Engineering 7 (2021) 1725–1731

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Microbiology—Article

Expanding the Scope of Multivariate Regression Approaches in Cross-Omics Research

Xiaoxi Hu [a,b], Yue Ma [a,c], Yakun Xu [a,c], Peiyao Zhao [a,d], Jun Wang [a,c,*]

[a] CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[b] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA

ARTICLE INFO

Article history:
Received 8 October 2019
Revised 14 March 2020
Accepted 25 May 2020
Available online 19 May 2021

Keywords:
Multivariate regression methods
Reduced rank regression
Sparsity
Dimensionality reduction
Variable selection

ABSTRACT

Recent technological advancements and developments have led to a dramatic increase in the amount of high-dimensional data and thus have increased the demand for proper and efficient multivariate regression methods. Numerous traditional multivariate approaches such as principal component analysis have been used broadly in various research areas, including investment analysis, image identification, and population genetic structure analysis. However, these common approaches have the limitations of ignoring the correlations between responses and a low variable selection efficiency. Therefore, in this article, we introduce the reduced rank regression method and its extensions, sparse reduced rank regression and subspace assisted regression with row sparsity, which hold potential to meet the above demands and thus improve the interpretability of regression models. We conducted a simulation study to evaluate their performance and compared them with several other variable selection methods. For different application scenarios, we also provide selection suggestions based on predictive ability and variable selection accuracy. Finally, to demonstrate the practical value of these methods in the field of microbiome research, we applied our chosen method to real population-level microbiome data, the results of which validated our method. Our method extensions provide valuable guidelines for future omics research, especially with respect to multivariate regression, and could pave the way for novel discoveries in microbiome and related research fields.

1. Introduction

Biology and medicine have long since entered the age of big data, accelerated by the development of analytical methods and diagnostic techniques in sequencing, proteomic analysis, metabolic analysis, and so forth. For example, next-generation sequencing in the last decade has made sequencing individual genomes and metagenomes feasible, even for single research laboratories, compared with the international efforts and years-long work on the Human Genome Project [1]. Microbiome research, which has also been accelerated by sequencing techniques [2], has revealed the importance of microbial communities in human health and diseases [3], among other fields. All these developments have produced an unprecedentedly large volume of data with high dimensionality, which in turn has promoted general research interest in multivariate regression analysis [4,5]. Multivariate regression aims to model the relationships between a set of responses and a set of features, in contrast to common regression, which usually depicts a one-to-one relationship [6,7]. Response variables (or dependent variables) are the outcomes of an experiment that researchers hope to explain, and predictor variables (or independent variables) are the controlled inputs that may cause the variation in the responses. For example, in a genomics study, the responses in such regressions could be human traits, and the features could be genetic or environmental factors. Therefore, multivariate regression can be applied in all aspects of our daily lives. For example, it is broadly used in economics to investigate the factors that influence stock returns [8]. It is also familiar and common in the field of biology [9]. For example, it is applied in clinical trials to help researchers explain the relationships between drug ingredients and pesticide effects [10]. Recently, in genomics research, including metagenomics research and in combination with metabolomics, proteomics, and so forth, multivariate

* Corresponding author.
E-mail address: junwang@im.ac.cn (J. Wang).

https://doi.org/10.1016/j.eng.2020.05.028

regression has played a significant role in understanding the association and potential causation of important traits [11].

Various attempts have been made to use multivariate methods to address specific challenges. Principal component analysis (PCA) is one of the oldest and best-known eigenvector-based multivariate analysis techniques [12]. It is widely used to find a linear combination of variables that describe the most variance using orthogonal transformation when the number of variables is large. By projecting the data to a lower-dimensional space showing the dominant gradients, PCA can reveal the internal structure of data [13,14]. In practice, principal component regression (PCR) is a linear regression model that uses PCA to estimate the regression coefficient matrix. Canonical correspondence analysis (CCA) is another approach that is often used to explain the relationships between two sets of variables by reducing dimensionality [15]. It aims to find a linear combination that can describe the maximum correlations between predictor variables and responses. Another method often used to find the relations between two matrices is partial least-squares (PLS) regression [16]. It summarizes covariance structure by projecting the response variables and predictor variables into a new space to build a linear regression model. Although the above methods are widely used in studies, three main statistical problems arise. The first problem is that traditional methods often ignore the possible interrelations between the response variables of observational data. The second problem is that some of the approaches do not allow variable selection, which is essential in exploratory experiments when the number of predictor variables is large. Third, some real databases often have large total numbers of variables and small sample sizes, leading to unreliable solutions [17,18].

Based on these considerations, we analyzed a class of new methods (reduced rank regression (RRR) and its extension) that improve the interpretability of regression models by considering the correlations between responses [19,20]. RRR is a data reduction method similar to PCA that creates new variables to summarize a large amount of information in the original data [21]. In particular, it defines a set of linear combinations of predictor variables to best explain the total variance in the response variables, and has many desirable characteristics such as simplicity, computational efficiency, and outstanding predictive performance. When the number of predictor variables is large, the selection of important variables is another issue of interest to researchers. Although some traditional methods are available—such as random forest, which makes predictions by constructing a multitude of decision trees—they are more suitable for cases with a single response variable. Therefore, some RRR-based approaches have adopted the ideas of group selection methods such as the group least absolute shrinkage and selection operator (group lasso) method, which is used for variable selection when determining the group structure among variables. In nutritional epidemiology and genetics, many reports use RRR approaches; yet comprehensive analysis of the properties of different RRR-based approaches, as well as their applicability to real data—especially metagenomic-centered data—remains to be conducted [22–25]. Here, we used simulated data with different dimensionalities to compare the performance of various RRR-based approaches with that of other multivariate regression methods with similar properties, examined their strengths and limitations under different scenarios, and finally applied them to large-scale public metagenomic datasets.

2. Methods

2.1. Description of the approaches tested in this study

In this study, we used the basic multivariate linear regression method as the starting point, and then included a few RRR-based approaches and other multivariate regression approaches for comparison. For each method, the definition and rationale are explained below.

2.1.1. Multivariate linear regression

A multivariate linear regression model is composed of multiple predictor variables $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_p$ and multiple response variables $\mathbf{Y}_1$, $\mathbf{Y}_2$, …, $\mathbf{Y}_q$. Each response variable is represented by a linear regression of the predictors; that is,

$$\mathbf{Y}_j = \sum_{k=1}^{p} \mathbf{X}_k c_{kj} + \epsilon_j,\ j = 1, 2, \cdots, q \tag{1}$$

where c_{kj} is the regression coefficient relating $\mathbf{X}_k$ to $\mathbf{Y}_j$, and ϵ_j is the error term with mean zero.

In addition, this formula can be rewritten with n observations, as follows:

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{2}$$

where $\mathbf{X}$ is an $n \times p$ predictor matrix, $\mathbf{Y}$ is an $n \times q$ response matrix, $\mathbf{C}$ is a $p \times q$ matrix of regression coefficients, and $\mathbf{E}$ is an $n \times q$ error matrix.

We estimate the coefficient matrix $\mathbf{C}$ based on the least squares criterion; that is

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{XC}\|^2 \tag{3}$$

where $\|\cdot\|$ denotes the Frobenius norm.

Using the ordinary least-squares (OLS) method, we obtain the estimate of $\mathbf{C}$, $\widehat{\mathbf{C}}$, which is calculated as following:

$$\widehat{\mathbf{C}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{4}$$

2.1.2. Reduced rank regression

However, the OLS method provides a rough estimate since it ignores the possible interrelationships between response variables and simply performs a separate estimation for each response variable. Therefore, here, we introduce RRR, which constrains the rank of coefficient matrix $\mathbf{C}$. We suppose $\mathbf{C}$ is of lower rank, $r = \mathrm{rank}(\mathbf{C}) \leq \min(p, q)$, and $\mathbf{C}$ can be expressed as a product of two rank r matrices, $\mathbf{C} = \mathbf{BA}^{\mathrm{T}}$, where $\mathbf{B}$ has $p \times r$ dimensions and $\mathbf{A}$ has $q \times r$ dimensions. The multivariate regression model (2) could be rewritten as

$$\mathbf{Y} = \mathbf{XBA}^{\mathrm{T}} + \mathbf{E} \tag{5}$$

In addition, $\mathbf{XB}$ has $n \times r$ dimensions representing a set of r linear combinations of $\mathbf{X}$, which can be interpreted as the latent factors driving the variation in $\mathbf{Y}$. Therefore, RRR helps reduce the dimensionality of the predictor variables and improves computing efficiency.

We can rewrite the optimization function (3); that is

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Y} - \mathbf{XBA}^{\mathrm{T}}\|^2 \tag{6}$$

The set of solutions $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ is given as

$$\widehat{\mathbf{A}} = \mathbf{V} \tag{7}$$

$$\widehat{\mathbf{B}} = \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}\mathbf{V} \tag{8}$$

where $\Sigma_{\mathbf{XX}} = (1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\Sigma_{\mathbf{XY}} = (1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, and $\mathbf{V}$ represents the eigenvectors of $\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}$ corresponding to the eigenvalues, in which $\Sigma_{\mathbf{YX}} = (1/n)\mathbf{Y}^{\mathrm{T}}\mathbf{X}$ [26].

2.1.3. Sparse reduced rank regression (SRRR)

SRRR is an extensive RRR approach that focuses on not only dimensionality reduction, but also variable selection [26,27]. It imposes the sparsity of the coefficient matrix by adding a penalty to the least-squares estimation, and thus has unique properties. Compared with RRR, which uses all predictor variables to build the latent factors, SRRR can be used to select the useful ones from a large number of variables and exclude the redundant ones by introducing a group lasso penalty [28]. Therefore, the optimization formula (6) can be rewritten as follows:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\mathbf{A}^{\mathrm{T}}\|^2+\sum_{i=1}^{p}\lambda_i\|\mathbf{B}^i\| \quad \text{s.t. } \mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I} \tag{9}$$

where λ_i is the penalty parameter. The constraint $\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}$ is applied to satisfy the identifiability conditions, where $\mathbf{I}$ denotes the identity matrix. In addition, if $\|\mathbf{B}^i\|$ is set to zero, the entire i-th row of matrix $\mathbf{B}$ will be zero, and the i-th predictor variable will be inactive.

By using the subgradient method or variational method, the optimization problem can be solved, but defining p penalty parameters (λ_i) by cross-validation (CV) could be time consuming. To reduce the number of tuning parameters, two strategies are usually used [26]:

(1) **Group lasso penalty:** Set all λ_i values equal to λ.

(2) **Adaptive weighting lasso penalty:** Calculate each λ_i based on the original data structure as $\lambda_i=1/\|\widetilde{\mathbf{C}}^i\|^{\gamma}\cdot\lambda$, where $\widetilde{\mathbf{C}}$ is a root-n consistent estimator of $\mathbf{C}$ and γ is a positive integer [29].

2.1.4. Subspace assisted regression with row sparsity (SARRS)

SARRS also focuses on solving the issues of low rankness and sparsity in the coefficient matrix [30]. This new estimation scheme can be used in adaptive sparse reduced rank multivariate regression and achieves the goals of dimensionality reduction and variable selection. Furthermore, compared with SRRR as discussed above, SARRS improves performance when the number of predictor variables exceeds the sample size.

During the process of optimizing the regression with group sparsity, two penalty functions can be used:

(1) **Group lasso penalty:** $\rho(\mathbf{B};\lambda)=\lambda\|\mathbf{B}\|$, where λ is the penalty parameter and $\mathbf{B}$ is the parameter matrix to be optimized.

(2) **Group minimax concave penalty (MCP):** $\rho(\mathbf{B};\lambda)=\lambda\cdot\int_0^{\|\mathbf{B}\|}(1-t/\gamma^{\lambda})_+\mathrm{d}t$, where γ is a positive integer greater than 1 and $(1-t/\gamma^{\lambda})_+$ denotes its positive part, that is $(1-t/\gamma^{\lambda})_+=(1-t/\gamma^{\lambda})\cdot 1_{\{(1-t/\gamma^{\lambda})\geq 0\}}$ [31].

2.1.5. Sparse partial least-squares regression (SPLS)

SPLS method is based on PLS and further encourages sparsity in the multidimensional direction in predictor space; thus, it achieves variable selection [17]. It first selects the predictor variables that have strong correlations with the responses, and then adds additional ones that have strong partial correlations. SPLS employs a different reduced rank structure than SRRR and does not directly focus on the prediction of the response variables, creating a possible weakness in its prediction.

2.1.6. Regularized multivariate regression for identifying master predictors (REmMap)

REmMap method is different from the above methods since it assumes not only that only some of the predictors are correlated with the responses, but also that these predictors may influence only some of the responses [32]. This is reasonable because in real situations, researchers often pay more attention to specific responses than to others. REmMap can fit multivariate regression models with high dimensionality and small sample sizes, and can introduce both overall sparsity and group sparsity into the coefficient matrix to detect master predictors.

2.1.7. Summary

The characteristics of the methods discussed in this article are summarized in Table 1.

2.2. Test of method performance based on simulated data

2.2.1. Simulation setups

To illustrate and compare the performance of SRRR, SARRS, SPLS, REmMap, and some traditional approaches (PCR, group lasso, and random forest), we first introduce a simulation study to generate data and analyze them using the above approaches. We use a similar simulation setup as Chen and Huang [26]. The central idea of the simulation study is to analyze some predictor variables that are correlated with the response variables and some that are uncorrelated. Then, we use these methods to examine which of them can most accurately determine the relationship and achieve good predictive performance.

We generate data with the multivariate linear equation $\mathbf{Y}=\mathbf{X}\mathbf{B}\mathbf{A}^{\mathrm{T}}+\mathbf{E}$. In this model, the $n\times p$ design matrix $\mathbf{X}$ follows a multivariate normal distribution $N(0,\Sigma_{\mathbf{X}})$, where $\Sigma_{\mathbf{X}}$ has diagonal elements of 1 and off-diagonal elements of ρ_x. Matrix $\mathbf{B}$ and matrix $\mathbf{A}$ comprise the coefficient matrix of the model. In $p\times r$ matrix $\mathbf{B}$, the first p_0 rows follow $N(0,1)$, and the remaining $p-p_0$ rows are zero. The $q\times r$ matrix $\mathbf{A}$ is generated from $N(0,1)$. Matrix $\mathbf{E}$ is a random noise matrix defined by $N(0,\sigma^2\Sigma_{\mathbf{E}})$, where σ^2 is the magnitude of the noise and $\Sigma_{\mathbf{E}}$ has diagonal elements of 1 and off-diagonal elements of ρ_e. Then, the $n\times q$ matrix $\mathbf{Y}$ is calculated from the above model.

We generate three sets of data including a training set, validation set, and test set. The training set is used to fit the models based on the various approaches. The validation set is used to tune the parameters inside the models and estimate the noise variance. Lastly, we use the data from the test set to evaluate the performance of the models we built.

To explore the methods' applicability in different situations, we conduct the simulation study using several different cases. First,

Table 1
Methods comparison.

Methods	Data reduction method (low rankness)	Variable selection (sparsity)	Explains interrelation between responses	Features
RRR	√	–	√	Restricts the rank of the regression coefficient matrix
SRRR	√	√	√	Uses a group lasso penalty to allow row-wise sparsity
SARRS	√	√	√	Suitable when the number of predictors exceeds the sample size
SPLS	√	√	√	Uses PLS to impose sparsity
REmMap	–	√	√	Each response has different relevant predictors
PCR	√	–	–	Projects the predictors into a lower-dimensional space
Group lasso	–	√	–	Enables variable selection considering group structure
Random forest	–	√	–	Used to rank the importance of predictor variables

since it is sometimes difficult for researchers to obtain sufficient samples to carry out trials, we want to test the performance of the approaches when applied to both small sample sizes and large sample sizes. Second, we are also interested in the influence of the number of variables. Real data such as microbiological data and genetic data often include high-dimensional predictor variables or response variables. Based on the above considerations, we simulate six cases as follows, where n is the sample size of the data and p and q are the numbers of variables in **X** and **Y**, respectively.

Case 1: Small sample size, $n < p$
- Case 1a: n = 20; p = 100; q = 25
- Case 1b: n = 20; p = 25; q = 25
- Case 1c: n = 20; p = 25; q = 100

Case 2: Large sample size, $n > p$
- Case 2a: n = 200; p = 100; q = 25
- Case 2b: n = 200; p = 25; q = 25
- Case 2c: n = 200; p = 25; q = 100

The simulation procedure and the methods discussed are coded in R using the spls, rrpack, remMap, pls, glmnet, and randomForest packages; the code for the SARRS method was provided by the authors of the respective packages. The computational procedure has been specified and listed in Appendix A.

2.2.2. Evaluation of various methods

In each case, we repeat the simulation procedure 500 times and use the following metrics to measure and compare the performance of the above multivariate regression methods:

(1) **Mean square error (MSE):** The MSE is used to show the predictive accuracy of these methods, and is defined as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mathbf{Y}}_i - \mathbf{Y}_i\right)^2 \tag{10}$$

where $\hat{\mathbf{Y}}_i$ is the predicted value of $\mathbf{Y}_i$.

(2) ***R* squared:** The R^2 index is the proportion of the variance in the response variables that can be explained by the predictor variables. It is often used to describe how well a model fits data. A larger R^2 indicates a better goodness of fit for the model.

(3) **Theil inequality coefficient (TIC):** The TIC is another indicator used to reflect the difference between the fitted values and the true values. It is defined as follows:

$$\mathrm{TIC} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mathbf{Y}}_i - \mathbf{Y}_i\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{Y}}_i^{\,2}} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i^2}} \tag{11}$$

The TIC ranges from zero to one, and a smaller TIC indicates a higher prediction accuracy.

(4) **Sensitivity (TPR) and specificity (SPC):** TPR and SPC are commonly used to evaluate the accuracy of variable selection. TPR is the ability to select the true relevant variables and is calculated as the ratio of the number of correct selections with respect to the total number of correlated input variables. SPC is the ability to select the true irrelevant variables and is calculated as the ratio of the number of correct selections with respect to the total number of uncorrelated input variables. A method that has both high TPR and high SPC means can select the relevant variables accurately.

(5) **Area under the curve (AUC):** The AUC is also used to measure the rate of correctly selecting the true relevant variables [33].

(6) **Overall rating:** The overall rating index is calculated by using the above evaluation metrics. The method with better performance has a higher overall score.

2.3. Application to real population-level microbiome data

To illustrate the practicality of the above methods for real-world problems, we apply them to data from the Belgian Flemish Gut Flora Project (FGFP; discovery cohort: n = 1106) from the work of Falony et al. [34]. This research is aimed at discovering the relationships between microbiota variation and environmental factors such as host features, geography, and medication intake. Sixty-six clinical and questionnaire-based variables are discussed as possible predictors, with 74 microbiome species as responses after selection.

We conducted CV to randomly split the data into a training set and a test set. A model was built to fit the data, and its performance was evaluated by the above metrics. We repeated the process 50 times to examine the stability of variable selection.

3. Results

3.1. Simulation study reveals the distinct properties of each method

In the simulation, we applied SRRR (with the group lasso penalty and adaptive weighting group lasso penalty), SARRS (with the group lasso penalty and group MCP penalty), SPLS, REmMap, PCR, group lasso, and random forest to the different cases and used CV to tune the low-rankness parameters for each approach. Their overall performance is shown in Fig. 1.

The heat map shows that all the methods perform worse in case 1 than in case 2, which is consistent with our prediction. In addition, it is clear that SARRS (with the group MCP penalty) best fits all the cases and, when the sample size becomes larger, SRRR (with

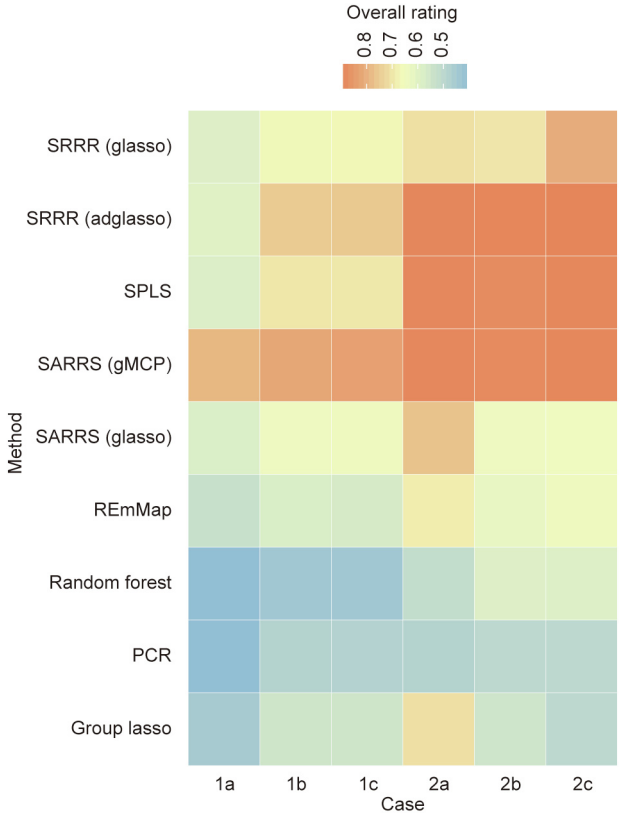

Fig. 1. Overall evaluation of all methods, shown as a heatmap. The x-axis denotes the different cases, and the y-axis denotes the methods. The color of each cell represents the corresponding overall rating. A higher overall rating indicates better performance. glasso: group lasso penalty; adglasso: adaptive weighting group lasso penalty; gMCP: group MCP penalty.

the adaptive weighting group lasso penalty) and SPLS are also applicable and perform equally well.

The performance of each method is measured by the criteria detailed above, and the result for case 1 is shown in Fig. 2.

In case 1a, we have an extremely small sample size, and the number of predictor variables is greater than that with the small size; therefore, most methods do not have good predictive and variable selection performance. Except for PCR, which could not select the relevant variables, the methods' SPCs are approximately 0.75, and their TPRs are approximately 0.55, indicating under selection. However, compared with the traditional methods (PCR, group lasso, and random forest), the new approaches discussed in this article all perform better, especially the SARRS method with the group MCP penalty. This method has the lowest MSE and TIC as well as the highest R^2, SPC, and AUC. This outcome is consistent with the discussion in the methodology section, which specifically noted that SARRS is the most suitable and accurate method when the number of predictor variables is much larger than the sample size.

In cases 1b and 1c, the number of predictor variables and the sample size become closer. We find that all the models fit the simulated data better than in case 1a due to the higher R^2 and the lower TIC. The plot also shows the superiority of SRRR in terms of prediction accuracy, since SPLS and REmMap have larger MSEs than SRRR and SARRS. Furthermore, regarding variable selection, we can see that SARRS (with the group MCP penalty), SRRR (with the adaptive weighting group lasso penalty), and SPLS have better performance, with much higher SPCs, than the others, indicating a balance in selecting the true relevant variables and avoiding overselection.

In case 2, we explore a situation with a large sample size; it is obvious that all methods have better performance than in case 1, as shown in Fig. 3.

We first discuss case 2a, in which there are many more predictor variables than response variables. Regarding predictive performance, the new methods that we are interested in all have an extremely low MSE for the coefficient matrix, even approaching 0, an average TIC of approximately 0.27, and an average R^2 of approximately 0.72, indicating a good model fit. Furthermore, the traditional methods still behave poorly in this aspect. Regarding variable selection, all methods have higher TPRs and AUCs than in case 1. SARRS (with the group MCP penalty), SRRR (with the adaptive weighting group lasso penalty), and SPLS also have extremely high SPCs, indicating that they could accurately select all of the relevant variables and reject all of the irrelevant variables.

In case 2b, we reduce the number of predictor variables to be the same as the number of response variables, thereby increasing the difference between the sample size and the number of variables. Under this circumstance, all the new methods perform well in terms of prediction. However, in variable selection, the outcomes of the methods are polarized. The best method is still SRRR (with the adaptive weighting group lasso penalty), followed by SARRS (with the group MCP penalty) and SPLS, whose selection accuracies approach 1. However, for the other methods, the SPC is generally low, indicating an overselection problem. As case 1c is similar to case 1b, case 2c is also similar to case 2b. However, compared with cases 1b and 1c, the increase in the sample size causes cases 2b and 2c to display much better performance in terms of both prediction accuracy and variable selection.

3.2. Application to real population-level microbiome data

The data characteristics of the case study, in which the sample size ($n = 1106$) is far greater than the number of variables ($p = 66, q = 74$), are consistent with case 2b in the above simula-

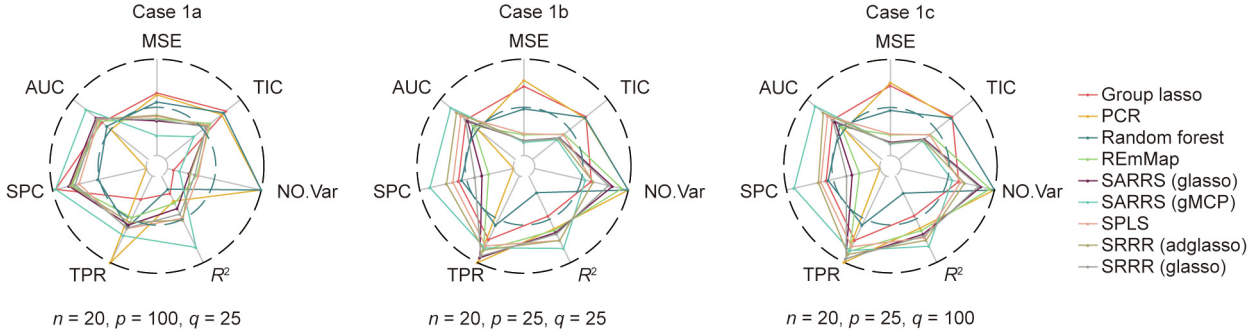

Fig. 2. Performance evaluation of all methods for cases 1a, 1b, and 1c, shown as a radar plot. The center of the circle indicates 0. For R^2, TIC, TPR, SPC, and AUC, the edge of the circle indicates 1. Therefore, if a method has a high R^2, TPR, SPC, and AUC, and a low TIC and MSE of the response matrix, we conclude that this method has good performance. The NO.Var index gives the number of variables selected with each method, where the center of circle indicates 0 and the edge indicates the number of predictor variables.

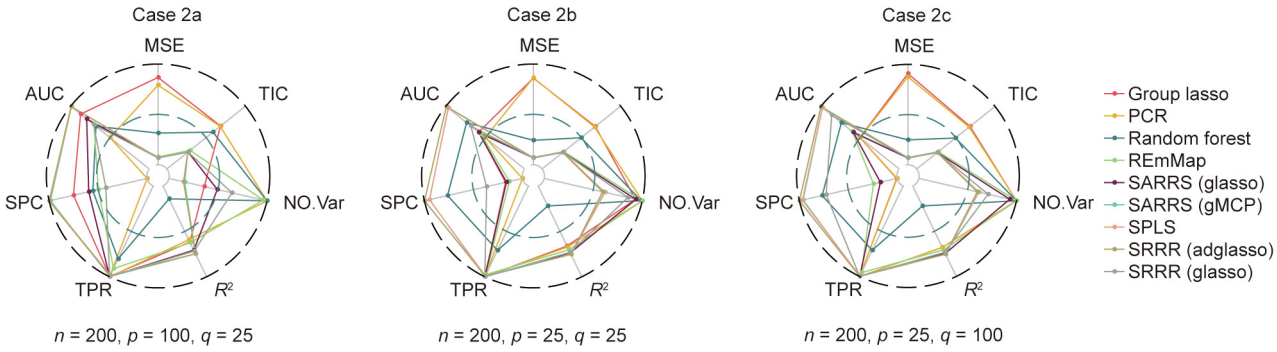

Fig. 3. Performance evaluation of all methods for cases 2a, 2b, and 2c, shown as a radar plot. The center of the circle indicates 0. For R^2, TIC, TPR, SPC, and AUC, the edge of the circle indicates 1. Therefore, if a method has a high R^2, TPR, SPC, and AUC, and a low TIC and MSE of the response matrix, we conclude that this method has good performance. The NO.Var index gives the number of variables selected with each method, where the center of circle indicates 0 and the edge indicates the number of predictor variables.

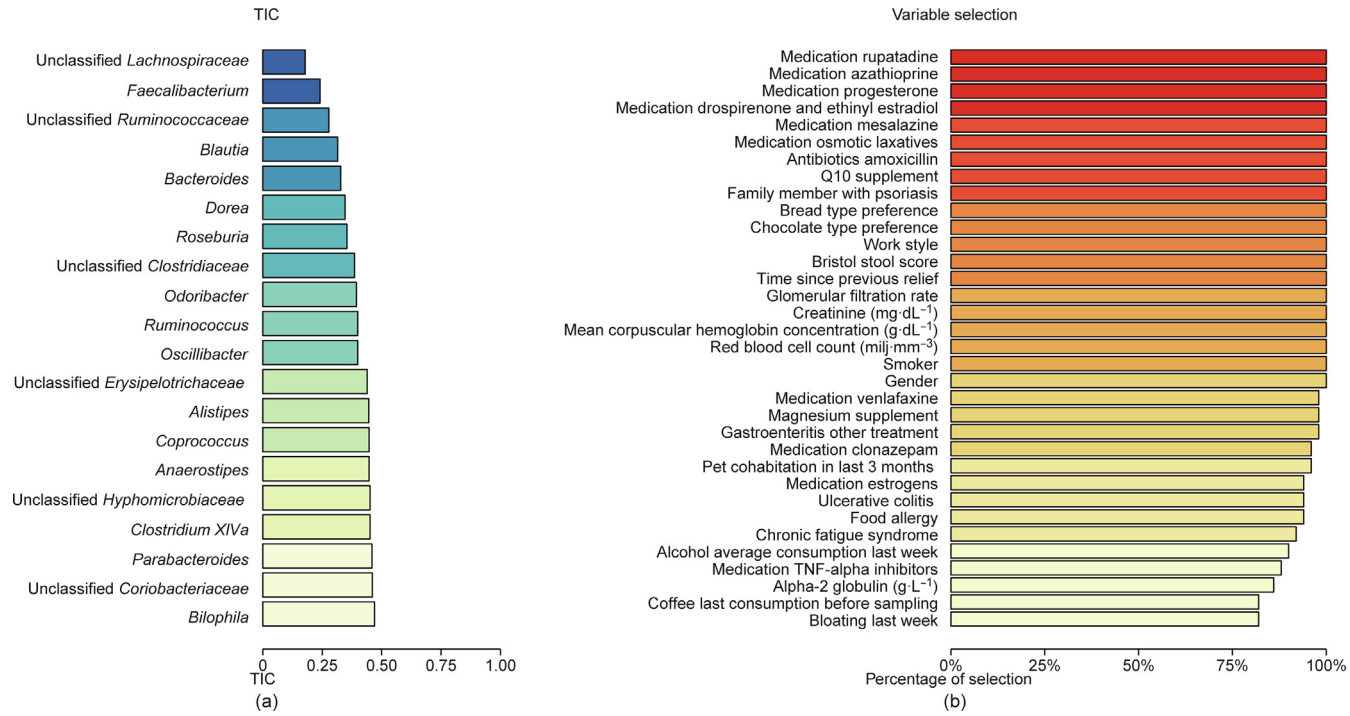

Fig. 4. Performance evaluation of the SRRR (with the adaptive weighting group lasso penalty) for real population-level microbiome data. (a) A bar plot using the TIC index to demonstrate the predictive performance of the SRRR model for each response variable. We display the 20 variables with the lowest TICs in this plot. (b) A plot showing the percentage of selection in 50 cross-validations for each predictor variable.

tion study. Based on the former discussion, we know that the most suitable method to apply for this dataset is SRRR (with the adaptive weighting group lasso penalty). Therefore, we build an SRRR model to analyze the relationships between the environmental indexes and bacterial composition, and discuss the predictive accuracy and variable selection outcomes. The results are shown in Fig. 4.

The average TIC of the model is 0.56, which is higher than the 0.26 for case 2b in the simulation study. However, since we know that the real data are noisier than the simulated data, we conclude that this TIC is acceptable but not convincing in terms of adequate model prediction. However, upon a closer examination of each response variable, we find that the variables with a low TIC are those reported in many previous studies, including the FGFP [34], such as *Faecalibacterium* (with a TIC of 0.24), *Blautia* (0.32), *Bacteroides* (0.33), *Roseburia* (0.35), and *Ruminococcus* (0.40). These are key butyric-producing bacteria that are involved in many diseases when they are at a low abundance, as low butyrate production by the microbiome leads to a higher level of inflammation and metabolic disorders [35,36]. Therefore, these variables were explained well by the predictors selected by SRRR and could be predicted by the coefficient matrix calculated by SRRR.

Finally, we examined the robustness of the variable selection. Since we repeated the CV procedure 50 times, the predictor variables that were selected in more than 80% of the cases are the most meaningful. Fig. 4(b) shows the 34 variables that were selected most frequently, including gender, smoker, red blood cell count (RBC), creatinine, stool score, mean corpuscular hemoglobin concentration (MCHC), and many kinds of medications. This outcome is consistent with the importance of the effects of medications, as discussed in a previous study [37].

4. Conclusion

As the volume and dimensionality of data increase in nearly every field of research, biomedical research will continue to be one of the most important and fast-developing areas. When extracting the maximum value from data, obtaining correct, useful, and meaningful associations between different measures or omics levels poses an important challenge [38]. Here, we examined some representative methods, including extensions of RRR and other multivariate regression methods, and used both simulated data and real microbiome-centered data to address the strengths and limitations of these methods, which might be instructive for future applications to microbiome and other related omics data.

We included a total of nine method–parameter combinations, including seven methods; furthermore, for two of them, two different penalties were used. We simulated data with different sample sizes/dimensionalities and compared the predictor and response variables with/without large differences in dimensionality. From the results when comparing case 1 and case 2, the importance of a large sample size became clear, which could greatly improve the performance of all methods. In particular, compared with case 1, SPLS has much better predictive accuracy in case 2, indicating that it is more applicable when the sample size is large. Under a situation similar to case 1, when the sample size is small, the best method is SARRS with a group MCP penalty, which has outstanding performance in terms of both prediction and variable selection. When the sample size is large, as in case 2, SARRS (with the group MCP penalty), SRRR (with the adaptive weighting group lasso penalty), and SPLS all perform very well; upon closer examination, SRRR (with the adaptive weighting group lasso penalty) performs slightly better than the other two methods.

We used this information to select the best method for a real scenario: namely, the published FGFP data, for which microbiome data and the environmental factors identified in the study are available. With roughly similar dimensionalities for the two types of variables, we decided to use SRRR with the adaptive weighting group lasso penalty. First, we identified the bacterial groups that are best explained by the environmental changes. The identified bacterial groups confirm previous assertions that the butyrate-producing bacteria are of great importance in human health and may serve as

a link to those environmental factors. In addition, since environmental factors were considered to be predictors (i.e., features selected to be associated with bacterial groups), we also managed to replicate the most important features in the published study, again demonstrating the reliability and robustness of the selected method.

Researchers should carefully choose a proper penalty when fitting models. For example, in the SRRR method, when $n > p$, the adaptive weighting group lasso penalty improves both the prediction accuracy and variable selection. When $n < p$, the adaptive weighting results in a lower TPR but higher SPC than those for the unweighted group lasso penalty. This could be explained by the fact that when we introduced weighting to the SRRR calculation procedure, the variables that were filtered out earlier had larger weights for their penalty terms and were no longer included in the model. Therefore, SRRR with an unweighted penalty will select more variables and lead to a high SPC.

In conclusion, we examined the applicability of several multivariate regression approaches and tested their performance under different omics scenarios, which in reality may differ vastly in their sample sizes and dimensionalities. Based on this, we were able to recommend the best method. Admittedly, our preliminary analysis could not be further expanded at this stage to incorporate the phylogenetic information between different measures (e.g., species) in many omics data, since this would require *a priori* information regarding the connectedness and similarity between those measures. We also used a renowned microbiome dataset to show that our method of choice can largely recapitulate the findings obtained by single-variate analysis and improve the consideration between variables and combined feature selection. These findings will facilitate the choice of methods in future, larger scale omics research, including microbiome-centered studies.

Acknowledgments

This project was supported by the National Key Research and Development Program of China (2018YFC2000500), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29020000), the National Natural Science Foundation of China (31771481 and 91857101), and the Key Research Program of the Chinese Academy of Sciences (KFZD-SW-219), "China Microbiome Initiative."

Compliance with ethics guidelines

Xiaoxi Hu, Yue Ma, Yakun Xu, Peiyao Zhao, and Jun Wang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2020.05.028.

References

[1] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291(5507):1304–51.
[2] Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010;11(1):31–46.
[3] Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature 2012;489(7415):220–30.
[4] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLOS Comput Biol 2015;11(5):e1004226.
[5] Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 2016;26(5):330–5.
[6] Izenman AJ. Modern multivariate statistical techniques: regression, classification, and manifold learning. New York: Springer-Verlag; 2008.
[7] Kharratzadeh M, Coates M. Sparse multivariate factor regression. In: Proceedings of the 2016 IEEE Statistical Signal Processing Workshop; 2016 Jun 26–29; Palma de Mallorca, Spain; 2016.
[8] Binder JJ. On the use of the multivariate regression model in event studies. J Account Res 1985;23(1):370.
[9] Kim KA, Jung IH, Park SH, Ahn YT, Huh CS, Kim DH. Comparative analysis of the gut microbiota in people with different levels of ginsenoside Rb1 degradation to compound K. PLoS ONE 2013;8(4):e62409.
[10] Peng Y, Li SN, Pei X, Hao K. The multivariate regression statistics strategy to investigate content-effect correlation of multiple components in traditional Chinese medicine based on a partial least squares method. Molecules 2018;23(3):545.
[11] Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25(6):968–76.
[12] Smith L. A tutorial on principal components analysis. Technical report. Dunedin: University of Otago; 2002 Feb. Report No.: OUCS-2002-12.
[13] Gleason PM, Boushey CJ, Harris JE, Zoellner J. Publishing nutrition research: a review of multivariate techniques—part 3: data reduction methods. J Acad Nutr Diet 2015;115(7):1072–82.
[14] Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. Mol Ecol 2016;25(5):1032–57.
[15] ter Braak CJF. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 1986;67(5):1167–79.
[16] Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1–17.
[17] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol 2010;72(1):3–25.
[18] Bunea F, She Y, Wegkamp MH. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. Ann Stat 2012;40(5):2359–88.
[19] Mukherjee A. Topics on reduced rank methods for multivariate regression [dissertation]. Ann Arbor: University of Michigan; 2013.
[20] D'Ambra L, Amenta P, Gallo M. Dimensionality reduction methods. Metodoloski Zveski 2005;2(1):115–23.
[21] Izenman AJ. Reduced-rank regression for the multivariate linear model. J Multivariate Analysis 1975;5(2):248–64.
[22] Hoffmann K, Schulze MB, Schienkiewitz A, Nothlings U, Boeing H. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. Am J Epidemiol 2004;159(10):935–44.
[23] Cespedes EM, Hu FB. Dietary patterns: from nutritional epidemiologic analysis to national guidelines. Am J Clin Nutr 2015;101(5):899–900.
[24] Vounou M, Nichols TE, Montana G; Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. NeuroImage 2010;53(3):1147–59.
[25] Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage 2012;60(1):700–16.
[26] Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am Stat Assoc 2012;107(500):1533–45.
[27] Chen L, Huang JZ. Sparse reduced-rank regression with covariance estimation. Stat Comput 2016;26(1–2):461–70.
[28] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Series B Stat Methodol 2006;68(1):49–67.
[29] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101(476):1418–29.
[30] Ma Z, Sun T. Adaptive sparse reduced-rank regression. 2014. arxiv:1403.1922.
[31] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Stat Sci 2012;27(4):481–99.
[32] Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann Appl Stat 2010;4(1):53–77.
[33] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
[34] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science 2016;352(6285):560–4.
[35] Wan Y, Wang F, Yuan J, Li J, Jiang D, Zhang J, et al. Effects of dietary fat on gut microbiota and faecal metabolites, and their relationship with cardiometabolic risk factors: a 6-month randomised controlled-feeding trial. Gut 2019;68(8):1417–29.
[36] Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vila AV, Võsa U, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat Genet 2019;51(4):600–5.
[37] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018;555(7698):623–8.
[38] Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational metaomics for microbial community studies. Mol Syst Biol 2013;9(1):666.