BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Sparse reduced-rank regression for integrating omics data

Haileab Hilafu[1*], Sandra E. Safo[2] and Lillian Haine[2]

*Correspondence: hhilafu@utk.edu
[1]Department of Business Analytics and Statistics, University of Tennessee, 37996 Knoxville, TN, USA
Full list of author information is available at the end of the article

Abstract

Background: The problem of assessing associations between multiple omics data including genomics and metabolomics data to identify biomarkers potentially predictive of complex diseases has garnered considerable research interest nowadays. A popular epidemiology approach is to consider an association of each of the predictors with each of the response using a univariate linear regression model, and to select predictors that meet a priori specified significance level. Although this approach is simple and intuitive, it tends to require larger sample size which is costly. It also assumes variables for each data type are independent, and thus ignores correlations that exist between variables both within each data type and across the data types.

Results: We consider a multivariate linear regression model that relates multiple predictors with multiple responses, and to identify multiple relevant predictors that are simultaneously associated with the responses. We assume the coefficient matrix of the responses on the predictors is both row-sparse and of low-rank, and propose a group Dantzig type formulation to estimate the coefficient matrix.

Conclusion: Extensive simulations demonstrate the competitive performance of our proposed method when compared to existing methods in terms of estimation, prediction, and variable selection. We use the proposed method to integrate genomics and metabolomics data to identify genetic variants that are potentially predictive of atherosclerosis cardiovascular disease (ASCVD) beyond well-established risk factors. Our analysis shows some genetic variants that increase prediction of ASCVD beyond some well-established factors of ASCVD, and also suggest a potential utility of the identified genetic variants in explaining possible association between certain metabolites and ASCVD.

Keywords: Integrative analysis, Multi-view data, Reduced rank regression, High dimensional data

Background

Advances in technologies and data collection processes have resulted in multiple high dimensional data types being measured on the same subjects. For instance, in biomedical research, these data types include genomics, metabolomics, proteomics, and transcriptomics. While each of these data types provide a different snapshot of the

underlying biological system, it is being increasingly recognized that combining these data types can reveal complex relationships that may not be unraveled from individual analyses. For instance, the integration of genomic and metabolomic/proteomic data can provide valuable insight into key genomic loci that influence human plasma levels associated with complex diseases [1]. This is of great interest because genomic studies including genome wide association studies (GWAS) have revealed that the majority of disease-causing single nucleotide polymorphisms (SNPs) lie in noncoding regions of the genome [2], making it difficult to know their functional implications. While individual genomic variants identified through GWAS can be tested experimentally, this approach is complicated by the modest effects of the identified variants and the fact that we may not know the specific gene driving the genomic association [3]. Integration of genomics data with other omics data can therefore enable us to identify genomic variants that could generate hypotheses for the genomic architecture of the underlying disease, or could identify variants that have the potential to improve clinical factors. Since the metabolome is considered as the end product of all genetic, epigenetic, and environment activities [4, 5], linking metabolite levels in human blood samples with genomics data can help shed light on complex disease-causing genomic variants. Additionally, tying genomic variants to metabolite levels can identify metabolites that can be used as biomarkers or potential targets for drug discovery [1]. A review of studies that combine genomics and metabolomics data can be found in [3]. In a recent study [1], genomics data were linked with protein levels known to be associated with cardiovascular disease (CVD) and many new gene locus-protein associations were unraveled, providing new insight into CVD risk pathophysiology [1].

However, integrating genomics and metabolomics data, for instance, to identify important disease-associated biomarkers is complicated by the high-dimensional nature of each omics data. The popular epidemiological approach to relate genomics with metabolomics or proteomics data is to consider an association of each of the genetic variant with each of the metabolites using a univariate linear regression model, and to select genetic variants that meet a priori specified significance level [1, 6, 7]. Specifically, the following linear regression model is considered:

$$\mathbf{y}_j = \mathbf{x}_i c_i + \mathbf{e}_j \tag{1}$$

where $\mathbf{y}_j, j = 1, \ldots, q$ is a $n \times 1$ vector of metabolic features or protein expression levels for n subjects, $\mathbf{x}_i, i = 1, \ldots, p$ is a $n \times 1$ vector of SNPs for n subjects, c_i is the unknown coefficient for the ith SNP, e_j is a vector of random noise, q denotes the number of responses (metabolic features or protein levels), and p denotes the number of predictors (SNPs). The above approach is limiting because larger sample size is usually required to identify associated biomarkers, which is costly. Furthermore, it assumes variables for each data type are independent, and thus ignores correlations that exist between variables both within each data type and across the data types. Additionally, genomic studies show that most genetic variants have modest effect on complex diseases, suggesting the need for methods that model multiple SNPs simultaneously in association studies. These limitations lead us to consider the following multivariate linear regression model

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

where $\mathbf{Y}$ is $n \times q$ matrix containing all the responses (e.g., all metabolites), $\mathbf{X}$ is $n \times p$ matrix of predictors (e.g., SNPs), $\mathbf{C}$ is a $p \times q$ matrix of unknown coefficients, and $\mathbf{E}$ is $n \times q$ matrix of random noise. Our goal is then to estimate the matrix of unknowns $\mathbf{C}$, and to identify multiple relevant predictors that are simultaneously associated with the responses, and which potentially could predict complex diseases. From a statistical point of view, the discovery of biomarkers is best cast as a variable selection problem, where "variable" refers to the genetic loci or metabolites. Variable selection in omics data is complicated by the high-dimensional nature of each of the omics data.

When we use (2) to model genomics and metabolomics/proteomics data, with a large number of responses and predictors, the number of unknown parameters that need to be estimated in $\mathbf{C}$, i.e. pq, can quickly exceed the sample size n. To overcome this problem, researchers have considered two important types of structural assumptions that induce lower-intrinsic-dimension on $\mathbf{C}$. The first is *low-rankness* where the rank of $\mathbf{C}$ is assumed to be much smaller than its matrix dimension of $min(p, q)$. That is, it is assumed that $rank(\mathbf{C}) = r < min(p, q)$. Then, counting the parameters in the singular value decomposition of $\mathbf{C}$, we observe that only $r(p + q - r)$ free parameters need to be estimated, which can be substantially lower than pq for low values of r. This structure is referred to as the *reduced-rank regression* (RRR) and has been widely used in variety of applications [8, 9]. Reduced-rank estimation is often obtained by introducing penalties that are proportional to the eigenvalues of the coefficient matrix or its rank, see, for example, [10–15].

The second structural assumption is the so called *sparsity* where only a small subset, s, out of the p predictors are assumed to contribute to the variation of the responses. Removing the ith predictor from model (2) is equivalent to setting the ith row in $\mathbf{C}$ to zero. Vectorizing both sides of model (2) yields a univariate response regression model. Thus, one can view the rows of $\mathbf{C}$ as groups of coefficients in the transformed model and set them to zero by any group selection method developed for univariate response regression models. Thus, the effective number of parameters is sq, which is smaller than the unrestricted pq, but may be higher than $r(p + q - r)$, especially if the rank of $\mathbf{C}$ is low. Proposals that use penalties that induce *row-sparsity* include, among others, [16–21]. For either structure, researchers have sought to understand how a given statistical estimation depends on the model parameters and on how to achieve optimal estimation without the knowledge of the rank r or the sparsity level s. In this article, we propose a new method that induces both *row-sparsity* and *low-rankness*, and leads to meaningful dimension reduction and variable selection.

Motivating data: an atherosclerosis disease study

We motivate our work using data from the Emory/Georgia Tech Predictive Health Institute (PHI) study. The PHI is longitudinal study of healthy employees from Emory University and Georgia Tech that began in 2005 with the aim of collecting health factors that could be used to optimize and maintain health rather than treating disease. With this in mind, we seek to identify genomic risk factors that are correlated with motabolite and which could be used for predicting 10-year risk of ASCVD. ASCVD is a chronic inflammatory disease as well as a disorder of lipid metabolism [22]. It is a complex disease of many risk factors including genetic risk factors. Many genetic studies have been conducted to identify genetic variants and genes that many be implicated in ASCVD [23]. However, the functional implications of these SNPs and genes are not well-understood.

Linking metabolomic data with genomic data can help shed light on genetic loci influencing ASCVD. Additionally, tying genetic loci to metabolomics can identify metabolites that can be used as biomarkers for ASCVD [1]. In light of the above, we seek to use metabolites to guide selection of SNPs that may be associated with ASCVD, and to also explore the potential utility of these genetic variants in explaining possible association between certain metabolites, and ASCVD risk.

Main contributions

This paper makes two main contributions. First, we propose a new computationally efficient convex formulation to estimate the coefficient matrix in (2) that takes advantage of the potential presence of *low-rankness* and *sparsity*. The proposed convex formulation is computationally efficient, and can be solved using readily available solvers. It is also shown to yield competitive numerical performances (in estimation, prediction, and variable selection) under a variety of model parameter settings when compared with state-of-the-art methods in the literature. Specifically, we observe that the superior results of the proposed method, in estimation and variable selection, are more pronounced when the number of responses and predictors are much higher than the sample size. This is encouraging to us since our motivating problem, and many integrative genomics analysis problems, fall under this regime. Second, atherosclerosis cardiovascular disease is a major health-economic burden in USA, and beyond, and the problem of identifying other non-traditional risk factors beyond well-established factors remains an important scientific problem and active research area. We aim to contribute to the body of knowledge in this field through the use of innovative statistical methods such as the ones proposed here. We therefore present careful analyses of data from healthy adults with low- vs moderate- to high-risk for developing atherosclerosis cardiovascular disease in the future using genomics, metabolomics, clinical, and demographic data, permitting us to identify genetic variants that increase atherosclerosis cardiovascular disease risk beyond established risk factors. Additionally, we explore the potential use of these genetic variants in explaining possible association of certain metabolites with atherosclerosis cardiovascular disease.

Method

Reduced-rank regression

Let $\{\mathbf{x}_i^\top, \mathbf{y}_i^\top\}_{i=1}^n$ denote an available n i.i.d. samples. In the sequel, we denote the predictor and response data matrices by $\mathbf{X}$ and $\mathbf{Y}$, respectively. Suppose that $\mathbf{C}$ is of lower rank, $r = \text{rank}(\mathbf{C}) < \min(p, q)$, and that we have a $q \times r$ orthonormal matrix $\mathbf{A}$ whose columns span the right singular subspace of $\mathbf{C}$. That is, we have a $q \times r$ orthonormal matrix $\mathbf{A}$ such that for some $p \times r$ matrix $\mathbf{B}$, $\mathbf{C} = \mathbf{B}\mathbf{A}^\top$. Then, post-multiplying both sides by $\mathbf{A}$, we can re-write model (2) as

$$\mathbf{Y}\mathbf{A} = \mathbf{X}\mathbf{B} + \mathbf{E}\mathbf{A}, \tag{3}$$

where $\mathbf{X}\mathbf{B}$ is of reduced dimension with only r components that can be interpreted as unobservable latent factors that drive the variation in the responses. This re-parametrization also indicates that $\mathbf{A}$ spans the right singular subspace of $\mathbf{Y}$. Therefore, if we had such a matrix $\mathbf{A}$, we would fit a lower dimensional regression of $\mathbf{Y}\mathbf{A}$ on $\mathbf{X}$ to obtain an estimate of $\mathbf{B}$, and use it to obtain an estimate for $\mathbf{C}$. In the literature, model (3)

is referred to as the *reduced-rank regression* model. Since the responses are modeled by $r(r < q)$ common latent factors, we achieve dimensionality reduction of the predictors and expect that this modeling exercise takes the correlations among the q responses into account. There are a number of approaches of obtaining such a matrix $\mathbf{A}$. For example, [17] and [24] consider the SVD, $\mathbf{C} = \mathbf{UDV}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are $p \times r$ and $q \times r$ matrices with orthonormal columns, respectively, and $\mathbf{D}$ is a $r \times r$ nonnegative diagonal matrix. [18] set $\mathbf{A}$ to be the r-dimensional right singular subspace of $\mathbf{Y}$. We exploit both these approaches in this paper. On the other hand, [20], propose to seek such an $\mathbf{A}$ that leads to the best approximation of the signal matrix $\mathbf{XC}$. That is, they seek a $q \times r$ orthogonal matrix $\mathbf{A}$ such that $\mathbf{XBA}^\top$ is the best rank r approximation of the signal matrix $\mathbf{XC}$.

The latent factors $\mathbf{XB}$ (r components) are lower dimensional than the original predictors. However, they are linear combinations of all the p original predictors. Therefore, while model (3) achieves dimension reduction of the predictors, it does not lead to variable selection. Recall that a predictor is unimportant in predicting the responses via model (2) if the corresponding row in $\mathbf{C}$ is zero. Thus, to achieve variable selection, row-sparse estimate of $\mathbf{C}$ is desirable. The following two key facts facilitate this using model (3): (*i*) if $\mathbf{C}$ has at most s non-zero rows, so does $\mathbf{B} = \mathbf{CA}$; (*ii*) the non-zero rows in $\mathbf{B}$ and $\mathbf{C}$ are the same. Thus, row-sparse estimate of $\mathbf{C}$ can be obtained by seeking row-sparse estimate of $\mathbf{B}$.

Our approach to sparse reduced-rank regression

Suppose that we have a matrix $\tilde{\mathbf{A}}$ as described above (we discuss how to obtain such a matrix in the implementation section below). We use the following optimization problem for a row-sparse estimation in reduced-rank regression (3):

$$\widehat{\mathbf{B}} = \min_{\mathbf{B}} \sum_{j=1}^{p} \left\| \mathbf{b}_j \right\|_2 \qquad \text{subject to} \qquad \max_{1 \le j \le p} \left\| \mathbf{X}_j^\top \left(\mathbf{Y}\tilde{\mathbf{A}} - \mathbf{XB} \right) \right\|_1 \le \tau, \tag{4}$$

where $\tau > 0$ is a tuning parameter that controls the sparsity level in $\widehat{\mathbf{B}}$. Large values of τ yield less sparse estimates and smaller values of τ yield more sparse estimates. The formulation in (4) can be thought of as a generalization of the dantzig selector [25] to a multivariate reduced-rank regression setting, with $\mathbf{Y}\tilde{\mathbf{A}}$ as the response and $\mathbf{X}$ as the predictor. This formulation, which yields row-sparse estimates, is desirable for the following reasons: (*i*) the solution to the optimization problem is unique up to a $r \times r$ orthogonal matrix; (*ii*) the set of important predictors obtained by solving the optimization problem is uniquely determined, where the important predictors are those that correspond to nonzero rows of the solution. Consequently, the solutions to (4) are determined up to an orthogonal transformation. Nonetheless, different solutions correspond to selection of the same set of predictors, hence the name *coordinate-independent* sparse reduced-rank regression (CISRRR).

Implementation

We focus on the reduced-rank version of our proposal via (4), as this is the more useful method for the high-dimensional setting. The proposed method can be viewed as a two stage estimation approach. In the first stage, we seek an estimate of the right singular subspace of $\mathbf{C}$, or the right singular subspace of $\mathbf{Y}$, say $\widehat{\mathbf{A}}$. In the second stage, we solve (4)

to obtain row-sparse estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$. The corresponding row-sparse estimate of $\mathbf{C}$ is then obtained as: $\widehat{\mathbf{C}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top$.

Algorithm 1 Coordinate-Independent Sparse Reduced-Rank Regression

Input: $\mathbf{X}, \mathbf{Y}, \tau, \tilde{\mathbf{A}}$ (initial $\widehat{\mathbf{A}}$).
Output: $\widehat{\mathbf{A}}(\tau), \widehat{\mathbf{B}}(\tau)$
Iterate Until Convergence:

1. Given current $\widehat{\mathbf{A}}$, to update $\widehat{\mathbf{B}}$, solve

$$\widehat{\mathbf{B}} = \min_{\mathbf{B}} \sum_{j=1}^{p} \|\mathbf{b}_j\|_2 \qquad \text{subject to} \qquad \max_{1\le j\le p} \left\| \mathbf{X}_j^\top \left(\mathbf{Y}\widehat{\mathbf{A}} - \mathbf{X}\mathbf{B}\right) \right\|_1 \le \tau. \tag{5}$$

2. With the current $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{A}}$, compute $\widehat{\mathbf{C}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^\top$ and update $\widehat{\mathbf{A}}$ by the r right singular vectors of $\mathbf{X}\widehat{\mathbf{C}}$.

The values of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ at convergence serve as the estimates $\widehat{\mathbf{A}}(\tau)$ and $\widehat{\mathbf{B}}(\tau)$.

To understand the motivation behind the updating in step 2, note that if we knew $\mathbf{C}$, we can think of the right singular subspace of $\mathbf{XC}$ as estimating the right singular subspace of $\mathbf{Y}$. Therefore, this step can be thought of as encouraging the updating $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{XC}$ approximates $\mathbf{Y}$. Next, we discuss two approaches to obtain the initial estimate $\tilde{\mathbf{A}}$.

1 We can use the sample estimate for the right singular subspace of $\mathbf{Y}$ as the initial estimate of $\mathbf{A}$. Let $\mathbf{Y} = \mathbf{UDV}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{D}$ is a diagonal $r \times r$ matrix, and $\mathbf{V}$ is $q \times r$. Set $\tilde{\mathbf{A}} = \mathbf{V}$ as the initial estimate.

2 Alternatively, we can use some regularization to obtain the initial estimate. Suppose that $\widehat{\mathbf{C}}^{\text{OLS}}$ is the nonsparse OLS estimate of $\mathbf{C}$. That is, $\widehat{\mathbf{C}}^{\text{OLS}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$. Perform a singular-value decomposition (SVD): $\widehat{\mathbf{C}}^{\text{OLS}} = \mathbf{UDV}^\top$. Then, use the first r columns of $\mathbf{V}$ as the estimate $\tilde{\mathbf{A}}$. In the event that $n < p$, we use the ridge-type estimator instead of the OLS estimator, i.e. $\widehat{\mathbf{C}}^{\text{Ridge}} = \left(\mathbf{X}^\top\mathbf{X} + \mathbf{I} * \sqrt{\log p/n}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$. For computational expediency, we also avoid having to invert the $p \times p$ matrix $\left(\mathbf{X}^\top\mathbf{X} + \mathbf{I} * \sqrt{\log p/n}\right)$. Instead, we use the tricks given in [26] and invert an $n \times n$ matrix. More specifically, let $\mathbf{X} = \mathbf{UDV}^\top$ represent the SVD of $\mathbf{X}$; that is, $\mathbf{V}$ is $p \times n$ with orthogonal columns, $\mathbf{U}$ is $n \times n$ with orthogonal columns, and $\mathbf{D}$ a diagonal matrix with elements $d_1 \ge d_2 \ge \cdots \ge d_n \ge 0$. [26] show that a computationally efficient estimate of the ridge estimator can be obtained as: $\widehat{\mathbf{C}}^{\text{Ridge}} = \mathbf{V}\left(\mathbf{R}^\top\mathbf{R} + \mathbf{I} * \sqrt{\log p/n}\right)^{-1}\mathbf{R}^\top\mathbf{Y}$, where the matrix $\mathbf{R} = \mathbf{UD}$ is $n \times n$, leading to a much less computationally expensive estimate.

All our empirical results are based on the first approach. However, we have conducted simulation studies to assess the performances using the second approach as well. The simulations showed that the two approaches yield comparable results. Once an initial $\tilde{\mathbf{A}}$ is specified, to solve the optimization problem in (5), we used CVX, a MATLAB package for specifying and solving convex optimization problems [27, 28]. Matlab codes that

implement this algorithm are provided as online supplementary material (see Additional file 2).

Tuning parameter selection

The tuning parameter τ in (5) controls the level of sparsity in $\widehat{\mathbf{B}}$, and hence in $\widehat{\mathbf{C}}$, and needs to be selected adaptively from the data. Notice that when $\tau > \max_{1\le j\le p}\left\|\mathbf{X}_j^\top\mathbf{Y}\widehat{\mathbf{A}}\right\|_1$, the optimization problem (5) yields a trivial solution, giving us an upper bound for τ. Therefore, we choose the optimal τ from the range $\left(0, \max_{1\le j\le p}\left\|\mathbf{X}_j^\top\mathbf{Y}\widehat{\mathbf{A}}\right\|_1\right)$ using K-fold cross validation. Specifically, for a given τ, we randomly split the available data $\{(\mathbf{Y},\mathbf{X})\}$ into K roughly equal-sized non-overlapping groups of observations, which we denote by $\{(\mathbf{Y},\mathbf{X})\}^k$, $k = 1,\ldots,K$. Let $\{(\mathbf{Y},\mathbf{X})\}^{-k}$ be the data matrix leaving out $\{(\mathbf{Y},\mathbf{X})\}^k$. With a given τ, we apply the proposed method on $\{(\mathbf{Y},\mathbf{X})\}^{-k}$ to obtain an estimate of the coefficient matrix $\widehat{\mathbf{C}}^k(\tau)$. Then, we compute the K-fold mean squared prediction error (MSPE) as follows:

$$\text{MSPE}(\tau) = \frac{1}{K}\sum_{k=1}^{K}\frac{\left\|\mathbf{X}^k\widehat{\mathbf{C}}^k(\tau) - \mathbf{Y}^k\right\|_{\text{F}}^2}{n_k q} \tag{6}$$

where n_k is the number of observations in $\{(\mathbf{Y},\mathbf{X})\}^k$. We do this for a number of τ values in the range (20 in our empirical studies) and select the optimal tuning parameter τ as:

$$\tau_{\text{opt}} = \min_{\tau}\left\{\text{MSPE}(\tau)\right\}. \tag{7}$$

Rank (r) selection: In our discussions so far, we have treated the rank (r) as known. In practice, it needs to be estimated from the data. There are many methods proposed to estimate r in the literature, see for instance [15] and [29], and the references therein. In our empirical studies, we again use cross-validation to estimate the rank. More specifically, we choose an estimate $\widehat{r}$ such that

$$\widehat{r} = \min_{r}\left\{\text{MSPE}\left(\tau_{\text{opt}}, r\right)\right\},$$

where $\text{MSPE}(\tau_{\text{opt}}, r)$ is the MSPE for a given r and the optimal tuning parameter as selected by (7). The value of r in practice is small, often times between 1 and 3. In our empirical simulations, we try values r in $\{1,\cdots,10\}$. If the optimal value of r obtained by the cross-validation approach is close to 10, one could expand the range. Our empirical results (Table 1 in Additional file 1) show that this approach works well.

Results

Simulation studies

In this section, we assess the finite sample performance of the proposed coordinate-independent sparse estimation method for reduced-rank regression (CISRRR). We assess, and compare, estimation, prediction and variable selection performances. Estimation and prediction performances are evaluated using, respectively,

$$\Delta = \left\|\mathbf{C} - \widehat{\mathbf{C}}\right\|_{\text{F}}^2/(pq) \quad \text{and} \quad \text{MSPE} = \left\|\mathbf{Y}_{\text{t}} - \widehat{\mathbf{Y}}_{\text{t}}\right\|_{\text{F}}^2/(n_t q) \tag{8}$$

where $\widehat{\mathbf{Y}}_{\text{t}} = \mathbf{X}_{\text{t}}\widehat{\mathbf{C}}$, n_{t} is the test set sample size, and $\|.\|_{\text{F}}$ represents the Frobenius norm. Variable selection performance is evaluated using true positive rate (TPR), the ratio of truly important variables that the method selects as important, and false positive rate (FPR), the ratio of unimportant predictors that the method selects as important. TPR

values close to one and FPR values close to zero indicate a better variable selection performance. In all our simulation settings, to minimize the effect of parameter tuning, we generate a large test set (with 10,000 observations), a strategy which was also employed by [16] and [18]. For our method, the tuning parameter candidates are taken at a grid of 20 equally spaced values between $(0,\tau_{\max}]$, where $\tau_{\max}$ is as defined in the tuning parameter selection section. Also, unless otherwise specified, the tuning parameter is chosen by 5-fold cross-validation as described in tuning parameter selection section. We repeat the simulation process 50 times and report results in the form of boxplots of the corresponding values.

We compare our method, i.e. Algorithm 1, with a number of state-of-the-art competing sparse estimation methods for multivariate linear regression. The first competing method we consider is the signal extraction approach for sparse multivariate response regression (SiER) by [20]. This method exploits the reduced rank structure by assuming there exist matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{C}=\mathbf{B}\mathbf{A}^{\top}$, and seeks such $\mathbf{A}$ and $\mathbf{B}$ that lead to the best rank r approximation of the signal matrix $\mathbf{XC}$. We use the `SiER` package in `R` to implement this method [20], using the "`cv.SiER`" function with 5-fold cross-validation to select the tuning parameters. The second competing method we consider is the regularized multivariate regression for identifying master predictors (remMap) by [30]. This method does not assume the reduced rank structure and solves a penalized least squares problem with both row-wise and element-wise sparsity imposed on the coefficient matrix. We use the `remMap` package in `R` to implement this method [30]. The third competing method is the subspace assisted regression with row-sparsity (SARRS) method by [18], which was extended to yield row and column sparse estimators in [19]. SARRS is carried out by Algorithm 1 in [18]. The fourth competing method is the sparse partial least squares (SPLS) method due to [31] which identifies sparse latent components by maximizing the covariance between them and the responses with sparsity inducing penalty imposed. We implement SPLS using the `spls` package in `R`. We use the function "`cv.spls`" with 5-fold cross-validation to select the tuning parameters, with the number of components K selected from $\{1,\cdots,10\}$ and the thresholding parameter η selected from $\{0.1,\cdots,0.9\}$. We note again that, for a fair comparison, we use the tuning parameter selection methods presented in the respective papers.

We compare the methods under different model parameter settings as characterized by the covariance matrix of the predictors, as well as different rank values, and *signal-to-noise* ratios. We adopt simulation settings from [16], which were also adopted by [18]. The rows of the design matrix $\mathbf{X}$ are i.i.d. random vectors sampled from a multivariate Gaussian distribution with zero mean vector and covariance matrix Σ, with $\Sigma_{ij}=\rho^{|i-j|}$. The coefficient matrix $\mathbf{C}\in\mathbb{R}^{p\times q}$ has the form

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_1\\ \mathbf{0}\end{bmatrix}=\begin{bmatrix}b\mathbf{B}_0\mathbf{B}_1\\ \mathbf{0}\end{bmatrix},$$

with $b>0$, $\mathbf{B}_0\in\mathbb{R}^{s\times r}$ and $\mathbf{B}_1\in\mathbb{R}^{r\times q}$, where all entries in $\mathbf{B}_0$ and $\mathbf{B}_1$ are i.i.d. random numbers from $N(0,1)$. Large value of b correspond to a large signal-to-noise ratio. We consider the following four cases.

1. $n>p=q$: $n=100, p=25, q=25, s=15, r=5, b=0.2, 0.4$, $\rho=0.1, 0.5, 0.9$.
2. $q<n<p$: $n=30, p=100, q=10, s=15, r=2, b=0.5, 1$, $\rho=0.1, 0.5, 0.9$.
3. $n<p=q$: $n=30, p=100, q=100, s=15, r=2, b=0.5, 1$, $\rho=0.1, 0.5, 0.9$.

4. $n < p < q$: $n = 30, p = 100, q = 1000, s = 15, r = 5, b = 0.5, 1, \rho = 0.1, 0.5, 0.9$.

We conduct additional simulations (refer to Figs. 1 and 2 in Additional file 1) where the error term matrix has entries from a non-Gaussian distribution. More specifically, we consider two additional noise distributions: $\mathbf{E}_{ij} \sim \sqrt{3/5}t_5$, and $\mathbf{E}_{ij} \sim$ 3U[-1,1], where "3U[-1,1]" refers to the sum of three uniform [-1,1] random variables, and t_ν stands for a t-distribution with ν degrees of freedom.

Simulation results

Figures 1, 2 and 3 report simulation results for the scenario where the noise matrix $\mathbf{E} \in \mathbb{R}^{n\times q}$ has i.i.d. $N(0,1)$ entries. Figure 1 reports the results for case 1 ($n = 100, p = q = 25$), from which, we make the following observations. In terms of estimation and prediction performances, remMap outperforms all the other methods, especially when $\rho = 0.1$ and 0.5. This is not surprising, as remMap does not impose the low-rankness assumption and the sample size dominates both p and q. However, our method (CISRRR) yields comparable results with remMap, and even outperforms it when $\rho = 0.9$ when remMap appears to struggle - again perhaps because it is marginal model. We see that all the methods also perform reasonably well and comparably $\rho = 0.1$. In terms of variable selection, our method yields the best results in all settings. It yields TPR values that are significantly higher than the TPR values of the other existing methods, and comparable (often better) FPR values. In fact, we see that the TPR values of our method are consistently around 1. We see that both SARRS and SiER struggle in this case, especially in terms of TPR. Figure

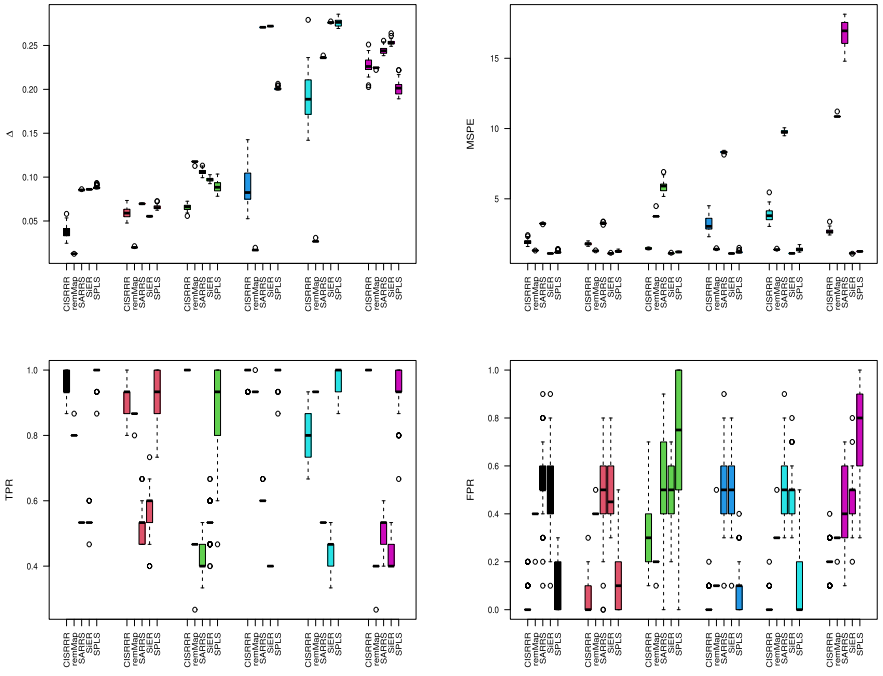

Fig. 1 Simulation results for Gaussian errors under case 1. Reported results are 50 independent replications. Δ assesses estimation performance. MSPE is mean squared prediction error; TPR is true positive rate; FPR is false positive rate. Black for $\rho = 0.1, b = 0.2$; Red for $\rho = 0.1, b = 0.4$; Green for $\rho = 0.5, b = 0.2$; Blue for $\rho = 0.5, b = 0.4$; Cyan for $\rho = 0.9, b = 0.2$; Purple for $\rho = 0.9, b = 0.4$

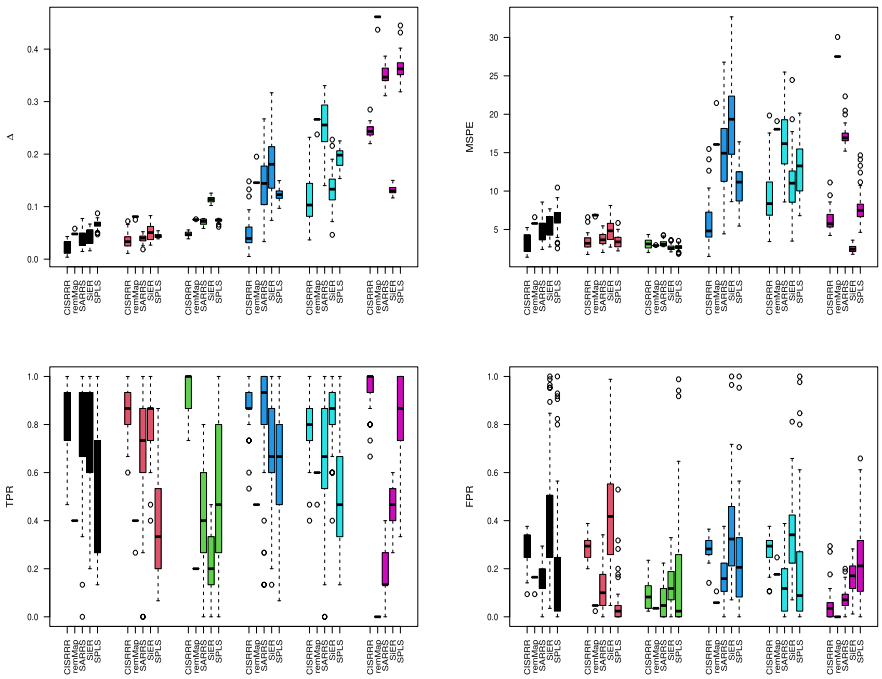

Fig. 2 Simulation results for Gaussian errors under case 3. Reported results are 50 independent replications. Δ assesses estimation performance. MSPE is mean squared prediction error; TPR is true positive rate; FPR is false positive rate. Black for $\rho = 0.1, b = 0.5$; Red for $\rho = 0.1, b = 1$; Green for $\rho = 0.5, b = 0.5$; Blue for $\rho = 0.5, b = 1$; Cyan for $\rho = 0.9, b = 0.5$; Purple for $\rho = 0.9, b = 1$

2 reports the results for case 3 ($n = 30, p = 100, q = 100$). Here, we see that CISRRR outperforms all the other methods in estimation, prediction, and TPR performances. Even though the performances in TPR appear comparable to that of SARRS and SiER, it is seen that our method yields more stable results - with less variability. With respect to FPR, our method is inferior to remMap and SPLS, and to a lesser degree SARRS, especially in the $\rho = 0.1$ and $\rho = 0.5$ cases. However, both remMap and SPLS yield very inferior TPR values, an indication that they both yield very sparse estimates. The performances for case 2 (not reported to save space) are similar to the performances for case 3. Figure 3 reports the results for case 4 ($n = 30, p = 100, q = 1000$). For this case, remMap did not produce results, since q is large and it does not make the reduced-rank assumption. Here again, we see that our method outperforms all the other methods in estimation, prediction, and TPR, in all the settings. The advantage of our method is, in fact, more pronounced in this case as it yields superior results consistently. Especially when the correlation structure among the predictors is higher ($\rho = 0.9$ vs $\rho = 0.5$ vs. $\rho = 0.1$), we see that the other methods performances deteriorate but our method continues to perform well. These observations are true for estimation, prediction and TPR. However, our method pays some price in terms of FPR as it does not outperform any of the other methods in terms of FPR. As we indicated earlier, SPLS yields more sparse results and thus have lower TPR and lower FPR values. Overall, our method is shown to yield competitive, and often times superior, results.

In all the simulation settings that we consider, the estimation and prediction performances of the methods are better when the correlation structure among the predictors

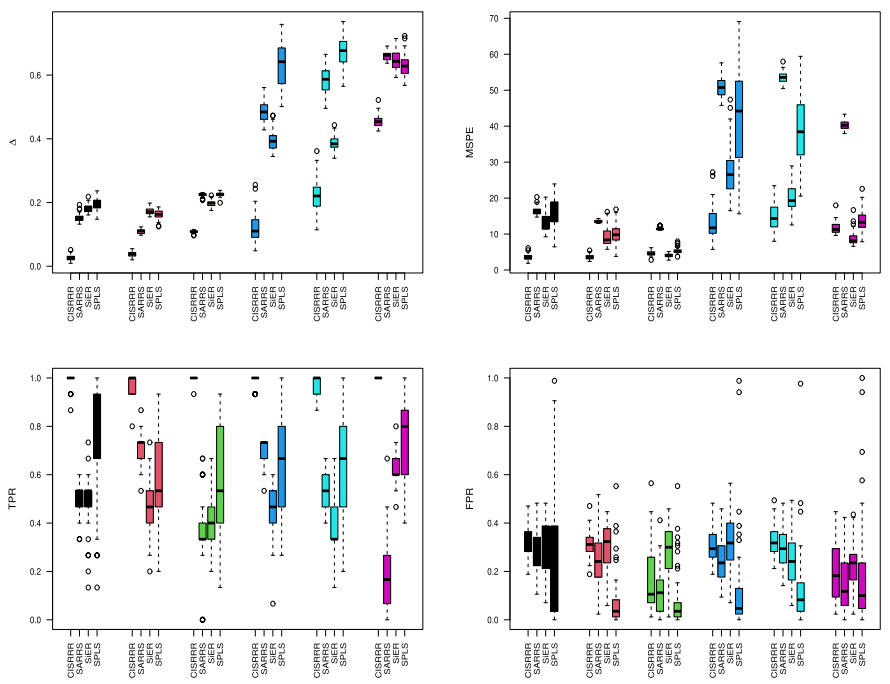

Fig. 3 Simulation results for Gaussian errors under case 4. Reported results are 50 independent replications. Δ assesses estimation performance. MSPE is mean squared prediction error; TPR is true positive rate; FPR is false positive rate. Black for $\rho = 0.1, b = 0.5$; Red for $\rho = 0.1, b = 1$; Green for $\rho = 0.5, b = 0.5$; Blue for $\rho = 0.5, b = 1$; Cyan for $\rho = 0.9, b = 0.5$; Purple for $\rho = 0.9, b = 1$

is weaker ($\rho = 0.1$ vs $\rho = 0.5$). Nevertheless, we see that our method continues to perform well, relative to the other methods. Overall, our methods yields competitive (often superior) results in estimation, prediction and TPR. In terms of FPR, it yields comparable performance, sometimes inferior to the best performing method. Furthermore, we observe that remMap performs well for the large n setting, and struggles when q is large since it does not induce the reduced-rank structure, as well as when ρ is large since it is a marginal model. Finally, it appears that, like remMap, SPLS performs well for the large n setting (case 1), but it struggles in variable selection when either p or q is large (cases 2, 3 and 4).

Real data analysis: the atherosclerosis disease study

Study goals and design: We apply the proposed method for simultaneous analysis of genetic (single nucleotide polymorphisms, SNPs) and metabolomics data. Data were obtained from the Emory University and Georgia Tech Predictive Health Institute (PHI) study. Our goal in this section is to identify relevant SNPs, and corresponding genes, that are simultaneously associated with metabolites, and which can be used to predict 10-year risk for atherosclerosis diseases (ASCVD). Specifically, we seek to use metabolites to guide selection of SNPs that may be associated with ASCVD, and to explore an indirect relationship between metabolites and ASCVD through the genetic variants.

SNP and metabolomics quality control and filtering: We obtained genetic and metabolomics data from the Emory PHI study. Several studies point to the association between biomarkers of inflammation, and the risk of CVD [32, 33]. As such, recent effort

has focused on identifying biomarkers of inflammation and characterizing their effect on CVD [34]. We therefore focused on inflammation-related genes, and the SNPs within these genes. Specifically, we pulled all SNPs in our data that were in gene regions found in the inflammation pathway from NCBI dbSNP; there were 262,157 SNPs. Of note, these SNPs may or may not be associated with ASCVD. For SNP quality control, please refer to the flow chart given in Fig. 3 in the web supplementary material (Additional file 1). We assumed an additive genetic model in which the genotypes were coded to count the number of minor alleles so "0" for both homozygous major, "1" for heterozygotes (1 major, 1 minor), and "2" for risk-allele homozygotes (minor alleles). We treated the genetic data as continuous.

We obtained metabolomics data on 6,010 m/z features. We removed features having more than 50% zeros and coefficient of variation $\geq$ 20%. This resulted in 272 m/z features for the analyses. Because of the skewed distributions of most metabolomic levels, we log2 transformed each feature. We standardized each feature to have mean zero and unit variance.

Application of the proposed and competing methods: We had matching genetic and metabolomics data on 121 subjects. We applied our method to the metabolomics ($\mathbf{Y}_{121\times272}$) and genetic ($\mathbf{X}_{121\times1988}$) data to identify subset of SNPs that are simultaneously associated with metabolomics data, and which potentially can predict ASCVD risk. We obtained 50 bootstrap training and testing datasets by sampling the dataset with replacement. Out of bag samples (samples in the original data but not in the bootstrap training sets) were considered as bootstrap testing sets. For each bootstrap dataset, we estimated the rank r of the coefficient matrix for the multivariate regression model using 5-fold cross-validation as described in the tuning parameter selection section. The rank of the coefficient matrix was estimated to be $\widehat{r} = 2$. Next, we apply the methods to the training data to obtain the estimated coefficient matrix, $\widehat{\mathbf{C}}$, which yields the predicted values for the test metabolite samples, $\widehat{\mathbf{Y}}_{\text{test}}$. We use $\mathbf{Y}_{\text{test}}$ and $\widehat{\mathbf{Y}}_{\text{test}}$ to compute the test MSPE, as given in (8). We record the number of non-zero rows of the estimated coefficient matrix (selected SNPs) for each method and for each bootstrap testing datasets. The averages are reported in Table 1.

Since our goal is to identify potential novel genetic variants that are linked with m/z features, and which could predict ASCVD risk, we considered the following analyses after identifying potential SNPs. The 10-year ASCVD risk score was dichotomized into low- vs moderate- to high-risk ASCVD. Specifically, ASCVD risk score $\geq$ 5.0% was considered moderate to high-risk, and ASCVD risk score $<$ 5% was considered low-risk [35, 36]. A weighted genetic-risk score (GRS) that utilizes the SNPs appearing at least 90% (>45

Table 1 Average MSPEs and average number of selected SNPS (non-zero rows) for the competing methods from 50 independent bootstrap replications

	CISRRR	**SARRS**	**SIER**	**SPLS**	**remMap**
MSPE	1.013	1.038	1.035	1.032	****
# Selected SNPs	72.820	146.00	274.4	278.420	****
SARRS	8				
SiER	6	5			
SPLS	2	3	7		

The bottom half of the table presents the number of overlapping SNPs in the respective top 15 SNPs for the methods

times) out of the 50 bootstrap datasets was calculated by multiplying the logarithm of the odds ratio for that particular SNP by 0, 1, or 2 depending on the number of risk alleles carried by each subject using the whole data. The log odds ratio from each bootstrap was estimated for each of the 15 SNPs, and a weighted mean of the estimates was used in calculating the genetic risk score. We further considered whether including the GRS to a model that used well-established risk factors (age and/or sex) improved predictive ability.

We explored causal association between metabolites and ASCVD risk, after adjusting for age and sex, by utilizing the GRS as an instrument in the causal pathway using Mendelian randomization [37]. Specifically, if the genetic risk score is statistically significantly associated with certain metabolites and is associated with ASCVD risk, then this would provide supportive evidence for a potential causal effect of that metabolite on ASCVD. In the Mendelian randomization analyses, we adopted the two-stage process outlined in [37]. In the first stage, we considered a linear regression model of each m/z features on the GRS controlled for age and sex, and we obtained fitted values for m/z features that showed significant association with the GRS (Benjamini-Hochberg False Discovery Rate [38] p-value $<= .0004$). In the second stage, these fitted values were included in a logistic regression model of ASCVD risk on the fitted values, and the effect on ASCVD risk was assessed after adjusting for age and sex.

Results

SNPs identified using the proposed and existing methods: We investigate the SNPs identified by the proposed method, and the corresponding genes, with respect to their potential effect on ASCVD. Table 1 reports the average MSPEs and average model size (number of selected SNPs). Our proposed method identified 15 SNPs that were selected in at least 90% of the 50 independent replications (see web supplementary material). Of these selected 15 SNPS, 8 appear in the top 15 most selected SNPs by SARRS, 6 appear in the top 15 most selected SNPs by SiER, and 2 appear in the top 15 most selected SNPs by SPLS.

Table 2 in the web supplementary file shows the least squares means for the 15 SNPs identified by our method - the predicted population means for 10-year ASCVD risk, after adjusting for age and sex. For instance, the *rs1286264* SNP located on chromosome 14 is an intron variant found in the protein coding gene Ribosomal Protein S6 Kinase A5 (RPS6KA5). From our data, individuals with two risk alleles of this polymorphism are more likely to have lower adjusted 10-year ASCVD risk score least squares means compared with individuals with normal alleles or 1 risk and 1 normal alleles. A weighted genetic risk score developed using the 15 SNPs was significantly associated with ASCVD risk, after adjustment for age and sex (p-value < 0.001) (Table 2 below). The predictive ability of the GRS + traditional risk factors was assessed with the area under the curve (AUC) from a receiver operating characteristics curve. We note from Table 2 that including the GRS improved AUC in both models. The difference between GRS + traditional risk factors model and only the traditional risk factor model were both statistically significant (p-value = 0.0134 for Model 1; p-value = 0.0104 for Model 2). Our findings suggest that a unit increase in the GRS increased the risk for ASCVD with an OR of 2.348 (95% CI: 1.599, 4.132) after controlling for age and sex. Intriguingly, when we dichotomize the GRS, with a high risk score >75th percentile, and low risk score $\leq$ 75th percentile, we find that the odds for ASCVD risk in the high genetic risk group was about 5 times the

Table 2 Predicting ASCVD risk with GRS (genetic risk score) and traditional risk factors

	Model 1			**Model 2**		
	OR	*p*-value	CI	OR	*p*-value	CI
GRS	2.354	<.001	(1.466, 3.779)	2.348	<.001	(1.460, 3.776)
age	1.155	<.001	(1.065, 1.254)	1.158	<.001	(1.067, 1.256)
sex (M vs F)				0.771	0.6954	(0.242, 2.459)
AUC	**0.8428 vs 0.743 (Ref)**			**0.8441 vs 0.743 (Ref)**		

Ref represents reference ROC; model with no risk score

odds for ASCVD risk in the low genetic risk group, after controlling for age and sex (OR = 5.076, *p*-value .0005, 95% CI: 1.77, 14.49).

Mendelian randomization exploratory analysis: Here, we sought to tie the genetic risk score to m/z features and to explore causal association of m/z features to ASCVD risk. The genetic risk score served as an instrument to estimate the effect of the m/z features on ASCVD risk, after adjustment for age and sex. In this exploratory analysis, our findings suggested that the Dehydroalanine compound (C02218) is a possible risk factor for ASCVD. Specifically, the amino acid, Dehydroalanine, belonging to the Cysteine (Cys) and methionine (Met) metabolism pathway increased ASCVD risk with an odds ratio of 23.204 (95% CI: 4.106, 131.124) per SD increase in the log2 predicted plasma levels, after controlling for age and sex. Some research studies have documented the negative health consequences including elevated risk for cardiovascular diseases with high-intakes of both Met and Cys [39–41]. Our findings suggest an indirect association between Dehydroalanine amino acid and ASCVD risk through these genetic variants.

Conclusion

We sought out to develop a method for identifying potential genetic variants that are associated with metabolites and have a predictive value beyond some established risk factors. We framed this as a two stage analysis: in the first stage we identified SNPs that were associated with the metabolites using dimension reduction techniques proposed in this article. In the second stage we used the selected SNPs as instruments to explore the cause-effect association of the selected metabolites with ASCVD. To handle the large number of SNPs and metabolites, we used sparse reduced-rank regression and proposed a new estimation method for the coefficient matrix using a group Dantzig type formulation. The proposed formulation is convex and can be solved using readily available solvers, such as the CVX toolbox in MATLAB. We carried out extensive simulation study to assess its finite sample performance, and compared it to other existing state-of-the-art methods.

In the second stage, we developed a genetic risk score comprised of 15 genetic variants and we assessed whether including the risk score in a model with well-established risk factors (age/sex) improved predictive ability. Our findings suggested a potential utility of the genetic risk score as it improved predictive ability. We used Mendelian randomization to explore association of a metabolite with ASCVD through the genetic risk score. Our analysis revealed a possible indirect association between the Dehydroalanine amino acid and ASCVD using the genetic risk score in the causal pathway, suggesting a potential role of Dehydroalanine on ASCVD risk through the genetic risk score. We note that our findings are just exploratory, as we lacked an independent data set to validate our results.

Nevertheless, our results add to the literature on possible genetic variants that could be used in addition to established risk factors to improve prediction of atherosclerosis cardiovascular disease.

In our proposed method, we only focus on sparsity on **B**; this amounts to selection of predictors. One would be able to induce sparsity on **A** to select responses. We have not pursued this idea in this approach, but we believe it will be interesting to do so in the future.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03606-2.

Additional file 1: Additional simulations and data analysis. We present additional simulations to assess the robustness of the proposed method when the error terms are non-Gaussian. In addition, in this file we report a flow-chart for the SNP selection process, and a table (Table 2) with the least squares means for the 15 SNPs identified by our method - the predicted population means for 10-year ASCVD risk, after adjusting for age and sex.

Additional file 2: Software. Matlab codes that implements the proposed CISERRR algorithm.

Abbreviations

ASCVD: Atherosclerosis cardiovascular diseases; SNPs: Single nucleotide polymorphisms; AUC: Area under receiver operating characteristic curves; GRS: Genetic risk score; MSPE: Mean squared prediction error; TPR: True positive rate; FPR: False positive rate

Acknowledgments

We are grateful to the Emory Predictive Health Institute for providing us with the genomics, metabolomics, and clinical data used in the real data analysis. We are also very grateful to the two reviewers and the editor for their constructive comments that greatly improved the paper.

Authors' contributions

HH conceived, deduced the algorithm, and designed the simulation study. SS and LH performed the real data analysis. HH and SS wrote the manuscript. All authors read and approved the final manuscript.

Funding

This research is partly supported by NIH grant 1KL2TR00249201 and T32HL129956. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Matlab codes for implementing the method are available as Additional File 2. The gene expression, metabolomics, and clinical data were provided to us by the Emory Predictive Health Institute, and therefore cannot be made publicly available. As part of their data agreement consent, "Emory University and the Predictive Health Initiative retain ownership rights to all provided data regardless of the purpose or outcome of any subsequent publications or collateral works". (http://predictivehealth.emory.edu/documents/CHDWB_EmoryUniversity_DataUseRequestForm.pdf) The data may be requested from https://redcap.emory.edu/surveys/?s=7PYMFLHYTL.

Ethics approval and consent to participate

The data access subcommittee of the Emory Predictive Health Institute granted access to use the data. See: http://predictivehealth.emory.edu/documents/CHDWB_EmoryUniversity_DataUseRequestForm.pdf.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Business Analytics and Statistics, University of Tennessee, 37996 Knoxville, TN, USA. [2]Division of Biostatistics, University of Minnesota, 55455 Minneapolis, MN, USA.

Received: 11 January 2020 Accepted: 16 June 2020
Published online: 03 July 2020

References

1. Benson MD, Yang Q, Ngo D, Zhu Y, Shen D, Farrell LA, Sinha S, Keyes MJ, Vasan RS, Larson MG, Smith JG, Wang TJ, Gerszten RE. Genetic architecture of the cardiovascular risk proteome. Circulation. 2018;137:1158–72.

2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5. https://doi.org/10.1126/science.1222794, http://science.sciencemag.org/content/early/2012/09/04/science.1222794.full.pdf.
3. Shah SH, Newgard CB. Integrated metabolomics and genomics: Systems approaches to biomarkers and mechanisms of cardiovascular disease. Circ Cardiovasc Genet. 2015;8(2):410–9. https://doi.org/10.1161/circgenetics.114.000223.
4. Griffin JL. The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? Philos Trans R Soc B Biol Sci. 2006;361(1465):147–61. https://doi.org/10.1098/rstb.2005.1734.
5. Krumsiek J, Bartel J, Theis FJ. Computational approaches for systems metabolomics. Curr Opin Biotechnol. 2016;39:198–206. https://doi.org/10.1016/j.copbio.2016.04.009, Systems biology • Nanobiotechnology.
6. Kettunen J, Tukiainen T, Sarin A-P, Ortega-Alonso A, Tikkanen E, Lyytikäinen L-P, Kangas AJ, Soininen P, Würtz P, Silander K, Dick DM, Rose RJ, Savolainen MJ, Viikari J, Kähönen M, Lehtimäki T, Pietiläinen KH, Inouye M, McCarthy MI, Jula A, Eriksson J, Raitakari OT, Salomaa V, Kaprio J, Järvelin M-R, Peltonen L, Perola M, Freimer NB, Ala-Korpela M, Palotie A, Ripatti S. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet. 2012;44(3):269–76. https://doi.org/10.1038/ng.1073.
7. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, Mewes H-W, Wichmann H-E, Weinberger KM, Adamski J, Illig T, Suhre K. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008;4(11):1000282. https://doi.org/10.1371/journal.pgen.1000282.
8. Reinsel GC, Velu RP. Multivariate reduced-rank regression: theory and applications; 1998.
9. Izenman A. Modern multivariate statistical techniques: Regression, classification, and manifold learning; 2008. Springer Texts in Statistics.
10. Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. J R Stat Soc Ser B. 2007;69:329–46.
11. Candes E, Plan Y. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. IEEE Trans Inform Theory. 2010;57:2342–59.
12. Negahban YS, Wainwright MJ. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann Stat. 2011;39:1069–97.
13. Rohde A, Tsybakov A. Estimation of high-dimensional low-rank matrices. Ann Stat. 2011;39:887–930.
14. Chen K, Dong H, Chan K. Reduced rank regression via adaptive nuclear norm penalization. Biometrika. 2013;100:901–20.
15. Bunea F, She Y, Wegkamp MH. Optimal selection of reduced rank estimators of high-dimensional matrices. Ann Stat. 2011;39:1282–309.
16. Bunea F, She Y, Wegkamp MH. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. Ann Stat. 2012;40:2359–88.
17. Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am Stat Assoc. 2012;107(500):1533–45.
18. Ma Z, Sun T. Adaptive sparse reduced-rank regression; 2014. https://arxiv.org/abs/1403.1922.
19. Ma Z, Ma Z, Sun T. Adaptive estimation in two-way sparse reduced-rank regression. Statistica Sinica. 2016. https://doi.org/10.5705/ss.202017.0073.
20. Luo R, Qi X. Signal extraction approach for sparse multivariate response regression. J Multivar Anal. 2017;153:83–97.
21. She Y. Selective factor extraction in high dimensions. Biometrika. 2017;104:97–110.
22. Castrillo A, Tontonoz P. PPARs in atherosclerosis: the clot thickens. J Clin Investig. 2004;114(11):1538–40.
23. Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. Nat Rev Genet. 2017;18:331.
24. Chen K, Chan K, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. J R Stat Soc Ser B. 2012;74:203–21.
25. Candes E, Tao T. The Dantzig selector: Statistical estimation when $\boldsymbol{p}$ is much larger than $\boldsymbol{n}$. Ann Stat. 2007;35:2313–51.
26. Hastie T, Tibshirani R. Efficient quadratic regularization for expression arrays. Biostatistics. 2004;5(3):329–40.
27. CVX-Research. Cvx: Matlab software for disciplined convex programming, version 2.0. 2012. http://cvxr.com/cvx.
28. Grant M, Boyd S. Graph implementations for nonsmooth convex programs. In: Blondel V, Boyd S, Kimura H, editors. Recent advances in learning and control. Lecture Notes in Control and Information Sciences, Springer-Verlag Limited; 2008. p. 95–110.
29. Bing X, Wegkamp M. Adaptive estimation of the rank of the coefficient matrix in high dimensional multivariate response regression models. Annals of Statistics. 2019;47:3157–84. https://arxiv.org/abs/1704.02381.
30. Peng J, Zhu J, Bergamaschi A, Han W, Noh D-Y, Pollack JR, Wang P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann Appl Stat. 2010;4:53–77.
31. Chun H, Keles S. Sparse partial least squares for simultaneous dimension reduction and variable selection. J R Stat Soc Ser B. 2010;72:3–25.
32. Golia E, Limongelli G, Natale F, Fimiani F, Maddaloni V, Pariggiano I, Bianchi R, Crisci M, D'Acierno L, Giordano R, et al. Inflammation and cardiovascular disease: from pathogenesis to therapeutic target. Curr Atheroscler Rep. 2014;16(9):435.
33. Willerson JT, Ridker PM. Inflammation as a cardiovascular risk factor. Circulation. 2004;109(21_suppl_1):2.
34. Stoner L, Lucero AA, Palmer BR, Jones LM, Young JM, Faulkner J. Inflammatory biomarkers for predicting cardiovascular disease. Clin Biochem. 2013;46(15):1353–71.

35. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation. 2014;129(25_suppl_2):49–73. https://doi.org/10.1161/01.cir.0000437741.48606.98.
36. Stone NJ. Preventing atherosclerotic cardiovascular disease using American College of Cardiology and American Heart Association Prevention Guidelines: some good news, but caveats remain. J Am Heart Assoc. 2016;5(8):004197. https://doi.org/10.1161/jaha.116.004197.
37. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for mendelian randomization. Stat Methods Med Res. 2017;26(5):2333–55. https://doi.org/10.1177/0962280215597579, PMID: 26282889.
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
39. Dong Z, Gao X, Chinchilli VM, Sinha R, Muscat J, Winkels RM, Richie Jr JP. Association of sulfur amino acid consumption with cardiometabolic risk factors: Cross-sectional findings from NHANES III. EClinicalMedicine. 2020;19:100248.
40. Suliman ME, Stenvinkel P, Heimbürger O, Bàràny P, Lindholm B, Bergström J. Plasma sulfur amino acids in relation to cardiovascular disease, nutritional status, and diabetes mellitus in patients with chronic renal failure at start of dialysis therapy. Am J Kidney Dis. 2002;40(3):480–8.
41. Wilcken D, Wilcken B. The pathogenesis of coronary artery disease. A possible role for methionine metabolism. J Clin Investig. 1976;57(4):1079–82.

Publisher's Note