

MULTI-VIEW ORTHOGONAL PROJECTION REGRESSION WITH APPLICATION IN MULTI-OMICS INTEGRATION

BY ZONGRUI DAI^{1,a}, YVONNE J. HUANG^{2,c} AND GEN LI^{*1,b}

¹*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan,* ^adaizr@umich.edu; ^bligen@umich.edu

²*Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan,* ^cyvjhuang@med.umich.edu

Multi-omics integration offers novel insights into complex biological mechanisms by utilizing the fused information from various omics datasets. However, the inherent within- and inter-modality correlations in multi-omics data present significant challenges for traditional variable selection methods, such as Lasso regression. These correlations can lead to multicollinearity, compromising the stability and interpretability of selected variables. To address these problems, we introduce the Multi-View Orthogonal Projection Regression (MVOVR), a novel approach for variable selection in multi-omics analysis. MVOVR leverages the unidirectional associations among omics layers, inspired by the Central Dogma of Molecular Biology, to transform predictors into an uncorrelated feature space. This orthogonal projection framework effectively mitigates the correlations, allowing penalized regression models to operate on independent components. Through simulations under both well-specified and misspecified scenarios, MVOVR demonstrates superior performance in variable selection, outperforming traditional Lasso-based methods and factor-based models. In real-data analysis on the CAARS dataset, MVOVR consistently identifies biologically relevant features, including the *Bacteroidaceae* family and key metabolites which align well with known asthma biomarkers. These findings illustrate MVOVR's ability to enhance variable selection while offering biologically interpretable insights, offering a robust tool for integrative multi-omics research.

1. Introduction. Multi-omics analysis provides a comprehensive understanding of biological mechanisms by integrating multiple types of data, such as genomics, transcriptomics, proteomics, and metabolomics. These datasets offer a novel insight into molecular processes and immunological research that cannot be found from any single modality alone (Chen et al. (2023); Chu et al. (2021); Clark et al. (2021)). Numerous studies have found that the fusion of different omics data can bring additional insights into the exploration of biomarkers, improving diagnostics, and therapy development (Gillenwater et al. (2021); Garg et al. (2024); Olivier et al. (2019); Hussein, Abou-Shanab and Badr (2024); Menyhárt and Győrffy (2021)). For example, in our recent study of CAARS data, we collected the gut microbiome and metabolome data of 51 patients to investigate the combined impact of these two omics layers on asthma development.

Asthma is a complicated respiratory disease which involves airway inflammation and allergic reactions (Gautam, Johansson and Mersha (2022)). As one of the most prevalent chronic airway diseases, it exhibits high heterogeneity, making diagnosis based on a single biomarker challenging (Chung (2016); Abdel-Aziz et al. (2020)). Increasing evidence suggests that the pathogenesis of asthma is closely linked to different omics data. For instance, potential host-microbiota interactions have been associated with an increased risk of asthma (Ruff, Greiling

*[Corresponding author indication should be put in the Acknowledgment section if necessary.]

Keywords and phrases: Multi-omics Integration, Variable selection, Latent Variables.

and Kriegel (2020)). Emerging collaborations are using multi-omics data to advance the understanding of asthma. For instance, multi-omics integration has explored the disease's heterogeneity and underlying pathology. This approach not only helps identify new patient stratification, but also paves the way for personalized treatment strategies (Zhang et al. (2024)). Lasso-based regression is a widely used approach for variable selection in multi-omics data. For example, IPF-LASSO (Integrative LASSO with Penalty Factors) and Priority-Lasso are both Lasso-based methods that account for the heterogeneity of multi-omics data by assigning distinct penalty and priority weights in the regression model (Klau et al. (2018); Boulesteix et al. (2017)). However, some studies have found that the performance of these models can be influenced when highly correlated predictors are present, as these methods do not account for the within- and inter-modality correlations inherent in multi-omics data (Castel, Zhao and Thoresen (2024)).

To illustrate this challenge, we measure the correlation between microbiome and metabolome through the Pearson correlation heatmap in the CAARS dataset. Microbiome data are aggregated to the family level and applied centered log ratio transformation. Metabolome data are centered and scaled. The heatmap shows that metabolome data have strong within-modality correlation compared to the microbiome. Additionally, there are some negative inter-modality correlations between the microbiome and metabolome (Figure.1).

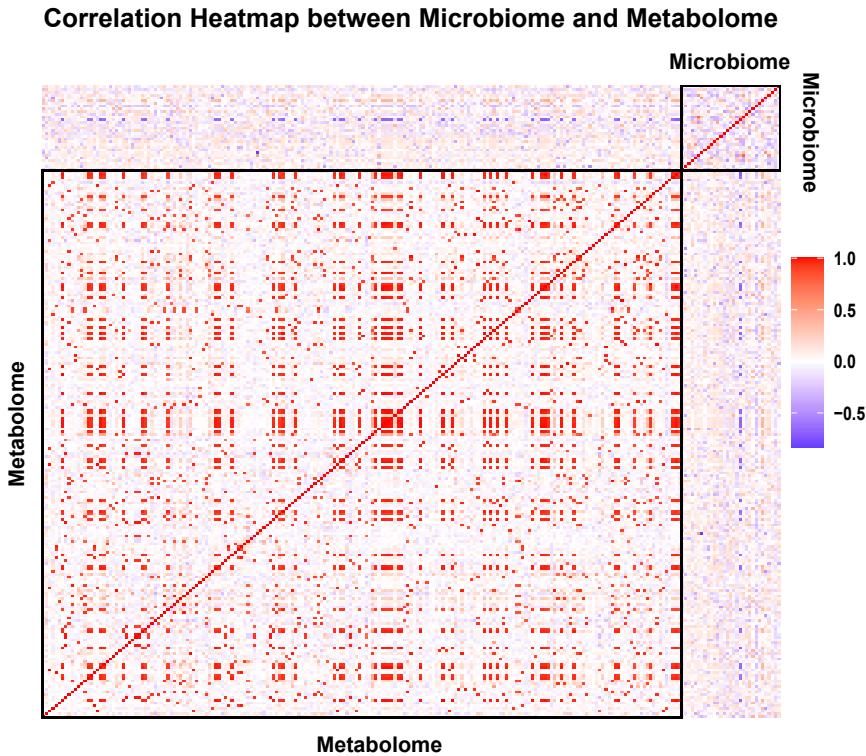


FIG 1. The Pearson Correlation Heatmap of Microbiome and Metabolome

The within- and inter-modalities correlations pose significant challenges for variable selection in traditional regression models. These correlations arise from latent associations between omics layers and intrinsic dependencies within each individual omic. Network-based approaches provide an effective way to capture these complex relationships, such as protein-protein and protein-RNA interactions (Richards, Eckhardt and Krogan (2021); Nasiri et al.

(2021); Szklarczyk et al. (2023)). Although these correlations can be quantified, they introduce complications when modeling the relationship between multi-omics and a response variable. Specifically, these strong correlations can lead to multicollinearity, making it difficult to distinguish the individual contributions of variables.

To better justify the problem, suppose there are two modalities $M_1 \in \mathbb{R}^{n \times p_1}$ and $M_2 \in \mathbb{R}^{n \times p_2}$ which are associated with response vector Y through β_1 and β_2 . A linear model could be constructed below:

$$Y = M_1\beta_1 + M_2\beta_2 + \epsilon_1$$

The existences of within- and inter-modality correlations indicate the latent relationships of M_1 and M_2 below:

$$\underbrace{M_1 = F\Lambda + U}_{\text{Within-Modality Correlation}} \quad \text{and} \quad \underbrace{M_2 = M_1B + E}_{\text{Inter-Modality Correlation}},$$

where F represents the latent factors in modality M_1 with rank r , and $F\Lambda$ captures the low-rank structure of M_1 , which is a commonly used approach to describe within-modality correlation. For instance, factor-based models frequently employ it to represent dependencies within predictors. The inter-modality correlation shows the M_2 modality can be represented by the linear combination of M_1 through the coefficient matrix B with error term E . Standard variable selection methods, such as Lasso regression, struggle in this setting because they will arbitrarily select among correlated predictors, potentially leading to unstable variable selection. Addressing this issue requires a method that not only accounts for the relationships between omics layers but also isolates the independent contributions of each omic.

The factor model has emerged as a powerful tool for correlated data by decomposing them into latent structures comprising factors and idiosyncratic components. For instance, factor-adjusted regularized regression can handle highly correlated data by identifying and removing the low-rank structure from data and retaining the idiosyncratic components for variable selection(Fan, Ke and Wang (2020)). Integrative Factor Regression is another factor decomposition-based model designed for multi-model dataset. It can extract modality-specific factors to account for the heterogeneity across modalities (Li and Li (2022)). However, factor-based models require data to have an approximate latent factor structure and only reduce correlations within individual modalities. Thus, correlations between modalities still exist.

An alternative approach is cooperative learning, which employs an agreement penalty based on contrastive learning. This method encourages different modalities to contribute similarly. By varying the hyperparameter of the agreement penalty, the solutions for this method include the early and late fusion approaches for multi-model data, providing robust performance for different settings (Ding et al. (2022a)). However, this method ignores the inherent correlations between the modalities and enforces different modalities to align the contribution. In multi-omics scenarios, this assumption may not always hold, as different omics layers can have distinct influences on the response variable. For example, miR-155 and miR-146a are well-known miRNAs that can suppress *E. coli*-induced inflammatory responses in neuroinflammation. This example suggests that the host transcriptome may have an opposing effect compared to the microbiome, leading to potential contradictions between the molecular signals originating from the host and those from the microbiota (Yang et al. (2021)).

To address these challenges, we introduce a novel Multi-View Orthogonal Projection Regression (MVOPR) for variable selection in multi-omics data. Unlike existing methods that impose specific structural assumptions on the data, our model leverages unidirectional associations among different omics to mitigate the correlations. Our approach is inspired by the Central Dogma of Molecular Biology, which states that DNA transcribes to RNA, and RNA

translates into protein, while the reverse process is impossible. For instance, once a protein is synthesized, it cannot alter its original RNA template. This inherent directionality in molecular interactions suggests that multi-omics relationships can be represented by a directed graph (digraph) with unidirectional pathways. Building on this biological insight, our method accounts for the dependencies by removing redundant correlations in a structured manner. Specifically, we employ an orthogonal projection framework that sequentially remove the effects of upstream omics layers on downstream ones. By transforming the original multi-omics data into an uncorrelated feature space, our approach ensures that variable selection methods, such as penalized regression, operate on independent components free from both within and across modality correlations. This enables MVOPR to overcome the limitations of standard Lasso-based approaches, noted above.

In this study, we demonstrate the effectiveness of MVOPR for multi-omics variable selection through both theoretical analysis and empirical validation. Our simulations and real-data analysis reveal that MVOPR consistently outperforms existing methods. We also show that when inter-modality correlation exists, the factor-based models will face different problems, unlike MVOPR. Importantly, even in cases where cross-modality correlations are absent, MVOPR remains robust and performs comparably to standard Lasso regression, demonstrating its adaptability across different correlation structures. By incorporating biological direction assumptions, our approach not only enhances variable selection performance but also aligns with the natural structure of molecular data, offering a robust framework for integrative multi-omics analysis.

The rest of the article is organized as follows. In Section 2, we present the MVOPR framework for both the two-modality and multiple-modality scenarios, followed by an introduction to three related methods for multi-modal data analysis. Section 3 provides a comparative analysis of MVOPR against other competing methods under various settings. In Section 4, we apply MVOPR to the CAARS dataset and evaluate its performance relative to alternative approaches.

2. Methodology.

2.1. MVOPR for Two Modalities. Suppose we have two modalities, $M_1 \in \mathbb{R}^{n \times p}$ and $M_2 \in \mathbb{R}^{n \times q}$. Let Y be the response vector of length n , assumed to be associated with M_1 and M_2 through regression coefficients $\beta_1 \in \mathbb{R}^p$ and $\beta_2 \in \mathbb{R}^q$. The relationship is modeled as:

$$(1) \quad Y = M_1\beta_1 + M_2\beta_2 + \epsilon_1,$$

where ϵ_1 is an error term assumed to be uncorrelated with M_1 and M_2 , with $E(\epsilon_1) = 0$ and $\text{Var}(\epsilon_1) = \sigma_{\epsilon_1}^2 I$. We further assume that M_2 is influenced by M_1 through a low-rank coefficient matrix $B \in \mathbb{R}^{p \times q}$ of rank r , with an error component $E \in \mathbb{R}^{n \times q}$ that is uncorrelated with M_1B and ϵ_1 :

$$(2) \quad M_2 = M_1B + E.$$

Using this inter-modality correlation (2), we can reformulate Model (1) as:

$$(3) \quad Y = M_1(\beta_1 + B\beta_2) + E\beta_2 + \epsilon_1.$$

If E is small, the model becomes almost unidentifiable, which affects the selection of the variables for M_2 . To handle this issue, we aim to remove the associated component M_1B from M_2 while retaining only the uncorrelated part E .

Let $U\Sigma V^T$ be the singular value decomposition (SVD) of M_1B , where U_r consists of the first r left singular vectors of U . Substituting this decomposition into Model (3) gives:

$$(4) \quad Y = M_1\beta_1 + E\beta_2 + U_r\gamma_1 + \epsilon_1,$$

where γ_1 is a nuisance parameter. Since U_r captures the principal directions of M_1B , it is highly correlated with M_1 . To eliminate this correlation, we project M_1 onto a subspace orthogonal to U_r .

Define the projection matrix $P = U_r U_r^T$, and let $P^\perp = I - P$ be its orthogonal complement. Transforming M_1 into $P^\perp M_1$ ensures that the new predictor no longer lies in the column space of M_1B , thereby breaking the correlation with U_r . The transformed MVOPR model for two modalities is then:

$$(5) \quad Y = M_1^* \beta_1 + M_2^* \beta_2 + U_r \gamma^* + \epsilon_1,$$

where $M_1^* = P^\perp M_1$ and $M_2^* = E$. In this formulation, the predictors M_1^* , M_2^* , and U_r are mutually uncorrelated.

This transformation enhances variable selection for both β_1 and β_2 by removing redundant correlations between modalities. The decomposition effectively subtracts the linear contribution of M_1 in M_2 and projects M_1 outside the column space of M_1B , ensuring no internal dependencies between M_1^* , M_2^* , and the nuisance components U_r .

2.2. MVOPR for Multiple Modalities. Extending the model to three modalities, let $M_1 \in \mathbb{R}^{n \times p_1}$, $M_2 \in \mathbb{R}^{n \times p_2}$, and $M_3 \in \mathbb{R}^{n \times p_3}$, with a known hierarchical dependency:

$$M_2 = M_1 B_{2,1} + E_2,$$

$$M_3 = M_1 B_{3,1} + M_2 B_{3,2} + E_3.$$

where $E_2 \in \mathbb{R}^{n \times p_2}$ and $E_3 \in \mathbb{R}^{n \times p_3}$ are independent error components, and $B_{2,1}$, $B_{3,1}$, and $B_{3,2}$ are low-rank coefficient matrices with ranks r_1 , r_2 , and r_3 , respectively.

The response Y is modeled as:

$$(6) \quad Y = M_1 \beta_1 + M_2 \beta_2 + M_3 \beta_3 + \epsilon_1.$$

To remove the associated components, define two projection matrices: $P_1 = U_1 U_1^T$ and $P_2 = U_2 U_2^T$ which based on $M_1 B'_1$ and $E_2 B_{3,2}$ separately. $B'_1 = (B_{2,1}, B_{3,1})$ is the concatenation of $B_{2,1}$ and $B_{3,1}$. The transformed modalities are:

$$M_1^* = (I - P_1) M_1, \quad M_2^* = (I - P_2) E_2, \quad M_3^* = E_3.$$

Transforming the modalities in model (6), the MVOPR model for three modalities is:

$$(7) \quad Y = M_1^* \beta_1 + M_2^* \beta_2 + M_3^* \beta_3 + U_1 \gamma_1 + U_2 \gamma_2 + \epsilon_1.$$

where γ_1 and γ_2 are nuisance parameters. The detailed derivations are provided in Supplementary Appendix A.

For more than three modalities, the transformation follows a similar procedure. Suppose there are k modalities with features p_1, p_2, \dots, p_k . If each modality M_j (for $j = 2, 3, \dots, k$) depends only on previous modalities:

$$M_j = \sum_{i=1}^{j-1} M_i B_{j,i} + E_j,$$

where E_j is independent noise, then the final regression model is:

$$(8) \quad Y = M_1 \beta_1 + M_2 \beta_2 + \dots + M_k \beta_k + \epsilon_1.$$

With the direction assumption above, we could derive the connection between response Y and all the modalities by the following algorithm.

Algorithm 1 Algorithm for multi-view regression on multiple modalities

-
- 1: **Input:** Multiple modalities M_1, M_2, \dots, M_k and response Y
 - 2: **Step.1:** Obtain the estimation of $B_{2,1}, B_{3,1}, \dots, B_{k,k-1}$
 - 3: **for** j in 1:m
 - 4: Regress $M_j \sim (M_1, \dots, M_{j-1})$. Calculate the residuals by $\hat{E}_j = M_j - \hat{M}_j$
 - 5: **Step.2:** Obtain the projection matrix P_1, \dots, P_{m-1}
 - 6: Calculate:
 - 7: $M_1 \hat{B}_1 = M_1(\hat{B}_{2,1}, \hat{B}_{3,1}, \dots, \hat{B}_{k,1})$
 - 8: $\hat{E}_2 \hat{B}'_2 = \hat{E}_1(\hat{B}_{3,2}, \hat{B}_{4,2}, \dots, \hat{B}_{k,2}), \dots, \hat{E}_{k-1} \hat{B}'_{k-1} = \hat{E}_{k-1} \hat{B}_{k,k-1}$.
 - 9: Obtain the SVD:
 - 10: $M_1 \hat{B}'_1 = U_1 \Sigma_1 V_1^T$ and projections $P_1 = U_1 U_1^T$ with rank r_1 .
 - 11: For $j \geq 2$, $\hat{E}_j \hat{B}'_j = U_j \Sigma_j V_j^T$ and projections $P_j = U_j U_j^T$ with rank r_j .
 - 12: **Step.3:** Transform the M_1, M_2, \dots, M_m by $M_1^*, M_2^*, \dots, M_m^*$
 - 13: $j = 1: M_1^* = P_1^\perp M_1$.
 - 14: $j = 2, \dots, k-1: M_j^* = P_j^\perp E_{M_j}$.
 - 15: $j = k: M_k^* = E_{M_k}$
 - 16: Obtain the nuisance variable $U = (U_1, U_2, \dots, U_{m-1})$
 - 17: **Step.4:** Obtain the estimation of β_1, \dots, β_k
 - 18: Solve the penalized optimization:
 - 19: $\min_{\beta, \gamma} \|Y - M_1^* \beta_1 - \dots - M_k^* \beta_k - U \gamma\|^2 + \sum_{j=1}^k P_\lambda(\beta_j)$
 - 20: **return** $\hat{\beta}_1, \dots, \hat{\beta}_k$
-

2.3. Related methods. To evaluate the relative performance of MVOPR, we consider several alternative models for multi-modality data. Specifically, we compare our method against Cooperative Regularized Linear Regression ([Cooperative Ding et al. \(2022b\)](#)), Integrative Factor Regression ([IntegFactor Li and Li \(2022\)](#)), and Factor-Adjusted Regularized Regression ([Factor Fan, Ke and Wang \(2020\)](#)).

2.3.1. Cooperative Regularized Linear Regression. Cooperative regularized linear regression is a widely used approach for multi-view learning. It integrates multiple modalities by imposing an agreement penalty that encourages the predictions from different modalities to be aligned. The level of agreement between modalities is controlled by the hyperparameter ρ . When $\rho = 0$, this method becomes traditional penalized regression with chosen penalty. When $\rho = 1$, it indicates a late fusion of all the modalities. Suppose there are two modalities M_1 and M_2 , its least square problem can be written as below:

$$(9) \quad \min_{\beta_1, \beta_2} \|Y - M_1 \beta_1 - M_2 \beta_2\|^2 + \frac{\rho}{2} \|M_1 \beta_1 - M_2 \beta_2\|^2 + \lambda_1 \|\beta_1\|_1 + \lambda_2 \|\beta_2\|_1$$

To simplify the optimization, λ_1 and λ_2 are equal in this study. Problem (9) is convex, we can transform the original data below:

$$(10) \quad \tilde{X} = \begin{pmatrix} M_1 & M_2 \\ -\sqrt{\rho}M_1 & \sqrt{\rho}M_2 \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \quad \tilde{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Based on the transformed data, the problem (9) is equivalent problem to the generic lasso problem below:

$$(11) \quad J(\theta_x, \theta_z) = \left\| \tilde{Y} - \tilde{X} \tilde{\beta} \right\|^2 + \lambda \left\| \tilde{\beta} \right\|_1$$

2.3.2. Factor-based Models. Factor-based methods assume that predictor M follows an approximate factor model

$$(12) \quad M = F\Lambda + U,$$

where F is a $K \times 1$ vector of latent factors, Λ is a $p \times K$ loading matrix, and U is the $p \times 1$ vector of idiosyncratic components. By separating the idiosyncratic components from the M , it de-correlates the original M to a weakly correlated element u . The regression model:

$$(13) \quad Y = M\beta + \epsilon$$

can then be reformulated as:

$$(14) \quad Y = U\beta + F\gamma + \epsilon, \quad \text{where } \gamma = \Lambda\beta \text{ is a nuisance parameter.}$$

To estimate the factors \hat{F} and idiosyncratic components \hat{U} from M , we adopt the method of Bai and Li ([Bai and Li \(2012\)](#)) and Fan et al. ([Fan, Liao and Mincheva \(2013\)](#)). The optimal number of latent factors K is selected based on Bai and Ng's information criteria method ([Bai and Ng \(2002\)](#)). Using these estimates, the least squares problem can be written as follows, where γ is the nuisance parameter.

$$\min_{\beta, \gamma} \|Y - \hat{U}\beta - \hat{F}\gamma\|^2 + \lambda\rho(\beta)$$

In Factor-Adjusted Regularized Regression, the multi-modal data is treated as a unified design matrix, and factor decomposition is applied globally across the entire dataset. Specifically, let $M = (M_1, \dots, M_m)$ represent the concatenation of all modalities ([Fan, Ke and Wang \(2020\)](#)). While Integrative Factor Regression targets multimodal data by modeling each modality separately, allowing for the extraction of modality-specific latent factors and idiosyncratic components. That is, for each modality M_i has its own factors F_i and idiosyncratic component U_i with $i \in 1, \dots, m$. Then, the regression model is fitted using the concatenated idiosyncratic components and latent factors, where $U = (U_1, \dots, U_m)$ and $F = (F_1, \dots, F_m)$ represent the concatenation of all modality-specific idiosyncratic components and latent factors, respectively.

In our setting, $M_2 = M_1B + E$ can be interpreted as an approximate factor model with $M_2 = F\Lambda + E$, where $F = U_r$ and $\Lambda = \Sigma V_r^T$, given that $M_1B = U_r\Sigma V_r^T$. However, despite this approximate factor structure, factor-based models are not well-suited for our problem. The decomposition used in Integrative Factor Regression will introduce a correlated nuisance parameter F , which may obscure the true effect of M_1 . Furthermore, when $M_1(I, B)$ lacks spiked eigenvalues, selecting an appropriate number of factors becomes challenging in Factor-Adjusted Regularized Regression. This often results in choosing an excessively large number of factors, distorting the contribution of M_1 and leading to suboptimal model performance. Factor-based models impose structural assumptions that may not align well with the dependencies present in multi-modal data. The risks associated with obscuring meaningful relationships, introducing highly correlated nuisance parameters, and improperly selecting the number of factors make these methods less effective in our problem setting. A more detailed discussion of this issue is provided in the Supplementary Material Appendix [A.2](#).

2.4. Estimation. To fit model (2), we first need to estimate the coefficient matrix \hat{B} that captures the relationship between M_1 and M_2 . Several well-established reduced-rank regression methods can be utilized for this estimation. For instances, row-sparse reduced-rank regression ([Chen and Huang \(2012\)](#)), sparse orthogonal factor regression ([Uematsu et al.](#)

(2019)), and multivariate reduced-rank linear regression (Chen, Dong and Chan (2013)) provide different sparsity assumptions for estimating \hat{B} . In general, the reduced-rank regression problem can be formulated as the following optimization problem:

$$(15) \quad \begin{aligned} & \min_{U,D,V} \|M_2 - M_1 UDV^T\|_F^2 + \lambda_1 \|D\|_1 + \lambda_2 \rho_a(UD) + \lambda_3 \rho_b(VD) \\ & \text{s.t. } U^T U = I, V^T V = I, B = UDV^T \end{aligned}$$

where the $U^T U = I, V^T V = I$ are introduced for identifiable purpose. ρ_a and ρ_b are penalty functions. They can be entry-wise L_1 norm or row-wise $L_{2,1}$ norm. $\lambda_1, \lambda_2, \lambda_3$ are the tuning parameters that control the magnitude of regularization. This framework generalizes several well-known methods: Row-sparse Reduced-Rank Regression when $\lambda_1 = \lambda_3 = 0$ and $\rho_a = \|\cdot\|_{2,1}$; Multivariate Reduced-Rank Linear Regression when $\lambda_1 = \lambda_2 = \lambda_3 = 0$; Sparse Orthogonal Factor Regression when all tuning parameters are nonzero. The tuning parameters and rank r are chosen based on the GIC (Fan and Tang (2013)). With the fitted model above, we could obtain the coefficient matrix \hat{B} and residual term \hat{E} .

Next, we estimate the P by the inner product of the first r left singular vectors U'_r from $M_1 \hat{B}$. Denote the estimation as \hat{P} . Then, transformed M_1 and M_2 can be estimated based on previous procedures. Once the transformed matrices are obtained, we estimate $\hat{\beta}_1$ and $\hat{\beta}_2$. This is done by solving the following penalized least squares problem:

$$(16) \quad \min_{\beta_1, \beta_2, \gamma_2} \|Y - \hat{M}_1^* \beta_1 - \hat{M}_2^* \beta_2 - U_r \gamma_2\|^2 + \lambda \rho(\beta_1) + \lambda \rho(\beta_2)$$

where ρ is a generic penalty function including the L1 norm, adaptive Lasso, MCP, and SCAD penalties. λ is a tuning parameter that controls the regularization power on both β_1 and β_2 .

For MVOPR with three modalities, the estimation of β_1, β_2 , and β_3 follows a similar penalized least squares approach:

$$(17) \quad \begin{aligned} & \min_{\beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2} \|Y - \hat{M}_1^* \beta_1 - \hat{M}_2^* \beta_2 - \hat{M}_3^* \beta_3 - U_1 \gamma_1 - U_2 \gamma_2\|^2 \\ & + \lambda \rho(\beta_1) + \lambda \rho(\beta_2) + \lambda \rho(\beta_3) \end{aligned}$$

where $\hat{M}_1^* = (I - \hat{P}_1) M_1$, $\hat{M}_2^* = (I - \hat{P}_1) \hat{E}_2$, and $\hat{M}_3^* = \hat{E}_3$. U_1 and U_2 are the left singular vectors with non-zero singular values of $M_1(\hat{B}_{2,1}\hat{B}_{3,1})$ and $E_2\hat{B}_{3,2}$.

3. Numerical analysis.

3.1. Variable selection on two modalities.

3.1.1. *E with identity covariance matrix.* To assess the performance of MVOPR in comparison to other methods, we carry out some simulations under different noise levels on ϵ_1 and ϵ_2 . In this simulations, suppose there are two modalities M_1 and M_2 with 300 features and 200 observations. M_1 is generated from multivariate normal distribution $MVN(0_p, \Sigma_{M_1})$ with identity covariance matrix. Assume M_2 is connected with M_1 through a low-rank row sparse coefficient matrix B with 95% rows as zeros with rank $r = 1$. Response Y is associated with both M_1 and M_2 through β_1 and β_2 . β_1 and β_2 are generated with 290 zeros and 10 non-zeros coefficients. The values of non-zero coefficients are sampled from uniform distribution $U(1, 2)$. We fix the signal-to-noise ratio (SNR) to be 100 for ϵ_1 . By varying the SNRs of ϵ_2 , we compare the variable selection performance of each model by AUC. Each AUC is calculated based on a 100 length grid of λ which controlling strength of sparsity. We conduct each simulation experiment 100 times for one SNR of ϵ_2 .

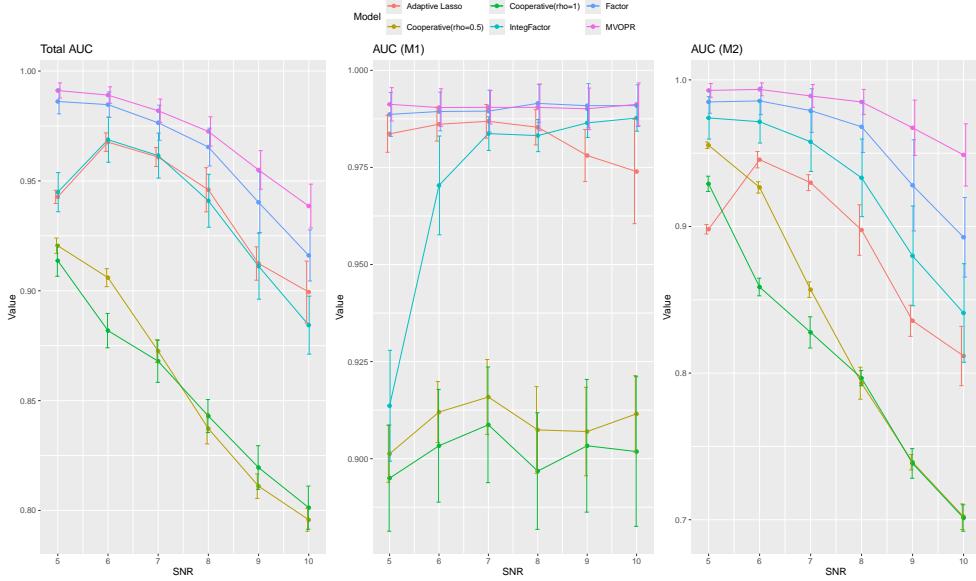


FIG 2. The AUC for each model by varying the SNR of ϵ_2 from 5 to 10

MVOPR outperforms other methods in terms of AUC across the entire range of SNR values (Figure 2). Factor-Adjusted Regularized Regression exhibits comparable performance to MVOPR in scenarios with low SNR. However, as the SNR increases, which corresponds to stronger correlations between M_1 and M_2 , MVOPR demonstrates clear superiority over Factor-Adjusted Regularized Regression. Especially, MVOPR has evident benefits on variable selection for M_2 when SNR is large. It shows the ability of MVOPR to better integrate information across modalities which allows it to maintain high AUC values even under more challenging conditions. In contrast, Factor-Adjusted Regularized Regression appears to struggle under high SNR conditions, likely due to its reliance on factor decomposition. Moreover, other competing methods, such as Integrative Factor Regression and Cooperative learning method, show declining performance as the SNR increases. These methods appear to be less effective in maintaining robust performance when faced with strong inter-modality correlations, highlighting the advantage of MVOPR in such scenarios.

Two alternative simulations are designed to show factor-based model may not be the ideal model to account for the inter-modality correlations. In the first simulation, M_1 and M_2 are generated from multivariate normal distribution with identity covariance matrix with 50 and 300 features. Each has 200 samples. Suppose M_2 is connected with M_1 through a low-rank row sparse coefficient matrix B 70% rows as zeros with rank $r = 9$. β_1 and β_2 only has 10 non-zero coefficients separately. The second simulation is used to showcase the performance of MVOPR under low-dimensional data compared. M_1 and M_2 are generated from multivariate normal distribution with identity covariance matrix with 200 samples and 50 features. Low-rank row sparse coefficient matrix B has 50% rows as zeros with rank $r = 3$. β_1 and β_2 has 25 non-zero coefficients separately. The SNR for ϵ_1 in both simulations are fixed as 100.

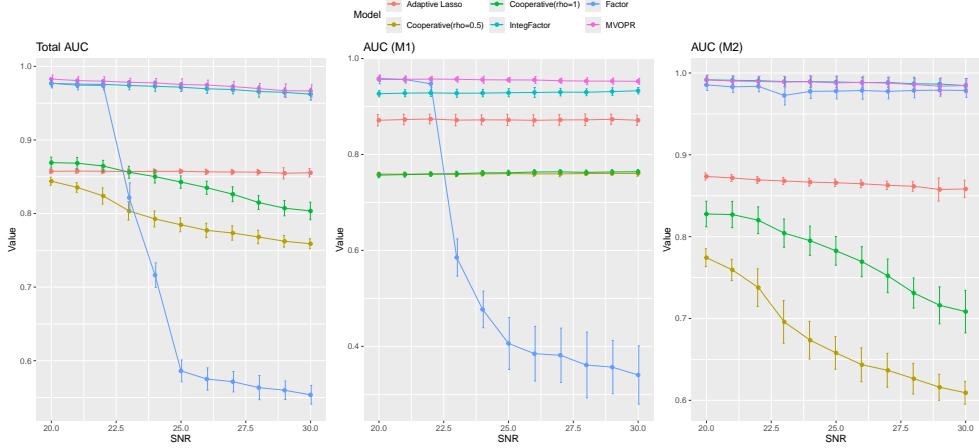


FIG 3. The AUC for each model by varying the SNR of ϵ_2 from 20 to 30 when M_1 and M_2 has different number of features.

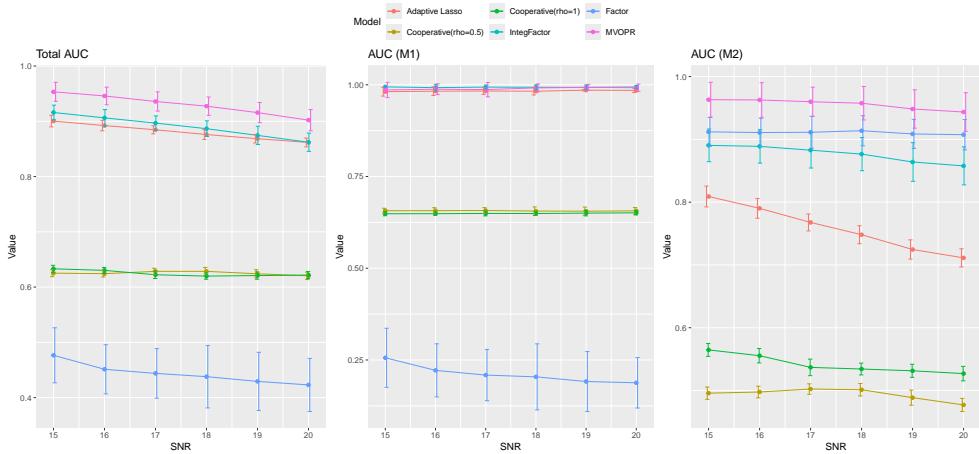


FIG 4. The AUC for each model by varying the SNR of ϵ_2 from 15 to 20 when M_1 and M_2 has 50 features respectively.

Under strong correlations between M_1 and M_2 , MVOPR performs consistently well in both M_1 and M_2 variable selections compared to other method (Figure.3, 4). Factor-Adjusted Regularized Regression shows comparable performance to MVOPR when SNR is smaller than 22. However, as inter-modality correlation become stronger, its variable selection ability for M_1 declines dramatically. This decline is likely attributed to the selection of excessively large number of factors, which overwhelms the meaningful signal and reduces its performance in isolating relevant variables from M_1 . This problems become more obvious under low-dimensional simulation, where (M_1, M_2) are more likely to be decomposed to a structure with excessive factors (Figure.4). Integrative Factor Regression consistently underperforms in selecting variables for M_1 , even when it selects zero number of factors for this modality. It can be attributed to its correlated nuisance variables with M_1 , which disrupts the variable selection for predictors. This limitation underscores the difficulty of maintaining a balance between factor decomposition and effective variable selection in the presence of strong inter-modal correlations (Figure.3).

3.2. Misspecified Case.

3.2.1. E with correlated structure. In real world settings, E may not always have independent covariance structure. To verify whether MVOPR can still work under this misspecified case, we consider two covariance patterns including auto-regressive (AR1) and compound symmetry (CS). In the simulations below, we generate two modalities M_1 and M_2 while each has 300 features and 200 observations. M_1 is generated from $MVN(0_p, \Sigma_{M_1})$ with identity covariance matrix. M_2 is associated with M_1 through a low-rank row sparse coefficient matrix B 50% rows as zeros with rank $r = 1$. Response Y is associated with both M_1 and M_2 through β_1 and β_2 . β_1 and β_2 are generated with 90 zeros and 10 non-zeros coefficients. The absolute values of non-zero coefficients are sampled from uniform distribution $U(1, 2)$. The SNR for ϵ_1 and ϵ_2 are fixed to be 3 and 5. We compare the variable selection performance of each model by AUC. In AR1 case, E is generated from $MVN(0_q, \Sigma_\rho)$. The diagonal elements of Σ_ρ are 1 with $cov(\epsilon_2^i, \epsilon_2^j) = \rho^{|i-j|}$. $\rho = 0.9$ and $\rho = 0.95$ conditions are included. The results are shown in Figure 5.A and Figure 5.B. Under this misspecified case, MVOPR still achieves a higher AUC than other methods. In compound symmetry case, E are generated from a $MVN(0_q, \Sigma_\mu)$. The diagonal elements of Σ_μ are 1 with $cov(\epsilon_2^i, \epsilon_2^j) = \mu$. We test the performance of each model under $\mu = 0.7$ and $\mu = 0.9$ condition. The results are shown in Figure 5.C and Figure 5.D. Under this condition, both MVOPR and factor-based models performs well compared to adaptive lasso.

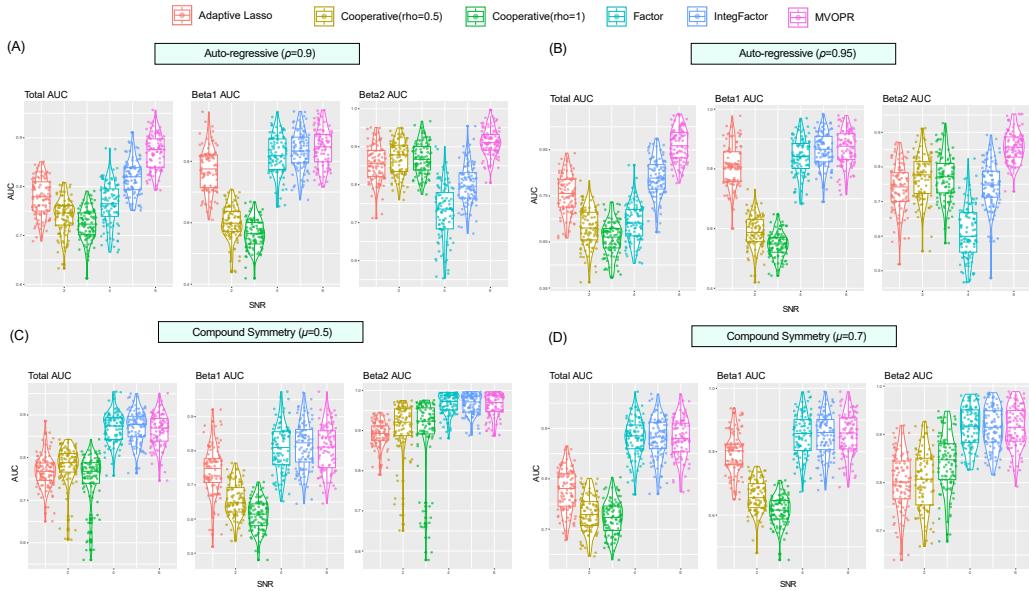


FIG 5. The AUC of each model when ϵ_1 has certain correlated structure. Fig.2.A-B showcase the AUC for each model under Auto-Regressive (AR1) covariance pattern. Fig.2.C-D showcase the AUC for each model under Compound Symmetry (CS) covariance pattern.

3.2.2. Null Experiment: No inter-modality correlation. To evaluate whether MVOPR can still perform well when the $M_2 = M_1B + \epsilon$ assumption is missing, we generate both M_1 and M_2 from $MVN(0_p, \Sigma_{M_{1,2}})$ independently. In this simulation, we treat $\Sigma_{M_{1,2}}$ as identity or auto-regressive ($\rho = 0.9$) covariance matrix. Suppose both M_1 and M_2 have 100 features and 200 samples. Y is associated with M_1 and M_2 based on β_1 and β_2 which are generated based on the same rule in the previous section 3.2.1.

Based on the results in Figure 6 and Figure 7, we notice that four models perform similarly under $\Sigma_{M_{1,2}} = I$. Meaning that even when the unidirectional assumption is missing, MVOPR can still work and share similar performance to other methods. However, if $\Sigma_{M_{1,2}}$ follows an auto-regressive ($\rho = 0.9$) covariance pattern, factor-based models exhibit weaker performance compared to adaptive Lasso and MVOPR. This may be attributed to the covariance structure of each modality, as the absence of spiked eigenvalues hinders the effectiveness of factor decomposition.

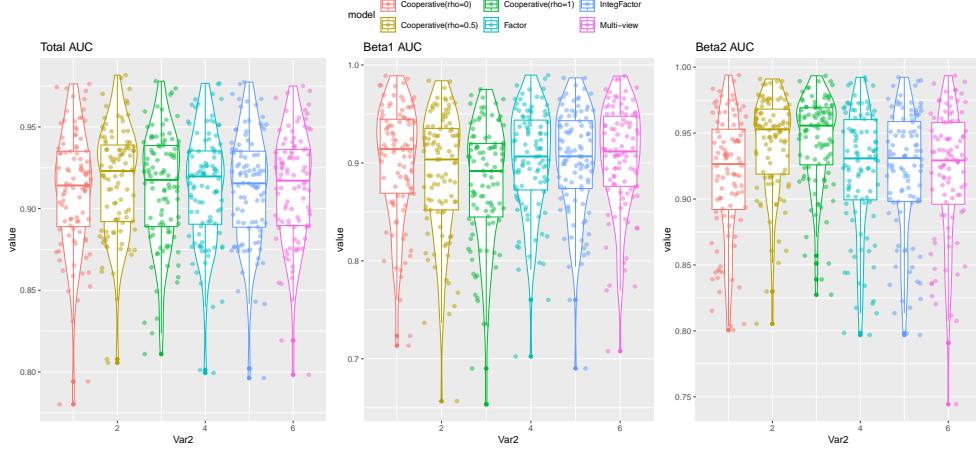


FIG 6. The AUC of each model when both M_1 and M_2 have diagonal covariance matrix without the $M_2 = M_1 B$ assumption

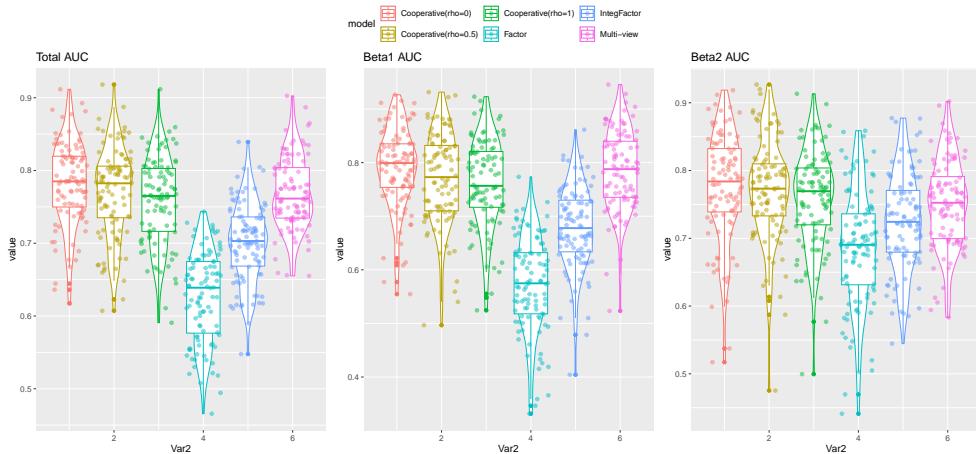


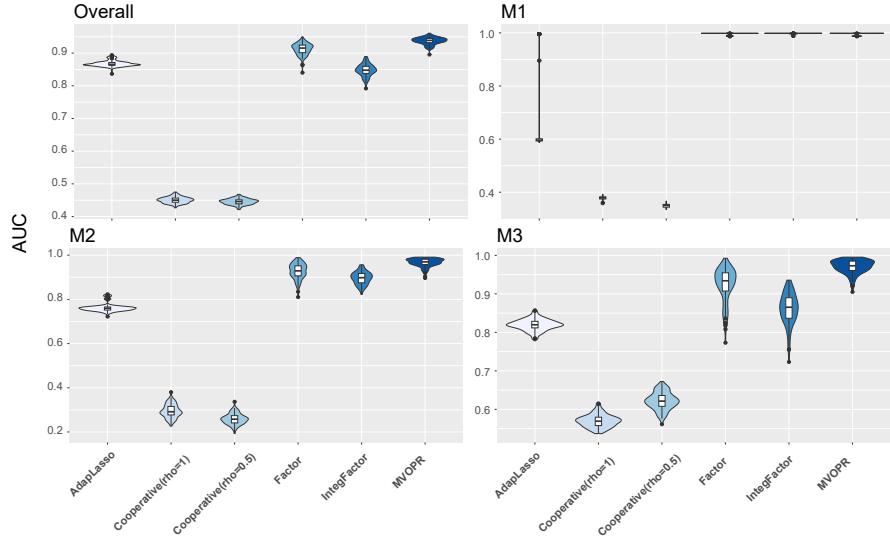
FIG 7. The AUC of each model when both M_1 and M_2 have autoregressive ($\rho = 0.9$) covariance matrix without the $M_2 = M_1 B$ assumption

3.3. Simulation for Multi-modalities. To evaluate the empirical performance of MVOPR on multi-modalities condition, we consider three modalities case from model (7). Suppose there are three modalities M_1 , M_2 , and M_3 . Each modality has the same number of variables $p = p_1 = p_2 = p_3 = 100$ with 100 observations. B_1 , B_2 , and B_3 are three low-rank coefficient matrix with rank $r_1 = 3$, $r_2 = r_3 = 1$. B_1 , B_2 , and B_3 are dense matrices with no row-wise

sparsity. Y is the response variable associated with M_1 , M_2 , and M_3 .

The estimations of \hat{B}_1 , \hat{B}_2 , and \hat{B}_3 are based on Multivariate Reduced-Rank Regression. E_2 and E_3 are generated from $MVN(0_p, \Sigma)$, while Σ follows the identity. In this simulation, the SNRs for E_2 and E_3 are fixed to be 10 and 20. We also consider a misspecified case where E_2 and E_3 has correlated covariance structures.

(A)



(B)

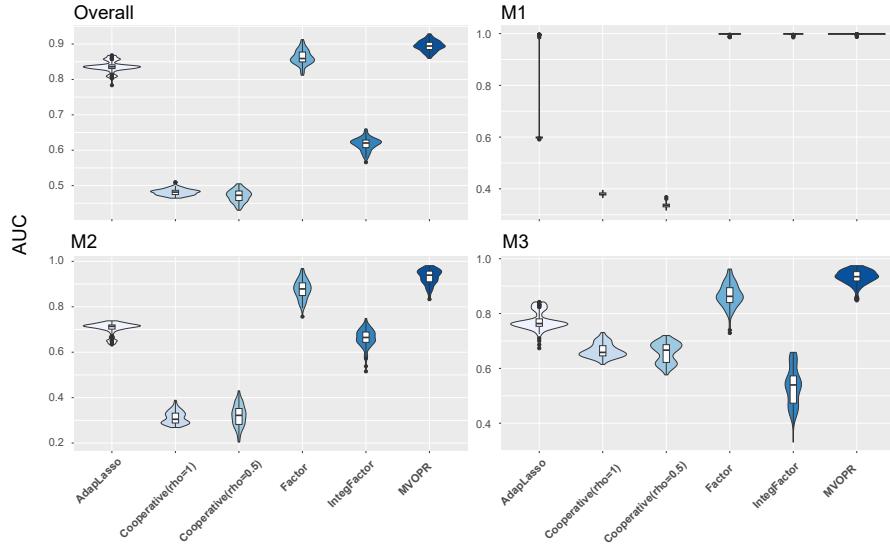


FIG 8. AUC of the variable selection in M_1 , M_2 , and M_3 . (A) The AUC distributions of MVOPR and other methods when E_2 and E_3 have identity covariance; (B) The AUC distributions of MVOPR and other methods when E_2 and E_3 have ARI covariance

Based on the Figure 8, we find that MVOPR outperforms other methods in terms of overall AUC, AUC in M_2 , and AUC in M_3 . MVOPR achieves the highest mean AUC values in these categories, indicating its superior performance in multi-modal integration. Among the

competing methods, factor-based models show some improvement in overall and AUC for M_2, M_3 compared to Adaptive Lasso. However, their performance is worse than MVOPR. Specifically, Integrative Factor Regression exhibits a notable decline in performance when E_2 and E_3 share a correlated covariance structure. This result suggests that factor-based models may struggle to capture the intricate inter-modality correlations. For Cooperative Learning, these models perform worse than adaptive Lasso, indicating that the agreement penalty may not always bring benefits to variable selection in these settings. The simulation results reveal the robustness and superiority of MVOPR in handling complex multi-omics settings. Even in misspecified scenarios, MVOPR consistently outperforms competing methods, demonstrating its reliability and effectiveness in capturing intricate correlations.

4. Real data analysis.

4.1. CAARS Data Analysis.

We conduct the MVOPR model on the CAARS data, collected from 55 patients. This dataset contains two omics layers: microbiome and metabolome. The study aims to understand how the omics data influence the continuous eosinophil count. To reduce the dimensionality of microbiome and metabolome, we only select the metabolites which have top 200 variance. 139 microbiome are aggregated to 31 family levels. Then, we normalize the microbiome by centered log ratio transformation and the metabolome data is centered and scaled. We use the square root of continuous eosinophil count as response.

Multivariate reduced rank regression is applied to estimate the coefficient matrix \hat{B} with the metabolome as a response and the microbiome as predictor. Original omics datasets are transformed based on the \hat{B} and residuals \hat{E} . To analyze the association between the square root of continuous eosinophil count and transformed data, L1 penalty is used for variable selection. Using a leave-one-out sample to qualify the robustness for variable selection. Out-sample MSE and stability indicators are used to qualify the performance of each model. If one feature is selected with non-zero coefficient among 85% iterations, it will be considered as a selected feature. Stability indicators are defined as follows: suppose the i th set of variables selected by the model during the i th iterations of leave-one-out sample as S_i . S_j and S_i are paired to calculate the stability. The two pairs are represented as i and j while $i \neq j$. Here are three stability indicators: Jaccard similarity coefficient, Otsuka–Ochiai coefficient, and Sørensen–Dice coefficient ([Kwon et al. \(2023\)](#)).

$$Jaccard(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \quad Ochiai(S_i, S_j) = \frac{|S_i \cap S_j|}{\sqrt{|S_i| \cdot |S_j|}} \quad Dice(S_i, S_j) = \frac{2|S_i \cap S_j|}{|S_i| + |S_j|}$$

In the results (Table [4.1](#)), MVOPR achieves the lowest Mean Squared Error (MSE) of 177.73 and highest Jaccard similarity (0.66), Otsuka–Ochiai coefficient (0.75), and Sørensen–Dice coefficient (0.74). Those results reflect its robustness in variable selection and model stability. Additionally, MVOPR successfully selects five features. Notably, these selected features also appear as non-zero coefficients in traditional Lasso regression during some iterations. However, their selection frequencies are low, and their confidence intervals cross zero, indicating a lack of significance and stability in the traditional Lasso approach. In contrast, MVOPR not only identifies these features consistently, but also provides stronger evidence of their significance. The Factor-Adjusted Regularized Regression and Integrative Factor Regression models have relatively worse stability and MSE compared to MVOPR. These results suggest that although factor-based models may partially account for the within-modality correlations, they may not fully leverage the inter-modality correlation as effectively as MVOPR.

| Models | MSE | Jaccard | Otsuka–Ochiai | Sørensen–Dice | Selected Features |
|-------------------------------|--------|---------|---------------|---------------|-------------------|
| Multi-view regression | 177.73 | 0.66 | 0.75 | 0.74 | 5 |
| Lasso regression | 223.44 | 0.27 | 0.35 | 0.32 | 0 |
| Factor Regression | 188.59 | 0.44 | 0.62 | 0.56 | 1 |
| Integrative Factor Regression | 414.49 | 0.41 | 0.53 | 0.47 | 1 |

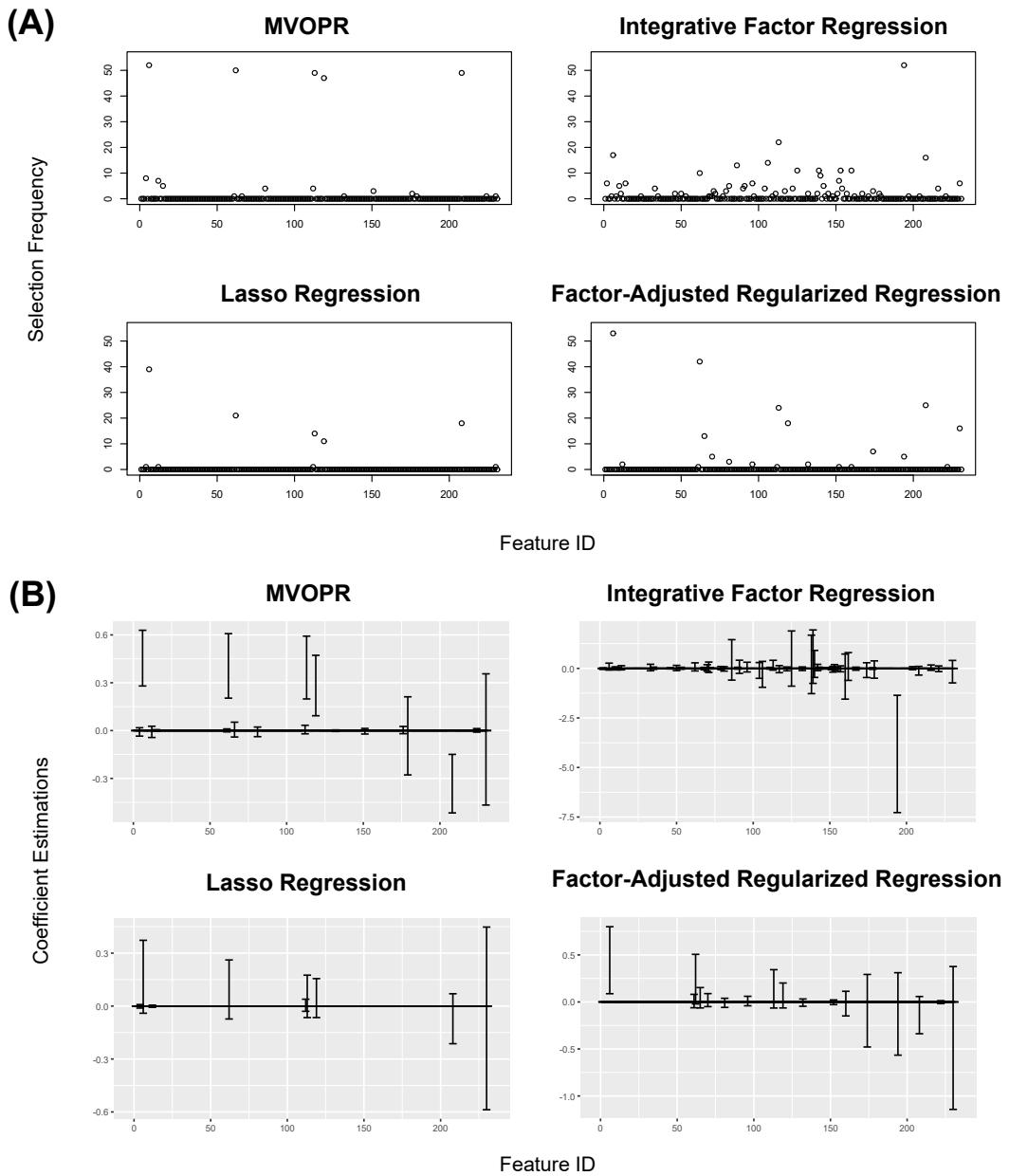


FIG 9. MVOPR for CAARS Data Analysis. (A) Selection frequency Microbiome (ID: 1-31) and Metabolome (ID: 32 - 231); (B) Confident Intervals for Coefficients Estimations; (C) Pearson Correlation Matrix between Selected Microbiome and Metabolome

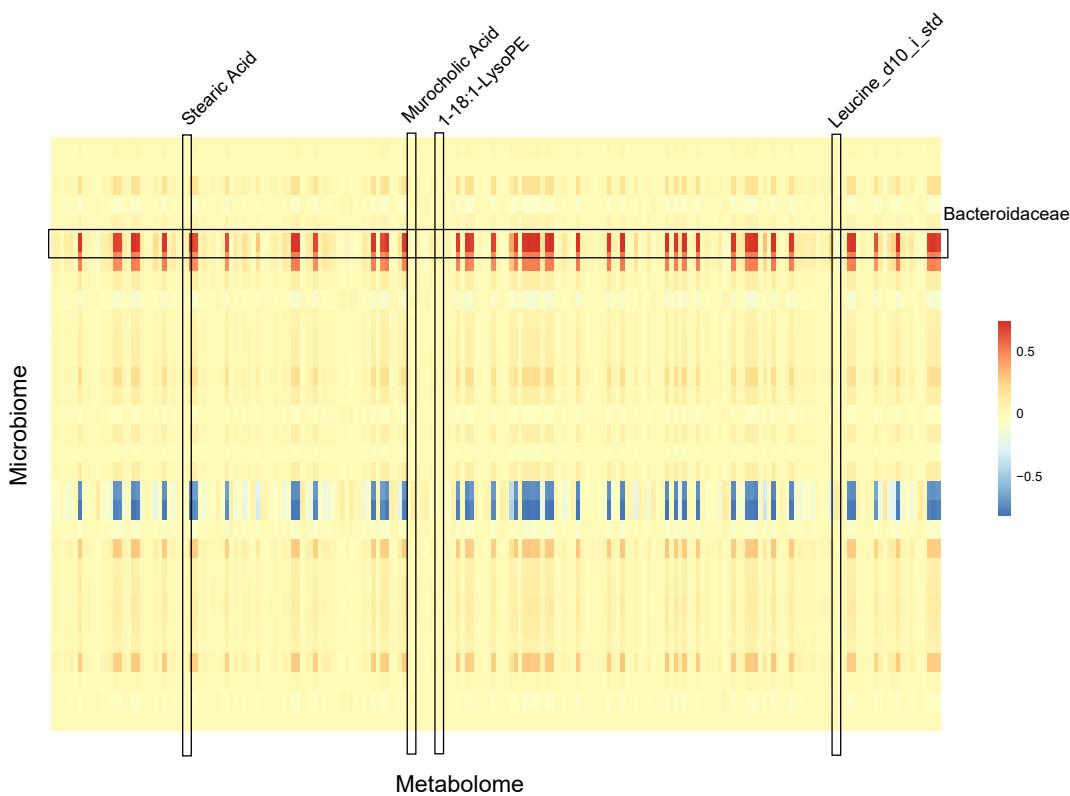


FIG 10. Pearson Correlation Matrix between Selected Microbiome and Metabolome within MVOPR

In MVOPR, *Bacteroidaceae* family is consistently selected as a nonzero coefficient across 52 iterations, showing a positive average effect on the square root-transformed continuous eosinophil count. This result aligns with previous research suggesting that an increased relative abundance of *Bacteroidaceae* is associated with asthma development (Zimmermann et al. (2019)). Within *Bacteroidaceae*, the genera *Bacteroides* plays a particularly important role in asthma pathophysiology. Some studies have identified *Bacteroides* as a key microbial component in asthma progression (Mahdavinia et al. (2023); Aslam et al. (2024); Fiuzo et al. (2024)). Among the four metabolites, stearic acid (tearic_acid_duplicate_2), murocholic acid (murocholic_acid_duplicate_2), 1-18:1-LysoPE (lyso_pe_18_1_9z_0_0_duplicate_2), and leucine (leucine_d10_i_std). Stearic acid has been previously identified as a biomarker for asthma, showing elevated levels in asthma patients. Studies by Tao et al. (2017, 2019) demonstrated that stearic acid exhibited excellent performance in distinguishing asthma patients from healthy controls. In our analysis, stearic acid also exhibited a positive correlation with the square root of the continuous eosinophil count, aligning with these findings. Muricholic acid has been linked to asthma in obesity models. A study by Barosova et al. (2023) found that obese mice with induced asthma had significantly higher muricholic acid levels compared to obese control mice. In our results, muricholic acid was positively correlated with eosinophilic inflammation, supporting its role in asthma pathophysiology. LysoPE (lysophosphatidylethanolamine) is a member of the lysophospholipids which is a large subclass of phospholipids. In previous studies under inflammatory diseases, researchers found that the signals of lysophospholipids was associated with the Chronic Obstructive Pulmonary Disease (Madapoosi et al. (2022)). In our analysis, leucine is identified as a significant metabolite associated with eosinophilic inflammation. Consistent with prior findings, higher

leucine levels have been reported in asthmatic individuals with elevated exhaled nitric oxide ($\text{FeNO} > 35$), a biomarker indicative of eosinophil-driven inflammation ([Comhair et al. \(2015\)](#)). This suggests that leucine may play a role in asthma pathophysiology, particularly in individuals with active eosinophilic reaction. The interplay between *Bacteroides* and these metabolites is further supported by lipidomic analyses. Notably, differences in lipid profiles among *Bacteroides* are largely driven by variations in plasmalogens, glycerophosphoinositol, and certain sphingolipids. These lipidomic distinctions may influence immune responses and inflammation, providing insight into the mechanisms by which *Bacteroides* species contribute to asthma pathogenesis ([Ryan, Joyce and Clarke \(2023\)](#); [Ryan et al. \(2023\)](#)). Above all, MVOPR demonstrates strong performance in real data analysis, effectively identifying key microbial and metabolic features associated with eosinophilic inflammation in asthma. By selecting the *Bacteroidaceae* family and relevant metabolites, MVOPR aligns well with established biological findings, highlighting its ability to capture meaningful microbiome-metabolome interactions. These results highlight the robustness and reliability of MVOPR in modeling complex multi-omics relationships, making it a powerful tool to uncover biomarkers in asthma.

5. Discussion. The MVOPR model presents a novel approach for multi-omics data integration by using the orthogonal projection framework to handle the correlated predictors, enhancing variable selection. Our model is effective under the unidirectional assumption, aligning well with the inherent biological pathways such as the Central Dogma of Molecular Biology. Traditional methods, such as Lasso-based regression and factor-based models, struggle in multi-omics settings due to the strong within- and inter-modality correlations. MVOPR effectively addresses these challenges by leveraging the unidirectional assumptions between omics layers and employing an orthogonal projection framework to mitigate multicollinearity problems.

Based on the results from simulations and real data analysis, MVOPR showcases superior performance over other competing methods. Unlike factor-based models, which require an approximate factor structure on predictors, MVOPR successfully eliminates redundant dependencies while preserving meaningful signals for variable selection. Even in scenarios where the inter-modality correlation assumption is violated, MVOPR maintains competitive performance, outperforming other methods. This suggests that MVOPR generalizes well beyond ideal conditions, making it a reliable tool for real-world applications.

However, in cases where the model is severely misspecified, such as incorrectly assuming directionality, performance of MVOPR can be affected. For instance, if the true causal direction is from M_1 to M_2 , but a model with reverse direction is fitted (M_2 to M_1), the estimated coefficient matrix \hat{B} may not be well-constructed, leading to poor projections and inaccurate variable selection. To ensure proper unidirectional modeling, a strong understanding of the latent relationships between modalities is crucial. This can be established through biological knowledge, such as the Central Dogma of Molecular Biology, or causal inference that helps to determine the correct directionality before model fitting.

In current analysis, MVOPR operates within a linear regression framework. However, some biological systems are inherently nonlinear and hierarchical, often involving complex interactions between different omics layers. Future extensions of MVOPR could incorporate non-linear model, such as kernel-based methods or deep-learning approaches, to capture these intricate dependencies more effectively.

When applying MVOPR to the CAARS dataset, we successfully identify microbial and metabolic markers linked to eosinophilic inflammation in asthma. Notably, MVOPR select some biomarkers which aligns with prior research. Compared to competing approaches, MVOPR demonstrate higher model stability and lower mean squared error (MSE) in real-data analysis. Traditional methods such as Lasso regression and factor-based models failed to

maintain consistent variable selection across iterations. In contrast, MVOPR achieved higher stability indicators (Jaccard, Otsuka–Ochiai, and Sørensen–Dice coefficients), suggesting improved stability in biomarker identification.

MVOPR represents a advancement in multi-omics variable selection, providing a robust, interpretable, and biologically relevant framework for multi-view data integration. By successfully mitigating within- and inter-modality correlations, MVOPR allows for more precise biomarker discovery, particularly in complex diseases such as asthma. As multi-omics datasets continue to develop, MVOPR offers a powerful and stable method for integrative analysis, providing novel framework for personalized medicine and targeted therapeutic strategies.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. The first author was supported by NSF Grant DMS-??-??????.

The second author was supported in part by NIH Grant ???????????.

APPENDIX A: EXTENSION TO MULTIPLE MODALITIES

A.1. Three modalities case. For three modalities case, we could first transform M_2 and M_3 into their residuals forms E_2 and E_3 . The model (8) would be rewritten as below:

$$Y = M_1\beta_1 + E_2\beta_2 + E_3\beta_3 + M_1(B_{2,1}\beta_2 + B_{3,1}\beta_3) + M_2B_{3,2}\beta_3 + \epsilon_1$$

while M_2 could be further decomposed to $M_1B_{2,1} + E_2$. Therefore,

$$Y = M_1\beta_1 + E_2\beta_2 + E_3\beta_3 + M_1(B_{2,1}\beta_2 + B_{3,1}\beta_3 + B_{2,1}B_{3,2}\beta_3) + E_2B_{3,2}\beta_3 + \epsilon_1$$

Since two nuisance variable $M_1(B_{2,1}, B_{3,1})$ and $E_2B_{3,2}$ are correlated with predictors M_1 and E_2 , we need to project those predictors to the orthogonal subspace. Suppose $M_1(B_{2,1}, B_{3,1}) = U_1\Sigma_1V_1^T$ and $E_2B_{3,2} = U_2\Sigma_2V_2^T$ with rank r_1 and r_2 . Two projection matrices are $P_1 = U_1U_1^T$ and $P_2 = U_2U_2^T$. Based on the projection, we have:

$$Y = (I - P_1)M_1\beta_1 + (I - P_2)E_2\beta_2 + E_3\beta_3 + U_1\gamma_1 + U_2\gamma_2 + \epsilon_1$$

In this form, we will have mutually uncorrelated predictors and nuisance variables in the regression. Below, we make a brief discussion about the assumptions on independence between predictors and nuisance Variables:

1. $E_3 \perp\!\!\!\perp (I - P_1)M_1$ and $E_3 \perp\!\!\!\perp (I - P_2)E_2$ hold since $E_3 \perp\!\!\!\perp M_1$ and $E_3 \perp\!\!\!\perp E_2$.
2. $(I - P_2)\epsilon_2 \perp\!\!\!\perp (I - P_1)M_1$ holds since $E_2 \perp\!\!\!\perp M_1$.
3. $(I - P_1)M_1 \perp\!\!\!\perp U_{1,r}$ and $(I - P_2)E_2 \perp\!\!\!\perp U_{2,r'}$ hold since the projection matrix P_1, P_2 are orthogonal to their complements $(I - P_1), (I - P_2)$.
4. $(I - P_2)E_2 \perp\!\!\!\perp U_{1,r}$ and $E_3 \perp\!\!\!\perp U_{1,r}$ since $E_2 \perp\!\!\!\perp M_1$ and $E_3 \perp\!\!\!\perp M_1$.
5. $(I - P_1)M_1 \perp\!\!\!\perp U_{2,r'}$ and $E_3 \perp\!\!\!\perp U_{2,r'}$ since $M_1 \perp\!\!\!\perp E_2$ and $E_3 \perp\!\!\!\perp E_2$.

A.2. Multiple modalities case. For multi-omics data with k modalities, we need to determine the order for each modality. Based on Central dogma of molecular biology, genomics will generally serve as the first modality which has the ability to influence all the downstream elements. Proteomics or metabolomics may serve as the last modality which can be regularized by upstream elements. Any omics between the first and last modality will serve as

intermediate modality such as transcriptome. After we have the sequential information for multi-omics, we could transform each modality except to their residual forms first.

$$Y = M_1\beta_1 + \sum_{j=2}^k E_j\beta_j + M_1B_1^*\gamma_1 + \sum_{i=2}^{k-1} E_iB_i^*\gamma_i + \epsilon_1$$

where $B_1^* = (B_{2,1}, B_{3,1}, \dots, B_{k,1})$. For any $2 \leq i \leq k-1$, $B_i^* = (B_{i+1,i}, B_{i+2,i}, \dots, B_{k,i})$. To remove the correlation between predictors and nuisance variables, we next project each predictors to the orthogonal subspace. Suppose we have SVD for each nuisance variable:

$$M_1B_1^* = U_1\Sigma_1V_1^T \quad E_iB_i^* = U_i\Sigma_iV_i^T$$

Assume U_1, U_2, \dots, U_{k-1} has rank r_1, r_2, \dots, r_{k-1} . Projection matrix $P_1 = U_1U_1^T, P_2 = U_2U_2^T, \dots, P_{k-1} = U_{k-1}U_{k-1}^T$. Then, the final model for MVOPR will be:

$$Y = P_1^\perp M_1\beta_1 + \sum_{i=2}^{k-1} P_i^\perp E_i\beta_i + E_k\beta_k + \sum_{j=1}^{k-1} U_j\gamma_j^* + \epsilon_1$$

The transformed modalities are mutually uncorrelated to each other in the regression.

APPENDIX: CONNECTION TO FACTOR BASED MODEL

Assume $M_1 \in \mathbb{R}^{n \times p}$ and $M_2 \in \mathbb{R}^{n \times q}$ are two omics data. Let Y denotes the response variable. Suppose Y is associated with M_1 and M_2 by $\beta_1 \in \mathbb{R}^p$ and $\beta_2 \in \mathbb{R}^q$. The interplay between M_1 and M_2 can be captured by a low-rank coefficient matrix B with rank r . E and ϵ_1 are the error matrix and vectors.

$$Y = M_1\beta_1 + M_2\beta_2 + \epsilon_1$$

$$M_2 = M_1B + E$$

Suppose B matrix has a SVD as $B = U_B\Sigma_BV_B^T$. Therefore, M_2 could be expressed based on an approximate factor model. $F = M_1U_B$ and $\Lambda = \Sigma_BV_B^T$. This structure aligns with the scenarios for Integrative Factor Regression and Factor-Adjusted Regularized Regression.

$$\begin{aligned} M_2 &= M_1U_B\Sigma_BV_B^T + E \\ &= F\Lambda + E \end{aligned}$$

Suppose M_1 doesn't follow approximate factor model structure. In Integrative Factor Regression, the factor decomposition for (M_1, M_2) will become (M_1, E) with factors F as nuisance parameter. However, since the factors are given by $F = M_1U_B$, it implies that F is a linear combination of M_1 and is therefore highly correlated with it. When we fit the regression $Y = M_1\beta_1 + E\beta_2 + F\gamma + \epsilon_1$, the true contribution of M_1 will be obscured by the correlated nuisance parameter F . When M_1 has some spiked eigenvalues and could be approximated by factor models, similar problem will still exist. Suppose $M_1 = F_1\Lambda_1 + U_1$, the decomposition of (M_1, M_2) will become (U_1, E) with nuisance parameters (F_1, F) . Since $F = M_1U_B = F_1\Lambda_1U_B + U_1U_B$, F will still be correlated to F_1 and U_1 .

In Factor-Adjusted Regularized Regression, the matrix $M = (M_1, M_2)$ is treated as a whole and decomposed accordingly. Since M follows the decomposition $M = M_1(I, B) + (0, E)$ which does not perfectly align with the model's assumption, the selection of the number of factors will be affected. When B has a rank r that is closed or equal to p , M could be decompose to $F\Lambda + (0, E)$ with p factors. In this case, the transformed M_1 will be nearly zero, as most of its information is absorbed by the factors. Similar issue will happen when $M_1(I, B)$ lacks spiked eigenvalues, leading to difficulties in distinguishing factor structure. This increases the risk of selecting an excessively large number of factors, potentially distorting the factor adjustment process.

REFERENCES

- ABDEL-AZIZ, M. I., NEERINCX, A. H., VIJVERBERG, S. J., KRANEVELD, A. D. and MAITLAND-VAN DER ZEE, A. H. (2020). Omics for the future in asthma. In *Seminars in immunopathology* **42** 111–126. Springer.
- ASLAM, R., HERRLES, L., AOUN, R., PIOSKOWIK, A. and PIETRZYK, A. (2024). The Link between Gut Microbiota Dysbiosis and Childhood Asthma: Insights from a Systematic Review. *Journal of Allergy and Clinical Immunology: Global* 100289.
- BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension.
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.
- BAROSOVA, R., BARANOVICOVA, E., HANUSRICHTEROVA, J. and MOKRA, D. (2023). Metabolomics in Animal Models of Bronchial Asthma and Its Translational Importance for Clinics. *International Journal of Molecular Sciences* **25** 459.
- BOULESTEIX, A.-L., DE BIN, R., JIANG, X. and FUCHS, M. (2017). IPF-LASSO: integrative L1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and mathematical methods in medicine* **2017** 7691937.
- CASTEL, C., ZHAO, Z. and THORESEN, M. (2024). Comparison of the LASSO and Integrative LASSO with Penalty Factors (IPF-LASSO) methods for multi-omics data: Variable selection with Type I error control. *arXiv preprint arXiv:2404.02594*.
- CHEN, K., DONG, H. and CHAN, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* **100** 901–920.
- CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Am. Stat. Assoc.* **107** 1533–1545.
- CHEN, C., WANG, J., PAN, D., WANG, X., XU, Y., YAN, J., WANG, L., YANG, X., YANG, M. and LIU, G.-P. (2023). Applications of multi-omics analysis in human diseases. *MedComm* **4** e315.
- CHU, X., ZHANG, B., KOEKEN, V. A., GUPTA, M. K. and LI, Y. (2021). Multi-omics approaches in immunological research. *Frontiers in Immunology* **12** 668045.
- CHUNG, K. F. (2016). Asthma phenotyping: a necessity for improved therapeutic precision and new targeted therapies. *Journal of internal medicine* **279** 192–204.
- CLARK, C., DAYON, L., MASOODI, M., BOWMAN, G. L. and POPP, J. (2021). An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimer's research & therapy* **13** 1–19.
- COMHAIR, S. A., McDUNN, J., BENNETT, C., FETTIG, J., ERZURUM, S. C. and KALHAN, S. C. (2015). Metabolomic endotype of asthma. *The Journal of Immunology* **195** 643–650.
- DING, D. Y., LI, S., NARASIMHAN, B. and TIBSHIRANI, R. (2022a). Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences* **119** e2202113119.
- DING, D. Y., LI, S., NARASIMHAN, B. and TIBSHIRANI, R. (2022b). Cooperative learning for multiview analysis. *Proc. Natl. Acad. Sci. U. S. A.* **119** e2202113119.
- FAN, J., KE, Y. and WANG, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics* **216** 71–85.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75** 603–680.
- FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75** 531–552.
- FIUZA, B. S. D., DE ANDRADE, C. M., MEIRELLES, P. M., DA SILVA, J. S., DE JESUS SILVA, M., SANTANA, C. V. N., PINHEIRO, G. P., MPAIRWE, H., COOPER, P., BROOKS, C. et al. (2024). Gut microbiome signature and nasal lavage inflammatory markers in young people with asthma. *Journal of Allergy and Clinical Immunology: Global* **3** 100242.
- GARG, M., KARPINSKI, M., MATELSKA, D., MIDDLETON, L., BURREN, O. S., HU, F., WHEELER, E., SMITH, K. R., FABRE, M. A., MITCHELL, J. et al. (2024). Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in the UK Biobank. *Nature Genetics* **56** 1821–1831.
- GAUTAM, Y., JOHANSSON, E. and MERSHA, T. B. (2022). Multi-omics profiling approach to asthma: an evolving paradigm. *Journal of personalized medicine* **12** 66.
- GILLENWATER, L. A., HELMI, S., STENE, E., PRATTE, K. A., ZHUANG, Y., SCHUYLER, R. P., LANGE, L., CASTALDI, P. J., HERSH, C. P., BANAEI-KASHANI, F. et al. (2021). Multi-omics subtyping pipeline for chronic obstructive pulmonary disease. *PloS one* **16** e0255337.
- HUSSEIN, R., ABOU-SHANAB, A. M. and BADR, E. (2024). A multi-omics approach for biomarker discovery in neuroblastoma: a network-based framework. *npj Systems Biology and Applications* **10** 52.

- KLAU, S., JURINOVIC, V., HORNUNG, R., HEROLD, T. and BOULESTEIX, A.-L. (2018). Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC bioinformatics* **19** 1–14.
- KWON, Y., HAN, K., SUH, Y. J. and JUNG, I. (2023). Stability selection for LASSO with weights based on AUC. *Scientific Reports* **13** 5207.
- LI, Q. and LI, L. (2022). Integrative factor regression and its inference for multimodal data analysis. *Journal of the American Statistical Association* **117** 2207–2221.
- MADAPOOSI, S. S., CRUICKSHANK-QUINN, C., OPRON, K., ERB-DOWNWARD, J. R., BEGLEY, L. A., LI, G., BARJAKTAREVIC, I., BARR, R. G., COMELLAS, A. P., COUPER, D. J. et al. (2022). Lung microbiota and metabolites collectively associate with clinical outcomes in milder stage chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine* **206** 427–439.
- MAHDAVINIA, M., FYOLEK, J. P., JIANG, J., THIVALAPILL, N., BILAVER, L. A., WARREN, C., FOX, S., NIMMAGADDA, S. R., NEWMARK, P. J., SHARMA, H. et al. (2023). Gut microbiome is associated with asthma and race in children with food allergy. *Journal of Allergy and Clinical Immunology* **152** 1541–1549.
- MENYHÁRT, O. and GYÓRFFY, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and structural biotechnology journal* **19** 949–960.
- NASIRI, E., BERAHMAND, K., ROSTAMI, M. and DABIRI, M. (2021). A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Computers in Biology and Medicine* **137** 104772.
- OLIVIER, M., ASMIS, R., HAWKINS, G. A., HOWARD, T. D. and COX, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *International journal of molecular sciences* **20** 4781.
- RICHARDS, A. L., ECKHARDT, M. and KROGAN, N. J. (2021). Mass spectrometry-based protein–protein interaction networks for the study of human diseases. *Molecular systems biology* **17** e8792.
- RUFF, W. E., GREILING, T. M. and KRIEGEL, M. A. (2020). Host–microbiota interactions in immune-mediated diseases. *Nature Reviews Microbiology* **18** 521–538.
- RYAN, E., JOYCE, S. A. and CLARKE, D. J. (2023). Membrane lipids from gut microbiome-associated bacteria as structural and signalling molecules. *Microbiology* **169** 001315.
- RYAN, E., GONZALEZ PASTOR, B., GETHINGS, L. A., CLARKE, D. J. and JOYCE, S. A. (2023). Lipidomic analysis reveals differences in *Bacteroides* species driven largely by plasmalogens, glycerophosphoinositols and certain sphingolipids. *Metabolites* **13** 360.
- SZKLARCZYK, D., KIRSCH, R., KOUTROULI, M., NASTOU, K., MEHRYARY, F., HACHILIF, R., GABLE, A. L., FANG, T., DONCHEVA, N. T., PYYSALO, S. et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51** D638–D646.
- TAO, J.-L., WANG, S.-C., TIAN, M., LIANG, H., XIE, T., LIN, L.-L. and DAI, Q.-G. (2017). Metabonomics of syndrome markers in Infantile Bronchial Asthma Episode. *Zhongguo Zhong xi yi jie he za zhi Zhongguo Zhongxiyi Jiehe Zazhi= Chinese Journal of Integrated Traditional and Western Medicine* **37** 319–325.
- TAO, J.-L., CHEN, Y.-Z., DAI, Q.-G., TIAN, M., WANG, S.-C., SHAN, J.-J., JI, J.-J., LIN, L.-L., LI, W.-W. and YUAN, B. (2019). Urine metabolic profiles in paediatric asthma. *Respirology* **24** 572–581.
- UEMATSU, Y., FAN, Y., CHEN, K., LV, J. and LIN, W. (2019). SOFAR: Large-Scale Association Network Learning. *IEEE Transactions on Information Theory* **65** 4924–4939. <https://doi.org/10.1109/TIT.2019.2909889>
- YANG, B., YANG, R., XU, B., FU, J., QU, X., LI, L., DAI, M., TAN, C., CHEN, H. and WANG, X. (2021). miR-155 and miR-146a collectively regulate meningitic *Escherichia coli* infection-mediated neuroinflammatory responses. *Journal of Neuroinflammation* **18** 114.
- ZHANG, W., ZHANG, Y., LI, L., CHEN, R. and SHI, F. (2024). Unraveling heterogeneity and treatment of asthma through integrating multi-omics data. *Frontiers in Allergy* **5** 1496392.
- ZIMMERMANN, P., MESSINA, N., MOHN, W. W., FINLAY, B. B. and CURTIS, N. (2019). Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: a systematic review. *Journal of Allergy and Clinical Immunology* **143** 467–485.