. We assume that the sparsity among k platforms is the same with respect to the predictors, that is, $\boldsymbol{\theta}_1^{(r)}, \ldots, \boldsymbol{\theta}_k^{(r)}$ are either all zero or all nonzero for each $r = 1, \ldots, p$. To select the influential predictors among all platforms, we add the group lasso penalty into the log-likelihood function, i.e., written as follows:

$$\tilde{Q}(\boldsymbol{B}) = \ell(\boldsymbol{B}) - n\sum_{j=1}^{p}\Omega_{\lambda_n}(||\boldsymbol{\Theta}^{(r)}||_F), \tag{5}$$

where $||\cdot||_F$ is the Frobenius norm (square root of sum of squares of all entries). Assuming the penalty function is the mixture of L_1 and L_2 norms, then a sparse estimate of $\boldsymbol{B}$ (denoted as $\hat{\boldsymbol{B}}$) can be obtained by solving the minimization problem as follows:

$$\min_{\boldsymbol{B}} \tilde{Q}(\boldsymbol{B}) = \min_{\boldsymbol{B}} -\ell(\boldsymbol{B}) + n\sum_{i,r,s}\lambda_{irs}|\theta_{irs}| + n\sum_{r=1}^{p}\lambda_g||\boldsymbol{\Theta}^{(r)}||_F. \tag{6}$$

Denote $p_1(\boldsymbol{B}) = n\sum_{i,r,s}\lambda_{irs}|\theta_{irs}|$ and $p_2(\boldsymbol{B}) = n\sum_{r=1}^{p}\lambda_g||\boldsymbol{\Theta}^{(r)}||_F$. The first penalty $p_1(\boldsymbol{B})$ encourages sparsity (individual coefficients to zero), and the second penalty $p_2(\boldsymbol{B})$ encourages group sparsity for each predictor to be zero across all platforms.

2.2. Parameter Estimation

We use an iteratively reweighted least squares approach and block coordinate descent per predictor $r, r = 1, \ldots, p$. For simplicity, we assume that the element-wise regularization parameters for the Lasso penalty are identical, $\lambda_{irs} = \lambda_e$. In practice, the covariance matrix $\boldsymbol{\Sigma}_i$ among responses is typically unknown for each platform $i = 1, \ldots, k$. A simple initial estimate, such as the empirical covariance of the model residuals, can be employed. The following algorithm is then implemented conditional on these estimated $\boldsymbol{\Sigma}_i$s. Denote $f(\boldsymbol{B}_i) = \frac{1}{2}\,\mathrm{tr}\left((\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)^\top\right)$. Then the optimization problem (6) becomes the following:

$$\min_{\boldsymbol{B}} Q(\boldsymbol{B}) = \min_{\boldsymbol{B}} \sum_{i=1}^{k} f(\boldsymbol{B}_i) + n \sum_{i,r,s} \lambda_e |\theta_{irs}| + n \sum_{r=1}^{p} \lambda_g ||\boldsymbol{\Theta}^{(r)}||_F. \tag{7}$$

Rewrite $\boldsymbol{X}_i\boldsymbol{B}_i = \sum_{r=1}^{p} \boldsymbol{X}_i^{(r)}\boldsymbol{\theta}_i^{(r)\top}$. For predictor r, by holding $\boldsymbol{\Theta}^{(j)}$ fixed for $j \neq r$, we define residuals without predictor r for platform i, $\boldsymbol{R}_i^{(r)} = \boldsymbol{Y}_i - \sum_{j\neq r} \boldsymbol{X}_i^{(j)}\boldsymbol{\theta}_i^{(j)\top} \in \mathbb{R}^{n\times m}$. Then the first gradient of f, with respect to $\boldsymbol{\theta}_i^{(r)}$, is $\nabla_{\boldsymbol{\theta}_i^{(r)}} f = -\boldsymbol{\Sigma}_i^{-1}\boldsymbol{R}_i^{(r)\top}\boldsymbol{X}_i^{(r)} \in \mathbb{R}^{m\times 1}$, where $\boldsymbol{X}_i^{(j)} \in \mathbb{R}^{n\times 1}$ is the column j of $\boldsymbol{X}_i, i = 1, \ldots, k$. Denote the following:

$$\boldsymbol{S}_i^{(r)} = -\nabla_{\boldsymbol{\theta}_i^{(r)}} f = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{R}_i^{(r)\top}\boldsymbol{X}_i^{(r)} \in \mathbb{R}^{m\times 1}. \tag{8}$$

By coordinate descent algorithm, we estimate $\boldsymbol{\theta}_i^{(r)}$ by fixing the rest. Let $\boldsymbol{\Theta}^{(j)}$ be fixed for $j \neq r$, then problem (7) can be transformed to be a minimization sub-problem for predictor r $(r = 1, \ldots, p)$ across k platforms, i.e., the following is calculated:

$$\min_{\boldsymbol{\theta}_i^{(r)}} Q_r(\boldsymbol{\theta}_i^{(r)}) = \sum_{i=1}^{k} \left(\frac{\|\boldsymbol{X}_i^{(r)}\|_2^2}{2}\,\boldsymbol{\theta}_i^{(r)\top}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\theta}_i^{(r)} - \boldsymbol{S}_i^{(r)\top}\boldsymbol{\theta}_i^{(r)} + \lambda_e n \|\boldsymbol{\theta}_i^{(r)}\|_1 \right) + \lambda_g n \|\boldsymbol{\Theta}^{(r)}\|_F. \tag{9}$$

Next, we will show how to solve the sub-problem (9). Define the element-wise soft-thresholding $\boldsymbol{S}_{i,soft}^{(r)} = \mathcal{S}_{n\lambda_e}(\boldsymbol{S}_i^{(r)})$, where $\mathcal{S}_\tau(x) = \mathrm{sign}(x)\max(|x| - \tau, 0)$, a soft-thresholding operator. In light of the block soft-thresholding solution of the group Lasso in [22], the predictor r is set to be inactive, $\boldsymbol{\theta}_i^{(r)} = \boldsymbol{0}$ across all platforms if the following is met:

$$T_r = \sum_{i=1}^{k} \|\boldsymbol{S}_{i,soft}^{(r)}\|_2^2 \leq n^2\lambda_g^2. \tag{10}$$

Otherwise, the predictor r is selected to be active, $\|\boldsymbol{\Theta}^{(r)}\|_F \neq 0$. If so, we can update $\boldsymbol{\theta}_i^{(r)}$ via a block soft-thresholding operator, as shown in [23]. Given our objective to select a union subset of active predictors across all platforms, an active predictor r may be relevant to platform i or other platforms. We therefore consider the following two scenarios for updating $\boldsymbol{\theta}_i^{(r)}$.

Assume that the predictor r in all platforms except platform i is inactive, that is, $\boldsymbol{\theta}_{i'}^{(r)} = \boldsymbol{0}$, for $i' \neq i$. Then the group Lasso norm reduces to the following: $\|\boldsymbol{\Theta}^{(r)}\|_F = \|\boldsymbol{\theta}_i^{(r)}\|_2$, and the sub-problem (9) simplifies to become

$$\min_{\boldsymbol{\theta}_i^{(r)}\in\mathbb{R}^m} \frac{\|\boldsymbol{X}_i^{(r)}\|_2^2}{2}\,\boldsymbol{\theta}_i^{(r)\top}\,\boldsymbol{\Sigma}_i^{-1}\,\boldsymbol{\theta}_i^{(r)} - \boldsymbol{S}_i^{(r)\top}\boldsymbol{\theta}_i^{(r)} + n\lambda_e\|\boldsymbol{\theta}_i^{(r)}\|_1 + n\lambda_g\|\boldsymbol{\theta}_i^{(r)}\|_2, \tag{11}$$

where the objective function is a sum of a smooth quadratic term and a composite non-smooth penalty $\lambda_e\|\cdot\|_1 + \lambda_g\|\cdot\|_2$. Its proximal mapping consists of sequential soft-

thresholding followed by block shrinkage. Then the minimizer of the row-wise objective with respect to $\boldsymbol{\theta}_i^{(r)}$ is given by the following:

$$\boldsymbol{\theta}_i^{(r)} = \boldsymbol{\Sigma}_i \frac{\mathcal{S}_{n(\lambda_e+\lambda_g)}(\boldsymbol{S}_i^{(r)})}{\|\boldsymbol{X}_i^{(r)}\|_2^2}. \tag{12}$$

Assume that the predictor r is active across all platforms, we conduct the closed-form updating for the nonzero group r as follows. A solution to problem (9) is a minimizer of the following:

$$\widetilde{Q}_r(\boldsymbol{\theta}_i^{(r)}) = \sum_{i=1}^{k} \Big(\frac{\|\boldsymbol{X}_i^{(r)}\|_2^2}{2} \boldsymbol{\theta}_i^{(r)\top} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\theta}_i^{(r)} - \boldsymbol{S}_{i,soft}^{(r)\top} \boldsymbol{\theta}_i^{(r)} \Big) + \lambda_g n \|\boldsymbol{\Theta}^{(r)}\|_F. \tag{13}$$

Take the first derivative of $\widetilde{Q}_r$ with respect to $\boldsymbol{\theta}_i^{(r)}$ and let them be zero, i.e., as follows:

$$\|\boldsymbol{X}_i^{(r)}\|_2^2 \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\theta}_i^{(r)} - \boldsymbol{S}_{i,soft}^{(r)} + \lambda_g n \frac{\boldsymbol{\theta}_i^{(r)}}{\|\boldsymbol{\Theta}^{(r)}\|_F} = \boldsymbol{0}. \tag{14}$$

We rewrite the above equation to group the terms linear in $\boldsymbol{\theta}_i^{(r)}$, written as follows:

$$\left(\|\boldsymbol{X}_i^{(r)}\|_2^2 \boldsymbol{\Sigma}_i^{-1} + \frac{\lambda_g n}{\|\boldsymbol{\Theta}^{(r)}\|_F} \boldsymbol{I}_m \right) \boldsymbol{\theta}_i^{(r)} = \boldsymbol{S}_{i,soft}^{(r)}. \tag{15}$$

which yields a closed-form of $\boldsymbol{\theta}_i^{(r)}$, i.e.,

$$\boldsymbol{\theta}_i^{(r)} = \left(\|\boldsymbol{X}_i^{(r)}\|_2^2 \boldsymbol{\Sigma}_i^{-1} + \frac{\lambda_g n}{\|\boldsymbol{\Theta}^{(r)}\|_F} \boldsymbol{I}_m \right)^{-1} \boldsymbol{S}_{i,soft}^{(r)}. \tag{16}$$

2.3. Algorithm

We estimate $\{\boldsymbol{B}_i\}_{i=1}^k$ via a nested alternating minimization procedure that consists of an outer loop updating covariance matrices and an inner loop performing row-wise block coordinate descent on regression coefficients with sparse group penalties. For the $(t+1)$-th outer loop, $t = 0, \ldots, T_{\max} - 1$, we perform a fixed number of inner loops to update the regression coefficients $\boldsymbol{B}_i^{(t,\ell)}, \ell = 0, 1, \ldots, \ell_{\max}$ while keeping the covariance matrices $\boldsymbol{\Sigma}_i^{(t)}$ fixed. When the inner loop arrives at the maximum iteration, we update $\boldsymbol{B}_i^{(t+1)}$, followed by the updating of $\boldsymbol{\Sigma}_i^{(t+1)}$ and recording the residuals. Convergence of the algorithm is then assessed at the outer-iteration level based on changes in $\boldsymbol{B}_i^{(t+1)}$ and $\boldsymbol{\Sigma}_i^{(t+1)}$. Each outer iteration, therefore, consists of a full inner loop for coefficient updates, followed by updating the regression covariance and convergence evaluation. The details are as follows:

Step 1: Initialization $\boldsymbol{B}_i^{(0)}$ and $\boldsymbol{\Sigma}_i^{(0)}$ for each platform $i, i = 1, \ldots, k$.

Step 2: (Outer loop) Update $\boldsymbol{B}_i^{(t+1)}$ across all platforms given $\boldsymbol{\Sigma}_i^{(t)}, \boldsymbol{B}_i^{(t)}$ at t-th iteration.

Substep 2.1: (Inner loop) Initialization: for predictor r $(r = 1, \ldots, p)$, $\boldsymbol{\theta}_i^{(r)(t,0)} = \boldsymbol{\theta}_i^{(r)(t)}, i = 1, \ldots, k$, where $\boldsymbol{\theta}_i^{(r)(t)}$ is the r-th row of $\boldsymbol{B}_i^{(t)}$.

Substep 2.2: Update $\boldsymbol{\theta}_i^{(r)(t,\ell+1)}$ across all platforms given $\boldsymbol{\theta}_i^{(r)(t,\ell)}$ and $\boldsymbol{\Sigma}_i^{(t)}$. Perform row-wise group screening. Let $\boldsymbol{\theta}_i^{(r)(t,\ell+1)} = \boldsymbol{0}$ if $T_r^{(t,\ell)} \leq n^2 \lambda_g^2$. Otherwise, predictor

r is added to the active row set. For each active predictor r, we conduct row-wise coefficient updating, i.e., calculated as follows:

$$\boldsymbol{\theta}_i^{(r)(t,\ell+1/2)} = \boldsymbol{\Sigma}_i^{(t)} \frac{\mathcal{S}_{n(\lambda_e+\lambda_g)}(\boldsymbol{S}_i^{(r)(t,\ell)})}{\|\boldsymbol{X}_i^{(r)}\|_2^2} \tag{17}$$

if $\|\boldsymbol{\theta}_{i'}^{(r)(t,\ell)}\|_1 = 0, i' \neq i$. If not, then the following is calculated:

$$\boldsymbol{\theta}_i^{(r)(t,\ell+1/2)} = \left(\|\boldsymbol{X}_i^{(r)}\|_2^2 \boldsymbol{\Sigma}_i^{-1(t)} + \frac{\lambda_g n}{\|\boldsymbol{\Theta}^{(r)(t,\ell)}\|_F} \boldsymbol{I}_m \right)^{-1} \boldsymbol{S}_{i,soft}^{(r)(t,\ell)}. \tag{18}$$

Substep 2.3: For each platform i, $\boldsymbol{B}_i^{(t,\ell+1)} = (\hat{\boldsymbol{B}}_{i,\widehat{\mathcal{A}}}, \boldsymbol{0}_{\widehat{\mathcal{A}}^c})$, where $\widehat{\mathcal{A}} = \{r : \|\boldsymbol{\theta}_i^{(r)(t,\ell+1/2)}\|_1 \neq 0\}$, and the following:

$$\widehat{\boldsymbol{B}}_{i,\widehat{\mathcal{A}}} = \left(\boldsymbol{X}_{i,\widehat{\mathcal{A}}}^\top \boldsymbol{X}_{i,\widehat{\mathcal{A}}} + \varepsilon \boldsymbol{I}_{|\widehat{\mathcal{A}}|} \right)^{-1} \boldsymbol{X}_{i,\widehat{\mathcal{A}}}^\top \boldsymbol{Y}_i. \tag{19}$$

Substep 2.4: Repeat substeps 2.2–2.3 until it reaches the maximum inner iterations $\ell_{\max}$.

Step 3: Update $\boldsymbol{\Sigma}_i^{(t+1)}$ via $\boldsymbol{\Sigma}_i^{(t+1)} = \frac{1}{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{B}_i^{(t+1)})^\top (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{B}_i^{(t+1)})$.

Step 4: Repeat steps 2–3 until it meets the termination conditions, i.e., the following:

$$R_B^{(t+1)} = \frac{1}{k} \sum_{i=1}^{k} \left\| \boldsymbol{B}_i^{(t+1)} - \boldsymbol{B}_i^{(t)} \right\|_F^2 \leq \delta, \tag{20}$$

$$R_\Sigma^{(t+1)} = \frac{1}{k} \sum_{i=1}^{k} \left\| \boldsymbol{\Sigma}_i^{(t+1)} - \boldsymbol{\Sigma}_i^{(t)} \right\|_F^2 \leq \delta, \tag{21}$$

where δ is a small number, say, 1×10^{-4}.

We remark that, for initialization, we set $\boldsymbol{B}_i^{(0)} = \boldsymbol{0.1}_{p\times m}$, $\boldsymbol{\Sigma}_i^{(0)} = \boldsymbol{I}_m$, for $i = 1, \ldots, k$. Tuning parameters are typically selected by evaluating a candidate set of values via cross-validation or an information criterion like the adjusted BIC. To enhance computational efficiency and avoid a costly two-dimensional grid search, we adopt a practical simplification by fixing the relationship between λ_g and λ_e, for instance, by setting $\lambda_g = \lambda_e$ or $\lambda_g = \lambda_e/2$. Consequently, we select the optimal λ_e via two-fold cross-validation from a common empirical candidate set for λ_e, such as $\{0.1, 0.2, \ldots, 2.0\}$ with an increment of 0.1 or smaller. The range and granularity of the candidate set can be adapted to the specific application.

3. Theoretical Property

The theoretical property is in light of [12], where their theories are built under the framework of multivariate linear regression. Herein, we extend them to multiple platforms. For platform i $(i = 1, \ldots, k)$, denote $\boldsymbol{J}_1(\boldsymbol{B}_i) = \{rs : |\theta_{irs}| \neq 0\}$ the index set of nonzero elements in $\boldsymbol{B}_i$, and $\boldsymbol{J}_2(\boldsymbol{B}) = \{r : ||\boldsymbol{\Theta}^{(r)}||_2 \neq 0, r = 1, \ldots, p\}$ the index set of nonzero rows in $\boldsymbol{B}$. We assume that the sparsity of all $\boldsymbol{B}_i, i = 1, \ldots, k$ is the same. Define $M_1(\boldsymbol{B}) = |\boldsymbol{J}_1(\boldsymbol{B}_i)|$ and $M_2(\boldsymbol{B}) = |\boldsymbol{J}_2(\boldsymbol{B})|$. For any matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times m}$ and $\boldsymbol{J}_1 \subseteq \{rs : 1 \leq r \leq p, 1 \leq s \leq m\}$, denote $\boldsymbol{\Delta}_{\boldsymbol{J}_1}$ as the projection of $\boldsymbol{\Delta}$ on the index set $\boldsymbol{J}_1$, which is a matrix with the same elements of $\boldsymbol{\Delta}$ on coordinates $\boldsymbol{J}_1$ and zeros on the complementary coordinates $\boldsymbol{J}_1^c$. Denote $\boldsymbol{\Delta}_{\boldsymbol{J}_2} = \{\boldsymbol{\Theta}_\Delta^{(r)}, r \in \boldsymbol{J}_2\}$, where $\boldsymbol{\Theta}_\Delta^{(r)}$ is a $p \times m$ matrix with the rth row same as $\boldsymbol{\Delta}$ and zeros on the other rows.

Let $\boldsymbol{B}_i^*$ be the true regression coefficient matrices in model (1), and $\hat{\boldsymbol{B}}_i$ be its estimated counterpart for $i = 1, \ldots, k$. Assume each column of the random error matrix follows a

multivariate normal distribution $N_m(\mathbf{0}, \mathbf{\Sigma}_i^*)$. Denote $u = M_1(\boldsymbol{B}^*)$ and $v = M_2(\boldsymbol{B}^*)$. The theoretical framework is built on a given covariance matrix for each platform i. In practice, $\mathbf{\Sigma}_i^*$ is usually unknown. So we instead use an estimator $\hat{\mathbf{\Sigma}}_i$. Before proceeding, we impose mild conditions on both design matrices $\boldsymbol{X}_i$ and covariance matrices $\mathbf{\Sigma}_i^*$ for all platforms.

Assumption 1. *Assume that the columns of $\boldsymbol{X}_i$ are centered such that the diagonal elements of the matrix $\boldsymbol{X}_i^\top \boldsymbol{X}_i/n$ are equal to 1 for all $i = 1, \ldots, k$. Let $\psi_{\max} > 0$ be the largest eigenvalue of $\boldsymbol{X}_i^\top \boldsymbol{X}_i/n$ for all $i = 1, \ldots, k$.*

Assumption 2. *For a given $\mathbf{\Sigma}_i \in \mathbb{R}^{m\times m}$, assume that there exists a $\lambda^* \in (0,1)$, such that $-\lambda^* \boldsymbol{I} \preceq \mathbf{\Sigma}_i - \mathbf{\Sigma}_i^* \preceq \lambda^* \boldsymbol{I}$ for all $i = 1, \ldots, k$. There exist positive constants $0 < C_1 < C_2$ such that $C_1 \geq \lambda^*$, and the eigenvalues of $\mathbf{\Sigma}_i^*$ are less than C_2 and larger than C_1 for all $i = 1, \ldots, k$.*

Assumption 3. *Let $\boldsymbol{J}_1 \subseteq \{rs : 1 \leq r \leq p, 1 \leq s \leq m\}$ and $\boldsymbol{J}_2 \subseteq \{1, \ldots, p\}$ be any index sets that satisfy $|\boldsymbol{J}_1| \leq u, |\boldsymbol{J}_2| \leq v$. Let $\tilde{\rho} = \{\rho_r, r = 1, \ldots, p\}$ be a set of positive numbers. For any nontrivial matrices $\mathbf{\Delta}_i \in \mathbb{R}^{p\times m}, i = 1, \ldots, k$, if it satisfies the following:*

$$\sum_{i=1}^{k} |\mathbf{\Delta}_{i,\boldsymbol{J}_1^c}|_1 + 2\sum_{r\in \boldsymbol{J}_2^c} \rho_r ||\mathbf{\Delta}^{(r)}||_2 \leq 3\sum_{i=1}^{k} |\mathbf{\Delta}_{i,\boldsymbol{J}_1}|_1 + 2\sum_{r\in \boldsymbol{J}_2} \rho_r ||\mathbf{\Delta}^{(r)}||_2, \tag{22}$$

where $\mathbf{\Delta}^{(r)}$ is built by the rth rows of $\mathbf{\Delta}_i$ for all $i = 1, \ldots, k$, then the following minimums exist and are positive, i.e., the following:

$$\kappa_1(u, v, \tilde{\rho}) = \min_{\boldsymbol{J}_1, \boldsymbol{J}_2, \mathbf{\Delta}\neq 0} \frac{\sum_{i=1}^{k} \left(\mathrm{tr}[\boldsymbol{X}_i \mathbf{\Delta}_i \mathbf{\Sigma}_i^{-1} (\boldsymbol{X}_i \mathbf{\Delta}_i)^\top]\right)^{1/2}}{n^{1/2} \sum_{i=1}^{k} ||\mathbf{\Delta}_{i,\boldsymbol{J}_1}||_F} > 0, \tag{23}$$

$$\kappa_2(u, v, \tilde{\rho}) = \min_{\boldsymbol{J}_1, \boldsymbol{J}_2, \mathbf{\Delta}\neq 0} \frac{\sum_{i=1}^{k} \left(\mathrm{tr}[\boldsymbol{X}_i \mathbf{\Delta}_i \mathbf{\Sigma}_i^{-1} (\boldsymbol{X}_i \mathbf{\Delta}_i)^\top]\right)^{1/2}}{n^{1/2} ||\mathbf{\Delta}_{\boldsymbol{J}_2}||_F} > 0. \tag{24}$$

We remark that these constants involved in assumptions represent standard regularity conditions in statistical theory. While their specific values are derived from the proofs and are not tuned in practice, their existence is required to establish the desired theoretical guarantees. Furthermore, we define $s_{\tilde{\rho}^*} = \sum_{r\in \boldsymbol{J}_2(\boldsymbol{B}^*)} \rho_r^2$, $\sigma^2 = \frac{C_1}{(C_1-\lambda^*)^2}$, $\lambda_g = \rho_r \lambda$ for $r = 1, \ldots, p, \rho = \min_r\{1, \rho_r\}$, and $\lambda = 2\sigma A\{\log(pmk)/n\}^{1/2}$ for some constant $A > \sqrt{2}$. Rewrite $\kappa_1 = \kappa_1(u, v, \tilde{\rho})$ and $\kappa_2 = \kappa_2(u, v, \tilde{\rho})$.

Theorem 1. *Under Assumptions 1–3, the following oracle bounds for the prediction error, the estimation error, and the order of sparsity hold, with probability at least $1 - (pmk)^{1-A^2/2}$, i.e., written as follows:*

$$\frac{1}{n}\sum_{i=1}^{k} \mathrm{tr}[(\boldsymbol{X}_i \boldsymbol{B}_i^* - \boldsymbol{X}_i \hat{\boldsymbol{B}}_i) \mathbf{\Sigma}_i^{-1} (\boldsymbol{X}_i \boldsymbol{B}_i^* - \boldsymbol{X}_i \hat{\boldsymbol{B}}_i)^\top] \leq 16\lambda^2 k \Big(\frac{u^{1/2}}{\kappa_1} + \frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)^2, \tag{25}$$

$$\sum_{i=1}^{k} ||\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*||_1 \leq \frac{24\lambda k^{1/2}}{1+\rho} \Big(\frac{u^{1/2}}{\kappa_1} + \frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)^2, \tag{26}$$

$$M_1(\hat{\boldsymbol{B}}_i) \leq \frac{64\psi_{\max}}{(C_1 - \lambda^*)} \Big(\frac{u^{1/2}}{\kappa_1} + \frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)^2, i = 1, \ldots, k. \tag{27}$$

The proof of Theorem 1 is given in the Appendix A. The three inequalities (25)–(27) guarantee the performance of the estimator $\hat{\boldsymbol{B}}$ in terms of prediction accuracy, estimation error, and sparsity. The first inequality (25) bounds the weighted mean squared prediction error across all k platforms. The bound scales with $\log(pmk)/n$ and depends on the sparsity and grouping structures. The second inequality (26) controls the cumulative ℓ_1-norm of

$\hat{\boldsymbol{B}}_i$ from $\boldsymbol{B}_i^*$ across all platforms. This bound grows slowly with $\sqrt{\log(pqm)/n}$, typical of high-dimensional settings, and depends similarly on the same sparsity and grouping structures. The third inequality (27) bounds $M_1(\hat{\boldsymbol{B}}_i)$, a measure of the model complexity or effective sparsity of the estimator. This ensures that the estimated model is not overly dense and depends on the maximum eigenvalue $\psi_{\max}$.

4. Simulations

To test our method, we simulated data from model (1) with $k = 2$, $m = 3$, and $(n, p) = \{(100, 100), (100, 200), (100, 500), (200, 200), (200, 500), (200, 1000)\}$. For each platform i, the design matrix $\boldsymbol{X}_i$ was generated with entries drawn independently from a p-variate normal distribution with mean zero and unit standard deviation. To introduce high correlation within each row of the error matrix $\boldsymbol{E}_i$, we set the diagonal elements of its covariance matrix $\boldsymbol{\Sigma}_i$ to 1 and the off-diagonal elements to 0.8. The coefficient matrix $\boldsymbol{B}_i$ was constructed with 10 active rows. Entries in these rows were sampled uniformly, either from $U(-1.5, -0.5)$ or $U(0.5, 1.5)$.

The evaluation metrics encompass three primary aspects: parameter estimation accuracy across all platforms, prediction accuracy, and feature selection accuracy. For each platform $i = 1, \ldots, k$, we measure the Frobenius norm of the deviation between the true regression coefficient matrix $\boldsymbol{B}_i^*$ and its estimate $\hat{\boldsymbol{B}}_i$, written as follows:

$$\boldsymbol{B}_{F,i} = \|\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i\|_F,$$

the deviation between the true covariance matrix $\boldsymbol{\Sigma}_i^*$ and its estimate $\hat{\boldsymbol{\Sigma}}_i$,

$$\boldsymbol{\Sigma}_{F,i} = \|\boldsymbol{\Sigma}_i^* - \hat{\boldsymbol{\Sigma}}_i\|_F,$$

and the root mean squared errors (RMSE), i.e.,

$$\text{RMSE}_i = \|\boldsymbol{X}_i\boldsymbol{B}_i^* - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i\|_F/\sqrt{n},$$

where $\boldsymbol{X}_i$ is out of samples for all platforms. The overall estimation accuracy of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ can be summarized by the average across all k platforms, written as follows:

$$\overline{\boldsymbol{B}_F} = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{B}_{F,i},\ \overline{\boldsymbol{\Sigma}_F} = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\Sigma}_{F,i},\ \text{and}\ \overline{\text{RMSE}} = \frac{1}{k}\sum_{i=1}^{k}\text{RMSE}_i.$$

Meanwhile, feature selection performance is evaluated using sensitivity, the proportion of truly active predictors that are correctly identified, written as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

and specificity, the proportion of truly inactive predictors that are correctly excluded,

$$\text{Specificity} = \frac{TN}{TN + FP},$$

where TP and TN represent the number of predictors correctly identified as active (non-zero true coefficients with non-zero estimates) and inactive (zero true coefficients with zero estimates), respectively. Conversely, FN and FP denote the number of predictors incorrectly identified as active (zero true coefficients with non-zero estimates) and inactive (non-zero true coefficients with zero estimates), respectively. These metrics are computed for each platform, and the results are aggregated across all k platforms to provide a comprehensive assessment of feature selection performance.

We compare our proposed method with two alternative approaches, FusionLearn and multivariate Group Lasso. The FusionLearn method, implemented in the R package FusionLearn [6], is designed for multi-platform data integration but treats responses as univariate. To adapt it to our multivariate setting, we apply it separately to each of the m responses across platforms, which ignores the covariance structure among responses. Conversely, the multivariate Group Lasso method, available via the glmnet package [19], is designed for multivariate regression within a single platform. We therefore apply it independently to the data from each platform, which does not leverage information shared across platforms.

The simulation results are given in Tables 2 and 3. As is seen in Table 2, the comparison of MM-HLR, FusionLearn, and GroupLasso reveals distinct performance in parameter estimation and prediction. In terms of parameter estimation of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, the proposed MM-HLR consistently provides estimates closest to the true ones, reflected by their smallest measures. In contrast, both FusionLearn and GroupLasso systematically overestimate the bias of parameters. The estimates of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ via GroupLasso are often excessively high, especially for smaller n and larger p. In terms of prediction performance, MM-HLR demonstrates a strong advantage, achieving the lowest RMSE in every (n, p) setting. Its RMSE values range from 0.38 to 0.58, which are much lower than those of the other two methods. These consistent superiorities confirm MM-HLR's excellent parameter estimation and predictive capability.

Table 2. Comparison of averaged parameter estimation accuracy ($\boldsymbol{B}_F, \boldsymbol{\Sigma}_F$) and prediction accuracy (RMSE) across different methods under various (n, p) configurations (standard errors in parentheses).

(n, p)	Methods	$B_{F,1}$	$B_{F,2}$	$\overline{B_F}$	$\mathbf{RMSE}_1$	$\mathbf{RMSE}_2$	$\overline{\mathbf{RMSE}}$	$\Sigma_{F,1}$	$\Sigma_{F,2}$	$\overline{\Sigma_F}$
$(100, 100)$	MM-HLR	0.5653 (0.13)	0.5830 (0.13)	0.5742 (0.09)	0.5598 (0.13)	0.5719 (0.11)	0.5659 (0.09)	0.4486 (0.24)	0.4328 (0.18)	0.4407 (0.13)
	FusionLearn	0.7115 (0.11)	0.7485 (0.13)	0.7300 (0.08)	0.7065 (0.11)	0.7444 (0.12)	0.7254 (0.08)	0.5298 (0.24)	0.5157 (0.19)	0.5228 (0.15)
	GroupLasso	1.0402 (0.13)	1.0629 (0.14)	1.0516 (0.08)	1.0522 (0.14)	1.0782 (0.14)	1.0652 (0.09)	0.9649 (0.41)	1.0862 (0.36)	1.0256 (0.30)
$(100, 200)$	MM-HLR	0.5821 (0.13)	0.5808 (0.13)	0.5814 (0.09)	0.5770 (0.13)	0.5764 (0.13)	0.5767 (0.10)	0.4282 (0.24)	0.4152 (0.24)	0.4217 (0.17)
	FusionLearn	0.7494 (0.11)	0.7583 (0.15)	0.7539 (0.08)	0.7437 (0.11)	0.7463 (0.14)	0.7450 (0.09)	0.6083 (0.23)	0.5350 (0.21)	0.5717 (0.15)
	GroupLasso	1.1937 (0.16)	1.1821 (0.15)	1.1879 (0.11)	1.1925 (0.17)	1.1590 (0.16)	1.1757 (0.12)	1.2531 (0.48)	1.3192 (0.46)	1.2862 (0.33)
$(100, 500)$	MM-HLR	0.5594 (0.11)	0.5944 (0.12)	0.5769 (0.09)	0.5566 (0.12)	0.5864 (0.12)	0.5715 (0.09)	0.3718 (0.20)	0.4083 (0.23)	0.3901 (0.15)
	FusionLearn	0.7646 (0.11)	0.7613 (0.09)	0.7629 (0.08)	0.7643 (0.12)	0.7593 (0.10)	0.7618 (0.08)	0.8426 (0.20)	0.7951 (0.19)	0.8189 (0.12)
	GroupLasso	1.2484 (0.15)	1.2021 (0.14)	1.2252 (0.11)	1.2593 (0.17)	1.1910 (0.15)	1.2252 (0.11)	1.6113 (0.49)	1.6055 (0.47)	1.6084 (0.38)
$(200, 200)$	MM-HLR	0.3998 (0.08)	0.3861 (0.08)	0.3930 (0.06)	0.3982 (0.08)	0.3864 (0.08)	0.3923 (0.06)	0.2941 (0.15)	0.2429 (0.12)	0.2685 (0.10)
	FusionLearn	0.5087 (0.09)	0.5190 (0.07)	0.5138 (0.06)	0.5062 (0.09)	0.5181 (0.07)	0.5122 (0.06)	0.3608 (0.17)	0.3300 (0.14)	0.3454 (0.11)
	GroupLasso	0.7809 (0.08)	0.7788 (0.08)	0.7798 (0.05)	0.7768 (0.09)	0.7678 (0.07)	0.7723 (0.05)	0.6529 (0.31)	0.6957 (0.28)	0.6743 (0.20)

Table 2. *Cont.*

(n, p)	Methods	$B_{F,1}$	$B_{F,2}$	$\overline{B_F}$	RMSE_1	RMSE_2	$\overline{\text{RMSE}}$	$\Sigma_{F,1}$	$\Sigma_{F,2}$	$\overline{\Sigma_F}$
$(200, 500)$	MM-HLR	0.3845 (0.08)	0.3871 (0.07)	0.3858 (0.05)	0.3828 (0.08)	0.3857 (0.08)	0.3843 (0.05)	0.2695 (0.13)	0.2657 (0.17)	0.2676 (0.10)
	FusionLearn	0.5516 (0.07)	0.5473 (0.09)	0.5495 (0.05)	0.5511 (0.07)	0.5428 (0.09)	0.5469 (0.05)	0.5278 (0.17)	0.5431 (0.17)	0.5355 (0.11)
	GroupLasso	0.8847 (0.10)	0.8629 (0.10)	0.8738 (0.06)	0.8806 (0.10)	0.8539 (0.10)	0.8673 (0.07)	1.0466 (0.41)	1.0352 (0.42)	1.0409 (0.32)
$(200, 1000)$	MM-HLR	0.4068 (0.09)	0.4044 (0.09)	0.4056 (0.06)	0.4084 (0.09)	0.3985 (0.08)	0.4035 (0.06)	0.2506 (0.14)	0.2579 (0.15)	0.2542 (0.10)
	FusionLearn	0.5893 (0.09)	0.5762 (0.10)	0.5827 (0.07)	0.5893 (0.09)	0.5707 (0.09)	0.5800 (0.07)	0.6546 (0.16)	0.6876 (0.15)	0.6711 (0.10)
	GroupLasso	0.9324 (0.10)	0.9642 (0.11)	0.9474 (0.08)	0.9336 (0.11)	0.9592 (0.11)	0.9464 (0.09)	1.1231 (0.44)	1.0740 (0.48)	1.0985 (0.35)

Table 3 presents the comparison of averaged feature selection accuracy (sensitivity and specificity) across different methods under various (n, p) configurations. All three methods achieve perfect sensitivity (1.0) across all pairs of (n, p), indicating their superior ability to select true predictors. However, the specificity values vary significantly among the three methods. MM-HLR achieves near-perfect or perfect specificity (0.9996 to 1.0) in all scenarios. GroupLasso has the lowest specificity overall, ranging from 0.7362 to 0.9396, indicating its weaker ability to exclude irrelevant predictors than the other two methods, while the ability of FusionLearn to exclude irrelevant predictors is in the middle, exhibiting the specificity values from 0.9166 to 0.9524.

Table 3. Comparison of averaged feature selection accuracy (sensitivity and specificity) across different methods under various (n, p) configurations (standard errors in parentheses).

(n, p)	Sensitivity			Specifitity		
	MM-HLR	**FusionLearn**	**GroupLasso**	**MM-HLR**	**FusionLearn**	**GroupLasso**
(100, 100)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.9996(0.00)	0.9166(0.02)	0.7362(0.08)
(100, 200)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.9996(0.00)	0.9257(0.02)	0.8152(0.06)
(100, 500)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.9496(0.01)	0.8944(0.04)
(200, 200)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.9326(0.02)	0.8392(0.05)
(200, 500)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.99996(0.00)	0.9394(0.01)	0.8962(0.04)
(200, 1000)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	1.0000(0.00)	0.9524(0.01)	0.9396(0.02)

5. Real Data Analysis

Zhang et al. [7] integrated three financial indices (S&P 500, Dow Jones, and VIX) as distinct platforms. Similarly, we consider a two-platform analysis ($K = 2$) of U.S. stock market data. Our first platform comprises the S&P 500 index and its volatility index (VIX), while the second consists of the NASDAQ-100 index and its volatility index (VXN). Thus, each platform provides a two-dimensional response vector ($m = 2$). We consider stocks that are actively traded in either the S&P 500 or NASDAQ-100 indices as predictors. The analysis uses weekly data from October 2022 to September 2025. To achieve approximately independent samples, log returns are calculated for all series. After data preprocessing and the removal of missing values, the final dataset contains 156 weekly observations and 510 predictors, which are common to both platforms. The correlations between two responses within each platform are -0.79 for Platform 1 (S&P 500 and VIX) and -0.66 for Platform 2 (NASDAQ-100 and VXN). The dataset is randomly partitioned into a training set of 100 samples ($n = 100$) and a test set comprising the remaining observations ($n_{test} = 56$).

Our objective is to identify a common set of predictors that are relevant to at least one of the responses in either platform.

Table 4 presents the prediction errors on the test samples for Platform i $(PE_i, i = 1, 2)$ and their average (PE), where $PE_i = \|\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i\|_F/\sqrt{n_{\text{test}}}$. We also report the number of selected predictors (*support*) to quantify the sparsity of the estimated model. The proposed MM-HLR method demonstrates superior overall performance compared to the benchmark methods, FusionLearn and GroupLasso, in terms of both prediction accuracy and model sparsity. Specifically, MM-HLR achieves the lowest total prediction error ($PE = 0.0828$), which is much lower than that of FusionLearn ($PE = 0.8903$) and approximately 20% lower than that of GroupLasso ($PE = 0.1067$). This advantage is consistent across both individual platforms. Furthermore, MM-HLR yields the most parsimonious model, selecting only 6 predictors. In contrast, FusionLearn and GroupLasso select 115 and 75 predictors, respectively, suggesting a higher risk of overfitting. This effective balance between prediction accuracy and model simplicity underscores the capability of the MM-HLR method to leverage both multivariate dependence and multi-platform structures within high-dimensional regression.

Table 4. The platform-specific prediction errors $(PE_i, i = 1, 2)$, the average prediction error (PE), and the number of selected predictors (*support*).

Method	$\boldsymbol{PE_1}$	$\boldsymbol{PE_2}$	***PE***	***Support***
MM-HLR	0.0993	0.0663	0.0828	6
FusionLearn	0.8895	0.8912	0.8903	91
GroupLasso	0.1257	0.0877	0.1067	75

6. Conclusions and Discussion

In this paper, we introduce the multi-platform high-dimensional multivariate linear regression model, designed to simultaneously model correlated multivariate responses across multiple platforms. The proposed framework explicitly accounts for within-platform response correlation while incorporating a group Lasso penalty to fuse information across platforms, thereby promoting both individual sparsity and structured grouping of predictors. The optimization algorithm, combining iteratively reweighted least squares with block coordinate descent, is developed to solve the resulting regularization problem efficiently. Theoretical guarantees are established for the estimator $\hat{\boldsymbol{B}}$, covering prediction accuracy, estimation error bounds, and support recovery (sparsity) under regularity conditions. Simulation studies under various scenarios demonstrate the outperformance of MM-HLR in parameter estimation, prediction, and variable selection, showing minimal bias, low variance, and strong robustness across all tested conditions. The superior performance of our approach is further validated through an analysis of real financial data, which confirms its effectiveness in leveraging both multivariate dependence and multi-platform structures within high-dimensional regression.

In this article, we impose a shared sparsity structure across platforms. Extending the framework to handle scenarios with only partial overlap would be a valuable direction for future work. Accordingly, we note that the method's performance may degrade if the common sparsity pattern is violated and that formally relaxing this assumption constitutes an important avenue for future research.

Author Contributions: Conceptualization, X.G., Y.W. and S.Q.; methodology, X.G., Y.W. and S.Q.; software, S.Q. and G.Z.; validation, S.Q.; formal analysis, S.Q.; investigation, S.Q.; resources, S.Q. and G.Z.; data curation, S.Q.; writing—original draft preparation, S.Q. and G.Z.; writing—review and editing, Y.W. and S.Q.; supervision, X.G., Y.W. and S.Q.; project administration, S.Q. All authors have read and agreed to the published version of the manuscript.

Funding: Shanshan Qin is supported by the National Natural Science Foundation of China (No.12201454) and the China Scholarship Council (No.202408120083). Xin Gao is supported by the Natural Science and Engineering Research Council of Canada (No. RGPIN-2024-06202) and Yuehua Wu is supported by the Natural Science and Engineering Research Council of Canada (No.RGPIN-2023-05655).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Theoretical Proofs

We provide the proof of Theorem 1. Before proceeding, we present Lemmas A1 and A2, which are required in the proof of Theorem 1.

Lemma A1. *Define a random matrix* $\boldsymbol{W}_i = \boldsymbol{E}_i\boldsymbol{\Sigma}_i^{-1} \in \mathbb{R}^{n\times m}$ *and a random variable* $V_{irs} = \boldsymbol{X}_i^{(r)\top}\boldsymbol{W}_i^{(s)}$, *where* $\boldsymbol{W}_i^{(s)} \in \mathbb{R}^n$ *is the sth column of* $\boldsymbol{W}_i$, $s = 1,\ldots,m; i = 1,\ldots,k$. *Define an event* $\mathcal{A}$ *and its complementary event* $\mathcal{A}^c$, *respectively, as follows:*

$$\mathcal{A} = \{||\boldsymbol{X}_i^\top \boldsymbol{W}_i||_\infty \le \frac{n\lambda}{2}, \textit{for all } i = 1,\ldots,k\},$$

$$\mathcal{A}^c = \{\textit{at least one } |V_{irs}| \ge \frac{n\lambda}{2}, i = 1,\ldots,k, r = 1,\ldots,p, s = 1,\ldots,m\}.$$

By Assumptions 1 and 2, the following holds:

$$Pr\{\mathcal{A}^c\} \le (pmk)^{1-A^2/2} \to 0,$$

for some constant $A > \sqrt{2}$.

Proof of Lemma A1. Since the rows of $\boldsymbol{E}_i$ follows $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i^*)$ independently and identically (iid), then the rows of $\boldsymbol{W}_i$ are iid with $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^*\boldsymbol{\Sigma}_i^{-1})$ random vectors. For the covariance matrix $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^*\boldsymbol{\Sigma}_i^{-1}$, by Assumption 2, we have the following:

$$-\lambda^*\boldsymbol{I} \preceq \boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_i^* \preceq \lambda^*\boldsymbol{I}, \boldsymbol{\Sigma}^* \preceq C_2\boldsymbol{I} \Rightarrow (C_1 - \lambda^*)\boldsymbol{I} \preceq \boldsymbol{\Sigma}_i \preceq (\lambda^* + C_2)\boldsymbol{I}$$
$$\Rightarrow \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^*\boldsymbol{\Sigma}_i^{-1} \preceq \boldsymbol{\Sigma}_i^{-1} + \lambda^*\boldsymbol{\Sigma}_i^{-2} \preceq \Big(\frac{1}{C_1 - \lambda^*} + \frac{\lambda^*}{(C_1 - \lambda^*)^2}\Big)\boldsymbol{I} = \frac{C_1}{(C_1 - \lambda^*)^2}\boldsymbol{I},$$

which means that the diagonal elements of $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^*\boldsymbol{\Sigma}_i^{-1}$ are less than $\sigma^2 \triangleq \frac{C_1}{(C_1-\lambda^*)^2}$. Since $\boldsymbol{W}_i^{(s)} \sim N_n(\boldsymbol{0}, \sigma_{is}^2\boldsymbol{I})$, where σ_{is}^2 is the sth diagonal element of $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^*\boldsymbol{\Sigma}_i^{-1}$. By Assumption 1, $Var(V_{irs}) = \boldsymbol{X}_i^{(r)\top} cov(\boldsymbol{W}_i^{(s)})\boldsymbol{X}_i^{(r)} = n\sigma_{is}^2 \le n\sigma$. Therefore $(n\sigma_{is}^2)^{-\frac{1}{2}}V_{irs} \triangleq Z$ are standard normal random variables for all i, r, s. We have the following:

$$\begin{aligned} Pr\{\mathcal{A}^c\} &\leq \sum_{r=1}^{p}\sum_{s=1}^{m}\sum_{i=1}^{k} Pr\{|V_{irs}| \geq \frac{n\lambda}{2}\} \\ &= p\sum_{s=1}^{m}\sum_{i=1}^{k} Pr\{|V_{irs}|/\sqrt{n}\sigma_{is} \geq \frac{\sqrt{n}\lambda}{2\sigma_{is}}\} \\ &\leq p\sum_{s=1}^{m}\sum_{i=1}^{k} Pr\{|Z| \geq \frac{\sqrt{n}\lambda}{2\sigma}\} \\ &\leq pmk\exp\big(-\frac{\lambda^2 n}{8\sigma^2}\big) = (pmk)^{1-A^2/2}, \end{aligned}$$

where $\lambda = 2\sigma A\{\log(pmk)/n\}^{1/2}$. □

Lemma A2. *Under Assumptions 1–3, for any $\boldsymbol{B}_i \in \mathbb{R}^{p\times m}, i = 1,\ldots,k$, with probability at least $1-(pmk)^{1-A^2/2}$, the following holds:*

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))^\top] + \lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i||_1 + 2\sum_{r=1}^{p}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)}||_F \\ \leq &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))^\top] + 4\lambda\sum_{i=1}^{k}\sum_{rs\in\boldsymbol{J}_1(\boldsymbol{B}_i)}|\hat{\theta}_{irs}-\theta_{irs}| + 4\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B})}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)}||_F \end{aligned} \tag{A1}$$

and

$$M_1(\hat{\boldsymbol{B}}_i) \leq \frac{4}{\lambda^2 n^2}||\boldsymbol{X}_i^\top\boldsymbol{X}_i(\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*)\boldsymbol{\Sigma}_i^{-1}||_F^2. \tag{A2}$$

Proof of Lemma A2. For any $\boldsymbol{B}_i \in \mathbb{R}^{p\times m}, i = 1,\ldots,k$, $Q(\hat{\boldsymbol{B}}) \leq Q(\boldsymbol{B})$. Plugging $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{B}_i^* + \boldsymbol{E}_i, i = 1,\ldots,k$ into the above inequality, results in the following:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))^\top] \\ \leq &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))^\top] \\ &+\frac{2}{n}\sum_{i=1}^{k}||\boldsymbol{X}_i^\top\boldsymbol{E}_i\boldsymbol{\Sigma}_i^{-1}||_\infty||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i||_1 \\ &+2\lambda\sum_{i,r,s}(|\theta_{irs}|-|\hat{\theta}_{irs}|) + 2\sum_{r=1}^{p}\lambda_g(||\boldsymbol{\Theta}^{(r)}||_F - ||\hat{\boldsymbol{\Theta}}^{(r)\top}||_F). \end{aligned} \tag{A3}$$

By Lemma A1, it holds that on event $\mathcal{A}$, the following is calculated:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))^\top] + \lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i||_1 + 2\sum_{r=1}^{m}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)}||_F \\ \leq &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))^\top] + 2\lambda\sum_{i=1}^{k}(||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i||_1 + ||\boldsymbol{B}_i||_1 - ||\hat{\boldsymbol{B}}_i||_1) \\ &+2\sum_{r=1}^{p}\lambda_g(||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)}||_F + ||\boldsymbol{\Theta}^{(r)}||_F - ||\hat{\boldsymbol{\Theta}}^{(r)}||_F) \\ \leq &\frac{1}{n}\sum_{i=1}^{k}\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\boldsymbol{B}_i))^\top] + 4\lambda\sum_{i=1}^{k}\sum_{rs\in\boldsymbol{J}_1(\boldsymbol{B})}|\hat{\theta}_{irs}-\theta_{irs}| \\ &+4\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B})}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)}||_F. \end{aligned} \tag{A4}$$

This completes the proof of the first inequality (A1) in Lemma A2.

To prove the second inequality (A2) in Lemma A2, we use the KKT conditions. For $\forall\, \hat{\theta}_{irs} \neq 0$, the following is written:

$$\boldsymbol{X}_i^{(r)\top}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s} = 2n\lambda sign(\hat{\theta}_{irs}) + \frac{2n\lambda_g\hat{\theta}_{irs}}{||\hat{\boldsymbol{\Theta}}^{(r)}||_F},$$

where $(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}$ is the s column of $\boldsymbol{\Sigma}_i^{-1}$, which implies that

$$\lambda \leq \frac{1}{n}|\boldsymbol{X}_i^{(r)\top}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}|. \tag{A5}$$

On the other hand, we have on event $\mathcal{A}$, written as follows:

$$\begin{aligned}\frac{1}{n}|\boldsymbol{X}_i^{(r)\top}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}| &\leq \frac{1}{n}|\boldsymbol{X}_i^{(r)\top}\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s} + \boldsymbol{X}_i^{(r)\top}\boldsymbol{E}_i(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}| \\ &\leq \frac{1}{n}|\boldsymbol{X}_i^{(r)\top}\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}| + \frac{1}{n}||\boldsymbol{X}_i^\top\boldsymbol{W}_i||_\infty \\ &\leq \frac{1}{n}|\boldsymbol{X}_i^{(r)\top}\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}| + \frac{\lambda}{2}. \end{aligned} \tag{A6}$$

Combining Equations (A5) and (A6) together, we can obtain the following:

$$\frac{\lambda}{2} \leq \frac{1}{n}|\boldsymbol{X}_i^{(r)\top}\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}|.$$

Therefore, for any platform i, the following is written:

$$\begin{aligned} M_1(\hat{\boldsymbol{B}}_i) &\leq \frac{4}{\lambda^2 n^2}\sum_{rs\in\boldsymbol{J}_1(\hat{\boldsymbol{B}}_i)}|\boldsymbol{X}_i^{(r)\top}\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)(\boldsymbol{\Sigma}_i^{-1})_{\cdot s}|^2 \\ &\leq \frac{4}{\lambda^2 n^2}||\boldsymbol{X}_i^\top\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}||_F^2. \end{aligned} \tag{A7}$$

This completes the proof of Lemma A2. □

Now, we detail the proof of Theorem 1.

Proof of Theorem 1. By the first inequality (A1) in Lemma A2 and letting $\boldsymbol{B} = \boldsymbol{B}^*$, we have, on event $\mathcal{A}$, the following:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^k \mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top] \\ &\leq 4\lambda\sum_{i=1}^k\sum_{rs\in\boldsymbol{J}_1(\boldsymbol{B}_i^*)}|\hat{\theta}_{irs} - \theta_{irs}^*| + 4\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B}^*)}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F \\ &\leq 4\lambda u^{\frac{1}{2}}\sum_{i=1}^k||(\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*)_{\boldsymbol{J}_1(\boldsymbol{B}_i^*)}||_F + 4(\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B}^*)}\lambda_g^2)^{\frac{1}{2}}||(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)_{\boldsymbol{J}_2(\boldsymbol{B}^*)}||_F, \end{aligned} \tag{A8}$$

which is derived by Cauchy–Schwarz inequality. On event $\mathcal{A}$, we also have the following:

$$\begin{aligned} &\lambda\sum_{i=1}^k||\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*||_1 + 2\sum_{r=1}^p\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F \\ &\leq 4\lambda\sum_{i=1}^k\sum_{rs\in\boldsymbol{J}_1(\boldsymbol{B}_i^*)}|\hat{\theta}_{irs} - \theta_{irs}^*| + 4\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B}^*)}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F, \end{aligned} \tag{A9}$$

which, together with the fact that

$$||\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*||_1 = \sum_{rs \in \boldsymbol{J}_1(\boldsymbol{B}_i^*)} |\hat{\theta}_{irs} - \theta_{irs}^*| + \sum_{rs \in \boldsymbol{J}_1^c(\boldsymbol{B}_i^*)} |\hat{\theta}_{irs} - \theta_{irs}^*|, \tag{A10}$$

$$\sum_{r=1}^{p} ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F = \sum_{r \in \boldsymbol{J}_2(\boldsymbol{B}^*)} ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F + \sum_{r \in \boldsymbol{J}_2^c(\boldsymbol{B}^*)} ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F, \tag{A11}$$

yield the following inequality,

$$\begin{aligned} &\lambda \sum_{i=1}^{k} \sum_{rs \in \boldsymbol{J}_1^c(\boldsymbol{B}_i^*)} |\hat{\theta}_{irs} - \theta_{irs}^*| + 2 \sum_{r \in \boldsymbol{J}_2^c(\boldsymbol{B}^*)} \lambda_g ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F \\ &\leq 3\lambda \sum_{i=1}^{k} \sum_{rs \in \boldsymbol{J}_1(\boldsymbol{B}_i^*)} |\hat{\theta}_{irs} - \theta_{irs}^*| + 2 \sum_{r \in \boldsymbol{J}_2(\boldsymbol{B}^*)} \lambda_g ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F. \end{aligned} \tag{A12}$$

Thus the inequality (22) in Assumption 3 holds with $\boldsymbol{\Delta}_i = \hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*, i = 1, \ldots, k$ and $\rho_r = \lambda_g / \lambda$. Therefore, we obtain the following:

$$\sum_{i=1}^{k} ||(\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*)_{J_1(\boldsymbol{B}^*)}||_F \leq \frac{\sum_{i=1}^{k} \left(\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top]\right)^{\frac{1}{2}}}{n^{1/2}\kappa_1}, \tag{A13}$$

$$||(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)_{J_2(\boldsymbol{B}^*)}||_F \leq \frac{\sum_{i=1}^{k} \left(\mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top]\right)^{\frac{1}{2}}}{n^{1/2}\kappa_2}. \tag{A14}$$

Plugging Equations (A13) and (A14) into (A8), we have the following:

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{k} \mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top] \\ &\leq \left(\frac{4\lambda u^{\frac{1}{2}}}{n^{\frac{1}{2}}\kappa_1} + \frac{4\left(\sum_{r \in J_2(\boldsymbol{B}^*)} \lambda_g^2\right)^{\frac{1}{2}}}{n^{\frac{1}{2}}\kappa_2}\right) \sum_{i=1}^{k} \left(\mathrm{tr}[\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top]\right)^{\frac{1}{2}} \\ &= \left(\frac{4\lambda u^{\frac{1}{2}}}{n^{\frac{1}{2}}\kappa_1} + \frac{4\lambda s_{\tilde{\rho}^*}^{\frac{1}{2}}}{n^{\frac{1}{2}}\kappa_2}\right) \sum_{i=1}^{k} \left(\mathrm{tr}[\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top]\right)^{\frac{1}{2}}, \end{aligned} \tag{A15}$$

which yields that, by Cauchy–Schwarz inequality,

$$\frac{1}{n} \sum_{i=1}^{k} \mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top] \leq 16\lambda^2 k \left(\frac{u^{\frac{1}{2}}}{\kappa_1} + \frac{s_{\tilde{\rho}^*}^{\frac{1}{2}}}{\kappa_2}\right)^2. \tag{A16}$$

Next, we prove the second term (26) in Theorem 1. Define $||\boldsymbol{B}||_{2,1} = \sum_{i=1}^{k} ||\boldsymbol{B}_i||_1 + \sum_{r=1}^{p} ||\boldsymbol{\Theta}^{(r)}||_F$. Hence, the following is calculated:

$$||\hat{\boldsymbol{B}} - \boldsymbol{B}^*||_{2,1} = \sum_{i=1}^{k} ||\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*||_1 + \sum_{r=1}^{p} ||\hat{\boldsymbol{\Theta}}^{(r)} - \boldsymbol{\Theta}^{(r)*}||_F \leq 2 \sum_{i=1}^{k} ||\hat{\boldsymbol{B}}_i - \boldsymbol{B}_i^*||_1.$$

Then we have the following:

$$\begin{aligned}(\lambda+\lambda\rho_r)||\hat{\boldsymbol{B}}-\boldsymbol{B}^*||_{2,1} &= \lambda||\hat{\boldsymbol{B}}-\boldsymbol{B}^*||_{2,1}+\rho_r\lambda||\hat{\boldsymbol{B}}-\boldsymbol{B}^*||_{2,1}\\ &\le 3\lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*||_1+\rho_r\lambda\sum_{r=1}^{p}||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)*}||_F\\ &\le 3\lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*||_1+\sum_{r=1}^{p}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)*}||_F\\ &\le 3\Big[\lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*||_1+2\sum_{r=1}^{p}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)*}||_F\Big]. \end{aligned} \tag{A17}$$

By (A9), we obtain the following:

$$\begin{aligned}\frac{1+\rho_r}{3}\lambda||\hat{\boldsymbol{B}}-\boldsymbol{B}^*||_{2,1} &\le \lambda\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*||_1+2\sum_{r=1}^{p}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)*}||_F\\ &\le 4\lambda\sum_{i=1}^{k}\sum_{rs\in\boldsymbol{J}_1(\boldsymbol{B}_i^*)}|\hat{\theta}_{irs}-\theta_{irs}^*|+4\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B}^*)}\lambda_g||\hat{\boldsymbol{\Theta}}^{(r)}-\boldsymbol{\Theta}^{(r)*}||_F\\ &\le 4\lambda u^{1/2}\sum_{i=1}^{k}||(\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*)_{\boldsymbol{J}_1(\boldsymbol{B}_i^*)}||_F+4\Big(\sum_{r\in\boldsymbol{J}_2(\boldsymbol{B}^*)}\lambda_g^2\Big)^{1/2}||(\hat{\boldsymbol{B}}-\boldsymbol{B}^*)_{\boldsymbol{J}_2(\boldsymbol{B}^*)}||_F\\ &\le \Big(\frac{4\lambda u^{1/2}}{n^{1/2}\kappa_1}+\frac{4\lambda s_{\tilde{\rho}^*}^{1/2}}{n^{1/2}\kappa_2}\Big)\sum_{i=1}^{k}\Big(\mathrm{tr}[\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i))^\top]\Big)^{1/2}\\ &\le \Big(\frac{4\lambda u^{1/2}}{n^{1/2}\kappa_1}+\frac{4\lambda s_{\tilde{\rho}^*}^{1/2}}{n^{1/2}\kappa_2}\Big)4(kn)^{1/2}\lambda\Big(\frac{u^{1/2}}{\kappa_1}+\frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)\\ &= 16k^{1/2}\lambda^2\Big(\frac{u^{1/2}}{\kappa_1}+\frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)^2. \end{aligned} \tag{A18}$$

Therefore, we obtain the following:

$$\sum_{i=1}^{k}||\hat{\boldsymbol{B}}_i-\boldsymbol{B}_i^*||_1\le\frac{1}{2}||\hat{\boldsymbol{B}}-\boldsymbol{B}^*||_{2,1}\le\frac{24\lambda k^{1/2}}{1+\rho}\Big(\frac{u^{1/2}}{\kappa_1}+\frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2}\Big)^2. \tag{A19}$$

Finally, we prove the third term (27) in Theorem 1. By (A2) in Lemma A2, we can obtain the following:

$$\begin{aligned}M_1(\hat{\boldsymbol{B}}_i) &= \frac{4}{\lambda^2n^2}||\boldsymbol{X}_i^\top\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}||_F^2\\ &= \frac{4}{\lambda^2n^2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)^\top\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}\Big)\\ &\le \frac{4}{\lambda^2n}\psi_{\max}\mathrm{tr}\Big(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)^\top\boldsymbol{X}_i^\top\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}\Big)\\ &= \frac{4}{\lambda^2n}\psi_{\max}\mathrm{tr}\Big(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-2}(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)^\top\boldsymbol{X}_i^\top\Big)\\ &\le \frac{4}{\lambda^2n}\frac{\psi_{\max}}{C_1-\lambda^*}\mathrm{tr}\Big(\boldsymbol{X}_i(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{B}_i^*-\hat{\boldsymbol{B}}_i)^\top\boldsymbol{X}_i^\top\Big). \end{aligned} \tag{A20}$$

Since a common sparsity structure is shared across all platforms, the following is assumed:

$$M_1(\hat{\boldsymbol{B}}_i) \leq \frac{4\psi_{\max}}{\lambda^2 n} \frac{1}{k(C_1 - \lambda^*)} \sum_{i=1}^{k} \mathrm{tr}[(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_i(\boldsymbol{B}_i^* - \hat{\boldsymbol{B}}_i))^\top]$$
$$\leq \frac{64\psi_{\max}}{(C_1 - \lambda^*)} \Big(\frac{u^{1/2}}{\kappa_1} + \frac{s_{\tilde{\rho}^*}^{1/2}}{\kappa_2} \Big)^2. \tag{A21}$$

This completes the proof of Theorem 1. □

References

1. Gao, X.; Carroll, R.J. Data integration with high dimensionality. *Biometrika* **2017**, *104*, 251–272. [CrossRef]
2. Liu, Q.; Xu, Q.; Zheng, V.W.; Xue, H.; Cao, Z.; Yang, Q. Multi-task learning for cross-platform siRNA efficacy prediction: An in-silico study. *BMC Bioinform.* **2010**, *11*, 1–16. [CrossRef]
3. Zhang, K.; Gray, J.W.; Parvin, B. Sparse multitask regression for identifying common mechanism of response to therapeutic targets. *Bioinformatics* **2010**, *26*, i97–i105. [CrossRef]
4. Zhong, Y.; Xu, W.; Gao, X. Heterogeneous multi-task feature learning with mixed $\ell_{2,1}$ regularization. *Mach. Learn.* **2024**, *113*, 891–932. [CrossRef]
5. Yang, S.; Kim, J.K. Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.* **2020**, *3*, 625–650. [CrossRef]
6. Gao, X.; Zhong, Y. FusionLearn: a biomarker selection algorithm on cross-platform data. *Bioinformatics* **2019**, *35*, 4465–4468. [CrossRef] [PubMed]
7. Zhang, G.; Wu, Y.; Gao, X. Bayesian Model Selection via Composite Likelihood for High-dimensional Data Integration. *Can. J. Stat.* **2024**, *52*, 924–938. [CrossRef]
8. Dai, G.; Müller, U.U.; Carroll, R.J. Data integration in high dimension with multiple quantiles. *Stat. Sin.* **2023**, *33*, 169–191. [CrossRef]
9. Simon, N.; Friedman, J.; Hastie, T. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv* **2013**, arXiv:1311.6529. [CrossRef]
10. Qian, J.; Tanigawa, Y.; Li, R.; Tibshirani, R.; Rivas, M.A.; Hastie, T. Large-scale multivariate sparse regression with applications to UK biobank. *Ann. Appl. Stat.* **2022**, *16*, 1891–1918. [CrossRef]
11. Peng, J.; Zhu, J.; Bergamaschi, A.; Han, W.; Noh, D.Y.; Pollack, J.R.; Wang, P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **2010**, *4*, 53–77. [CrossRef]
12. Li, Y.; Nan, B.; Zhu, J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **2015**, *71*, 354–363. [CrossRef]
13. Molstad, A.J.; Zhang, X. Conditional probability tensor decompositions for multivariate categorical response regression. *J. Am. Stat. Assoc.* **2025**, 1–25. [CrossRef]
14. Chen, B.; Chen, C. Fast optimization methods for high-dimensional row-sparse multivariate quantile linear regression. *J. Stat. Comput. Simul.* **2024**, *94*, 69–102. [CrossRef]
15. Petrella, L.; Raponi, V. Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *J. Multivar. Anal.* **2019**, *173*, 70–84. [CrossRef]
16. Zhang, Q.; Huang, W.; Jin, C.; Zhao, P.; Shu, Y.; Shen, L.; Tao, D. Multinoulli Extension: A Lossless Yet Effective Probabilistic Framework for Subset Selection over Partition Constraints. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025.
17. Zhong, Y.; Gao, X.; Xu, W. Robust multitask feature learning with adaptive Huber regressions. *Can. J. Stat.* **2025**, *53*, e70022. [CrossRef]
18. Kawano, S.; Fukushima, T.; Nakagawa, J.; Oshiki, M. Multivariate regression modeling in integrative analysis via sparse regularization. *Jpn. J. Stat. Data Sci.* **2025**, 1–28. [CrossRef]
19. Friedman J.; Hastie, T.; Tibshirani, R.; Narasimhan, B; Tay, J.K.; Simon, N.; Yang, J. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R Package Version 4.1-7. 2023. Available online: https://CRAN.R-project.org/package=glmnet (accessed on 14 January 2026)
20. Cox, D.R.; Reid, N. A note on pseudolikelihood constructed from marginal densities. *Biometrika* **2004**, *91*, 729–737. [CrossRef]
21. Gao, X.; Song, P.X.K. Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Stat. Sin.* **2011**, *21*, 165–185.

22. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
23. Parikh, N.; Boyd, S. Proximal algorithms. *Found. Trends Optim.* **2014**, *1*, 127–239. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.