Molecular Group and Correlation Guided Structural Learning for Multi-Phenotype Prediction

Xueping Zhou,¹ Manqi Cai,¹ Molin Yue,¹ Juan Celedón,² Ying Ding,¹ Wei Chen²,* and Yanming Li³,*

¹Department of Biostatistics, University of Pittsburgh, Pittsburgh, 15216, PA, USA, ²Department of Pediatrics, University of Pittsburgh Medical Center Children’s Hospital of Pittsburgh, Pittsburgh, 15224, PA, USA and ³Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas, 66160, KS, USA
*Corresponding author. Wei Chen wei.chen@chp.edu, Yanming Li yli8@kumc.edu

Abstract

We propose a supervised learning algorithm to perform feature selection and outcome prediction for genomic data with multi-phenotypic responses. Our algorithm particularly incorporates the genome and/or phenotype grouping structures and phenotype correlation structures in feature selection, effect estimation, and outcome prediction under a penalized multi-response linear regression model. Extensive simulations demonstrate its superior performance over its competing methods. We apply the proposed algorithm to two omics studies. In the first study, we identified novel association signals between multivariate gene expressions and high-dimensional DNA methylation profiles, providing biological insights into how CpG sites regulate gene expressions. The second study is for cell type deconvolution. Using the proposed algorithm, we were able to achieve better cell type fraction predictions using high-dimensional gene expression data.

Key words: association study, cell-type deconvolution, DNA methylation, feature selection, genomic grouping structure, multi-type prediction, multivariate regression

Introduction

Rapid advances in modern bio-technologies have yielded massive omics data, including genetics, transcriptomics, proteomics, and epigenetics among many others. Such datasets are high-dimensional, where the number of features exceeds the sample size. In association studies, it is often that only a small proportion of the features are associated with the outcomes of interest. To identify the truly associated features among high-dimensional candidates, regularization methods are commonly employed to shrink the coefficients of noises to zeros. Another intrinsic characteristic of omics data is the genome hierarchy grouping structures. For example, single nucleotide polymorphisms (SNPs) can be grouped by their harboring genes, and genes can be grouped by functional pathways. Regularized regression methods with high-dimensional predictors that take into consideration of the response covariance structure. Wilms and Croux[32] later proposed a multivariate group Lasso which performs signal selection between high-dimensional multivariate responses and multiple predictors while taking into account the grouping structures on both. MSGLasso outperforms the univariate response approaches in terms of outcome prediction and feature selection. However, MSGLasso does not explicitly consider the response covariance structure. Wilms and Croux[32] later proposed a multivariate group Lasso with covariance estimation that takes into account the covariance structure of the responses. However, it only considered the group-level sparsity, but not the within-group sparsity.
In this study, we propose an extension of MSGLasso – the Brilliant. Biological gRoup guNeeded muLtiVariate muLtiPle ineAR regression with peNalization. Same to MSGLasso, Brilliant utilizes the intrinsic genome hierarchy grouping structures in model fitting and predictions. It selects both the important group-level and therein the important individual-level association signals. Different to MSGLasso, Brilliant explicitly incorporates the covariance structure of multivariate responses. By doing so, the proposed Brilliant method improves the performance of feature selection, outcome prediction, and coefficient estimation significantly compared to MSGLasso and other univariate response approaches.

The rest of the paper is structured as follows. In section 2, we introduce the detailed algorithm of Brilliant. Section 3 uses simulation studies to demonstrate the performance of Brilliant. Section 4 presents two real data applications. A discussion is given in Section 5.

Method

Let $n$ denote the number of samples, consider multivariate linear regression model

$$Y = XB + E,$$  \hspace{1cm} (1)

where $Y = (y_1, \ldots, y_q) \in \mathbb{R}^{n \times q}$ is the response data matrix of $n$ samples and $q$ responses, and $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ is the predictor matrix of $n$ samples and $p$ predictors. $B \in \mathbb{R}^{p \times q}$ is the coefficient matrix, and $E \in \mathbb{R}^{n \times q}$ is the matrix of random errors. Assume that each row vector of $E$ follows a multivariate normal distribution $N(0, \Sigma)$ with a zero mean vector and a $q \times q$ covariance matrix $\Sigma$.

Suppose $X$ has $P$ groups and $Y$ has $Q$ groups. They together introduce $P \times Q$ intersection block groups [21] on the coefficient matrix $B$. Let $\mathcal{G}$ be the set of all $P \times Q$ groups and $B_g$ be a single block group in $\mathcal{G}$, $g \in 1, \ldots, G = P \times Q$. For a group $B_g$, denote its $L_2$ norm by $||B_g||_2 = \sqrt{\sum_{j \in B_g} \beta_{jk}^2}$.

Brilliant considers the following multi-responses optimization problem.

$$\arg\min_{B} \frac{1}{2n} \text{tr}((Y - XB)^T(Y - XB)\hat{\Omega}) + \lambda_1 \sum_{1 \leq j \leq p, 1 \leq k \leq q} |\beta_{jk}| + \lambda_2 \sum_{g \in \mathcal{G}} ||B_g||_2,$$  \hspace{1cm} (2)

where $\hat{\Omega}$ is a working version of the precision matrix $\Omega = \Sigma^{-1}$. When $\Omega$ is a diagonal matrix, Brilliant reduces to MSGLasso. Non-negative tuning parameters $\lambda_1$ and $\lambda_2$ are for the Lasso penalty and group Lasso penalty, respectively. The non-negative $w_{g, \phi}$ are adaptive weights for each individual group. In this paper, we set $w_{g, \phi}$ equal to total number of entries in the group $B_g$.

The working precision matrix $\hat{\Omega}$ can be estimated from the data. For example, using the centralized responses, one can calculate $\hat{\Omega}$ by directly invert the sample covariance matrix in a low dimensional setting or using graphical lasso [12] in a high-dimensional setting. In this paper, we propose to use a block diagonal working precision matrix with the diagonal blocks obtained by inverting the sample covariance matrices of responses within each group and the off-diagonal blocks set to be zero. We show that using an adequately specified $\hat{\Omega}$, Brilliant can outperform MSGLasso.

Let $B$ be the solution to (2) and $\beta_{jk}$ be its entry on the $j$th row and $k$th column. For a given grouping structure $\mathcal{G}$ on $B$ and a given set of tuning parameters, an iterative algorithm adopted from the mixed coordinate descent algorithm [21] for coefficient estimation is summarized as in Algorithm 1. In the original mixed coordinate descent algorithm [21], the outcome covariance structure was not explicitly incorporated into the estimation. The Brilliant specifically incorporates the outcome covariance structure by integrating the user-specified working precision matrix into the mixed coordinating algorithm. In Algorithm 1, the estimates of $\hat{\beta}_{jk}$ by Brilliant depend on $\hat{\omega}_{jk}$ and $S_{jk}$, where $\hat{\omega}_{jk}$ is the $kk$-th element of the working precision matrix $\hat{\Omega}$ and $S_{jk} = x_j^T(Y - XB_{(\bar{\cdot} - jk)})\hat{\Omega}_{k}$ with $B_{(\bar{\cdot} - jk)}$ being the $jk$-th element of $B$ replaced by zero and $\hat{\Omega}_{k}$ being the $kk$th column vector of the working precision matrix. Incorporating the outcome covariance structure can be beneficial especially when the covariance structure of the outcomes is priory known or when the outcomes are highly correlated within or between groups. It is important to note that Brilliant incorporates both outcome correlations from within group and across groups, which is evidenced in $S_{jk}$. Please refer to Appendix 6 for a detailed proofs of the updating formulas in algorithm 1.

Simulation Studies

In our simulation studies, we compared the performance of Brilliant with that of the MSGLasso using the R package MSGLasso [21], and of the univariate method Lasso using the package glmnet [11]. The total number of predictors was set to 500. These predictors were separated into 10 non-overlapping groups with each group containing 50 predictors. Predictors within the same group were correlated with either a compound symmetry (CS) or a first-order auto-regression (AR1) structure with a correlation coefficient $\rho = 0.5$. Predictors from different groups were set to be independent. The predictor vector for each subject was independently generated from multivariate normal distributions with the specified correlation structures, marginal variances 1, and zero means. The total number of responses was set to be 100. Responses were separated into four non-overlapping groups and each group included 20 responses. For a given grouping structure $\mathcal{G}$, we further assume the variables within a group are correlated, while variables from different groups are uncorrelated. That is equivalent to say that the true covariance matrix $\Sigma$ assumes the form of a block diagonal matrix. The error vector for each subject was also independently generated from multivariate normal distributions with the same correlation structure as the predictor vector and a correlation coefficient $\rho$ taking value in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The covariance structures of each response group are specified in Table 1.

The coefficient matrix $B$ was created to have a group structure with $P \times Q$ intersection groups. To achieve group-level sparsity of $B$, we randomly selected 1/5 of all groups to be non-zero groups, or groups with informative features. To obtain within-group sparsity, we randomly selected 1/2 of the entries within a non-zero group and set to 0s. The other 1/2 non-zero values were drawn from a uniform distribution $U\{[-3, 0) \cup (0, 3]\}$. $Y$ was generated based on the multivariate multiple linear regression framework described in (1).
The performance of outcome prediction, parameter estimation and feature selection was assessed. Outcome prediction was evaluated by average of the squared differences between the referenced cell fractions and their predicted values, or the mean square error (MSE), defined in (3), where $n$ denotes the sample size in the test set. Feature selection and coefficient estimation were evaluated based on the estimated regression coefficient matrix $\hat{\mathbf{B}}$. For coefficient estimation, mean absolute estimation error (MAEE) (4) was assessed. For feature selection, we calculated precision, recall and F1 score as formulated in (5)-(7), where |$S$| is cardinality of a set $S$.

$$\text{MSE} = \frac{1}{n \times q} \sum_{j=1}^{n} \sum_{k=1}^{q} (\hat{y}_{jk} - y_{jk})^2.$$  \hspace{1cm} (3)

$$\text{MAEE} = \frac{1}{p \times q} \sum_{j=1}^{p} \sum_{k=1}^{q} |\beta_{jk} - \hat{\beta}_{jk}|.$$  \hspace{1cm} (4)

$$\text{Precision} = \frac{|\{j : 1 \leq j \leq p, 1 \leq k \leq q, \beta_{jk} \neq 0 \text{ and } \hat{\beta}_{jk} \neq 0\}|}{|\{j : 1 \leq j \leq p, 1 \leq k \leq q, \hat{\beta}_{jk} \neq 0\}|}.$$  \hspace{1cm} (5)

$$\text{Recall} = \frac{|\{j : 1 \leq j \leq p, 1 \leq k \leq q, \beta_{jk} \neq 0 \text{ and } \hat{\beta}_{jk} \neq 0\}|}{|\{j : 1 \leq j \leq p, 1 \leq k \leq q, \beta_{jk} \neq 0 \text{ and } \hat{\beta}_{jk} \neq 0\}|}.$$  \hspace{1cm} (6)

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$  \hspace{1cm} (7)

Figure 1 and Figure 7 summarize the performance of Brilliant, MSGlLasso and Lasso when the covariance structure

---

**Table 1.** Within-group covariance structures of $X$ and $E$ in each simulation setting.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Covariance in $X$</th>
<th>Covariance in $E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CS, $\rho=0.5$</td>
<td>CS, $\rho=0.1$</td>
</tr>
<tr>
<td>2</td>
<td>CS, $\rho=0.3$</td>
<td>CS, $\rho=0.1$</td>
</tr>
<tr>
<td>3</td>
<td>CS, $\rho=0.5$</td>
<td>CS, $\rho=0.1$</td>
</tr>
<tr>
<td>4</td>
<td>CS, $\rho=0.7$</td>
<td>CS, $\rho=0.1$</td>
</tr>
<tr>
<td>5</td>
<td>CS, $\rho=0.9$</td>
<td>CS, $\rho=0.1$</td>
</tr>
<tr>
<td>6</td>
<td>AR1, $\rho=0.5$</td>
<td>AR1, $\rho=0.1$</td>
</tr>
<tr>
<td>7</td>
<td>AR1, $\rho=0.3$</td>
<td>AR1, $\rho=0.1$</td>
</tr>
<tr>
<td>8</td>
<td>AR1, $\rho=0.5$</td>
<td>AR1, $\rho=0.1$</td>
</tr>
<tr>
<td>9</td>
<td>AR1, $\rho=0.7$</td>
<td>AR1, $\rho=0.1$</td>
</tr>
<tr>
<td>10</td>
<td>AR1, $\rho=0.9$</td>
<td>AR1, $\rho=0.1$</td>
</tr>
</tbody>
</table>

* CS: compound symmetry.
* AR1: first-order auto-correlation.
* $\rho$: correlation coefficient.
is CS or AR1, respectively. Brilliant outperforms MSGLasso and univariate lasso when the correlation \( \rho \) within the same error group is 0.3 or greater. The other multivariate approach MSGLasso gives the second best prediction performance (Figure 1A) after Brilliant. This confirms that taking into account the multivariate response grouping structure can improve prediction compared to univariate response approach. It also shows that incorporating the responses’ covariances can further improve prediction. For coefficient estimation (Figure 1B), it shows that by explicitly incorporating the responses’ covariances, Brilliant obtains the lowest estimation error, especially when the responses are highly correlated. For feature selectio, as compared to MSGLasso and Lasso, Brilliant also gives higher precision, recall and F1 scores when response correlations range from moderate to high (Figure 1C-E). MSGLasso outperforms Brilliant when the correlation \( \rho \) is low (\( \rho = 0.1 \)). This is due to the estimation error from calculating the precision matrix (inverting the working covariance matrix) in Brilliant.

**Real Data Application**

We demonstrate the potential use and performance of Brilliant using the real data from our in-house multi-omics Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVAPR) study [10] and the Framingham Heart Study (FHS) [24, 22]. In the EVAPR study, we collect phenotype, bulk RNA-seq (RNA-seq) gene expression and DNA methylation (DNAm) data from Puerto Rican children aged 9-20 years. Here, we focus on the 218 non-asthma controls, who have no missing data on all the predictors and outcomes. The FHS dataset is downloaded from dbGaP (phs000007.v32.p13), which contains blood microarray expression for 4,100 samples.

**Association analysis of DNA methylation and gene expression**

In the first real data example, we applied Brilliant and its competing methods to the bulk RNA-seq gene expression and DNAm data from the nasal epithelium tissue samples of the 218 subjects to identify the susceptibility CpG site associated with expressions of atopy-related genes.

The multivariate responses were expressions of 152 genes in four atopy-related biological pathways from the Gene Ontology Consortium [3]. Previous studies have shown that cytokine [26] and histone [1] play important roles in atopy. Here, we include 37 genes from the “positive regulation of cytokine production” pathway and 22 genes from the “histone methylation” pathway. We also include 55 genes in “energy derivation by oxidation of organic compounds” pathway since oxidative stress is directly related to atopic diseases [6]. Moreover, 39 genes from the “transport along microtubule” pathway were included due to their potential role in atopic diseases [15]. The average of within-pathway pairwise Pearson correlations was 0.715 (standard error= 0.168, details are given in Table 2 and Figure 2). DNAm information on 1,897 CpG sites around the 152 genes (within the region of 1,000 base pairs up/down streams) were extracted. We further chose the top 500 CpGs with the largest

---

**Fig. 1.** Simulation result of \( p = 500 \) with 10 \( \mathbf{X} \) groups, \( q = 100 \) in 5 \( \mathbf{Y} \) groups. Covariance structure in each \( \mathbf{X} \) group follows CS with correlation coefficient \( \rho = 0.5 \). Covariance structure in each \( \mathbf{Y} \) group follows CS with \( \rho = 0.1, 0.3, 0.5, 0.7 \) or 0.9. Training sample size is 200 and test sample size is 100. A. Mean squared prediction error (MSE) with lower value indicating better prediction performance; B. Mean absolute estimation error of coefficient matrix (MAEE) with lower value suggesting better estimation; C. F1 score for feature selection; D. Precision for feature selection; E. Recall for feature selection; For C-E, a higher value indicates better feature selection performance.
Table 2. Average of pairwise Pearson correlations for genes within each outcome group.

<table>
<thead>
<tr>
<th>Gene Ontology Biological Process</th>
<th>Number of Genes</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histone methylation</td>
<td>22</td>
<td>0.690 (0.155)</td>
</tr>
<tr>
<td>Energy derivation by oxidation of organic compounds</td>
<td>54</td>
<td>0.618 (0.162)</td>
</tr>
<tr>
<td>Transport along microtubule</td>
<td>39</td>
<td>0.860 (0.081)</td>
</tr>
<tr>
<td>Positive regulation of cytokine production</td>
<td>37</td>
<td>0.772 (0.102)</td>
</tr>
</tbody>
</table>

* Averages of pairwise Pearson correlations (standard errors).

variances in their methylation levels ($M$ values). The log-transformed gene expression TPM values and log-transformed methylation $M$ values were then used as the input responses and predictors for the Brilliant and MSGLasso. The groups of the gene responses were based on their belonging pathways. The groups of CpG sites were also based on the belonging pathways of their harboring or closest nearby genes. As a result, there were four gene groups and four CpG groups, which together yield 16 ($= 4 \times 4$) interaction block groups on the underlying to be estimated regression coefficient matrix. The sizes of CpG groups vary from 91 to 143.

Fig. 2. Pearson correlation heatmap of 152 genes. The vertical axis is indexed by gene names. The horizontal axis is indexed by the pathway names that the gene belongs to.

A three-fold nested cross-validation (CV) procedure was implemented for Brilliant and the competing methods. The inner CV was for selecting the optimal tuning parameters and the outer CV was for generating prediction results and avoiding overfitting. The Brilliant method outperforms both Lasso and MSGLasso. The groups of the gene responses were based on their belonging pathways. The groups of CpG sites were also based on the belonging pathways of their harboring or closest nearby genes. As a result, there were four gene groups and four CpG groups, which together yield 16 ($= 4 \times 4$) interaction block groups on the underlying to be estimated regression coefficient matrix. The sizes of CpG groups vary from 91 to 143.

Table 3. Prediction performance of Brilliant and its competing methods in the EVAPR Nasal gene expression and methylation association study.

<table>
<thead>
<tr>
<th></th>
<th>Brilliant</th>
<th>Lasso</th>
<th>MSGLasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Squared Errors Per Gene</td>
<td>38.91</td>
<td>41.23</td>
<td>41.42</td>
</tr>
</tbody>
</table>

* “Sum of Squared Errors Per Gene” is the total sum of squared errors divided by the number of genes.

Fig. 3. Venn diagram of numbers of association signals detected by Brilliant and the competing methods in EVAPR study.

Brilliant detected 7,600 potential associations between CpG methylations and gene expressions. MSGLasso and Lasso identified 2,575 and 1,728 associations, respectively. Among the 7,600 potential associations selected by Brilliant, 408 were also detected by both Lasso and MSGLasso (Figure 3). In the association block between the CpG group – “histone methylation” pathway and the gene group – “energy derivation by oxidation of organic compounds” pathway, Brilliant detected 96 association signals between DNAm and gene expression (Figure 9). There were 143 CpG sites in the “histone methylation” pathway and 54 genes in the “energy derivation by oxidation of organic compounds” pathway. The top association signals uniquely detected by Brilliant are listed in Table 4. Among them, majority of the associated CpG sites are located near genes known as key transcriptional and/or epigenetic regulators, with the associating genes being their targeted regulating genes. For instance, the association between cg05608790 and NDUFb9 was detected by Brilliant ($\beta_{\text{brilliant}} = -0.0095$). This reflects the regulation of cg05608790 on its neighboring gene AUTS2 ($\beta_{\text{brilliant}} = -0.0150$), which is an epigenetic regulator of a transcriptional network including NDUFb9 [16]. Similarly, the association between cg19832721 and NDUFA4 ($\beta_{\text{brilliant}} = -0.0104$) reflects the regulational effects of cg19832721 on its nearby gene KANSL1 ($\beta_{\text{brilliant}} = 0.0114$), which is a key gene for producing the histone acetyltransferase 8 (KAT8, also known
as MOF) regulatory nonspecific lethal complex. This complex is critical for acetylation of nucleosomal histone H4 at lysine 16, a unique histone mark controlling chromatin structure, and therefore involved in the regulation of transcription [28, 29]. For this coefficient block, Lasso detects an additional 237 potential CpG-Gene associations. In contrast, MSGLasso detected only 7 association signals within this block.

Cell-type deconvolution for bulk gene expression in blood cell sample

Tissue-level omics studies are based on the average effects across heterogeneous collection of cells, such as studies on peripheral blood include signals from all underlying cell types. Therefore, the tissue-level association studies are prone to be confounded by the variation in cell composition across samples [19]. To eliminate such confounding, different algorithms of cell-type deconvolution have been proposed to infer the fractions of cell types in bulk gene expression data collected from tissue samples [4]. Among available methods, the reference-based cellular deconvolution are believed to be the most accurate and reliable approach [4, 17].

In the second real data application, we used both the EVAPR and FHS datasets to predict cell type proportions for blood neutrophil (Neutro), monocyte (Mono), lymphocyte (Lymph) and eosinophil (Eosino) using the bulk gene expression data from blood cell samples. The multivariate responses for both datasets were the fractions of the four cell types for each subject, the values of which were measured through complete blood cell counts [20]. We put all four cell type fractions into one response group for both datasets. Predictors in EVAPR dataset were expression levels of the signature genes of each cell type based on the widely used leukocyte gene signature matrix reference (LM22) [25]. Those genes were put into four predictor groups based on their reference cell type groups. More specifically, there were 64 signature genes in neutrophil group, 53 in lymphocyte group, 40 in monocyte group, and 39 genes in eosinophil group. The transcripts per million (TPM) counts of predictors were first undertaken a log transformation before applying the Brilliant algorithm. Predictors in FHS dataset contains blood microarray gene expressions from 4,100 samples. Those genes were also grouped into four groups based on the leukocyte signature genes available in the dataset (Details in Appendix 9). The microarray gene expression data were also log transformed.

Prediction performance of Brilliant were evaluated within each and also across the two datasets. For internal prediction within each datasets, a three-fold nested cross-validation was performed on each dataset to select the optimal hyperparameters and avoid overfitting. For cross dataset prediction, FHS data were used as the training data and EVAPR data were used as the test data. A three-fold cross-validation was performed on the FHS data to select the optimal hyperparameters. The prediction results are shown in Figure 4, where the scatter plots between the predicted and referenced cell fractions and their Pearson correlations were presented. The correlations range from 0.896.

The predicted cell type fractions from Brilliant were compared with other commonly used reference-based cell-type deconvolution methods, as in the study by Cai et al. [7], including CIBERSORT/CIBERSORTx [25], DeconRNASeq [13], DCQ [2], DeCompress [5], DSA [36], dtangle [17],EnsDeconv (Ensemble Deconvolution) [7], GEDIT (Gene Expression Deconvolution Interactive Tool) [23], EPIC (Estimating the Proportions of Immune and Cancer cells) [27], FARDEEP (Fast And Robust DEconvolution of Expression Profiles) [14], hspe (hybrid-scale proportions estimation) [18], lCeDT (Immune Cell Deconvolution in Tumor tissues) [33], and SCDC methods including non-negative least squares (SCDC NNLS) and inverse sum of squares (SCDC ISSE) [9]. All the results were based on the LM22 reference. The FHS dataset was used as the training data, and the EVARP dataset was used as the test data. The comparison results are given in Figure 6, where the Spearman’s rank correlation, a convention performance evaluation metric in the cell-type deconvolution

<table>
<thead>
<tr>
<th>Cpg Site</th>
<th>Gene</th>
<th>Chr</th>
<th>Cpg Position</th>
<th>Cpg Nearby Gene</th>
<th>β</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg04948765</td>
<td>AUTS2</td>
<td>7</td>
<td>7:70257926</td>
<td>AUTS2</td>
<td>0.0182</td>
<td>I-1</td>
</tr>
<tr>
<td>cg04948765</td>
<td>NDUFV2</td>
<td>18</td>
<td>7:70257926</td>
<td>AUTS2</td>
<td>0.0138</td>
<td>I-2</td>
</tr>
<tr>
<td>cg05608790</td>
<td>AUTS2</td>
<td>7</td>
<td>6:9427388</td>
<td>AUTS2</td>
<td>-0.0150</td>
<td>I-1</td>
</tr>
<tr>
<td>cg05608790</td>
<td>NDUFV2</td>
<td>8</td>
<td>6:9427388</td>
<td>AUTS2</td>
<td>-0.0095</td>
<td>I-2</td>
</tr>
<tr>
<td>cg05608790</td>
<td>PDHB</td>
<td>3</td>
<td>6:9427388</td>
<td>AUTS2</td>
<td>-0.0111</td>
<td>I-2</td>
</tr>
<tr>
<td>cg20134287</td>
<td>AUTS2</td>
<td>7</td>
<td>7:70061455</td>
<td>AUTS2</td>
<td>-0.0208</td>
<td>I-1</td>
</tr>
<tr>
<td>cg20134287</td>
<td>NDUFV2</td>
<td>1</td>
<td>7:70061455</td>
<td>AUTS2</td>
<td>-0.0192</td>
<td>I-2</td>
</tr>
<tr>
<td>cg20134287</td>
<td>PDHA1</td>
<td>X</td>
<td>7:70061455</td>
<td>AUTS2</td>
<td>0.0152</td>
<td>I-2</td>
</tr>
<tr>
<td>cg19832721</td>
<td>KANSL1</td>
<td>17</td>
<td>17:4429866</td>
<td>KANSL1</td>
<td>0.0114</td>
<td>I-1</td>
</tr>
<tr>
<td>cg19832721</td>
<td>NDUFV2</td>
<td>7</td>
<td>17:4429866</td>
<td>KANSL1</td>
<td>-0.0104</td>
<td>I-2</td>
</tr>
<tr>
<td>cg19832721</td>
<td>NDUFV2</td>
<td>8</td>
<td>17:4429866</td>
<td>KANSL1</td>
<td>-0.0122</td>
<td>I-2</td>
</tr>
<tr>
<td>cg18829443</td>
<td>TET3</td>
<td>2</td>
<td>2:74227916</td>
<td>TET3</td>
<td>-0.0171</td>
<td>I-1</td>
</tr>
<tr>
<td>cg18829443</td>
<td>NDUFV2</td>
<td>3</td>
<td>2:74227916</td>
<td>TET3</td>
<td>0.0181</td>
<td>I-2</td>
</tr>
</tbody>
</table>

* β is the estimated regression coefficient. “Chr” stands for chromosome.
* Cpg positions were mapped to hg19 reference genome.

Table 4. Top association signals between DNA methylation and gene expression uniquely selected by Brilliant.
field, between the predicted and referenced cell fractions for each of the aforementioned methods. Although not designed as a cell-type deconvolution algorithm, Brilliant achieves the highest average Spearman’s correlation (the vertical bars in Figure 6) across the four cell types. We also calculated the mean square of errors (MSE) between the predicted and referenced cell fractions for each of the methods (Figure 10 in the Appendix). Brilliant gave the smallest average MSE across cell types among all methods.

The feature selection performance of Brilliant was evaluated by comparing the estimated coefficient matrix in the final model (built on FHS data) with the LM22 reference gene matrix. The heatmap in Figure 5 displays the coefficient matrix estimated by Brilliant. The matrix has been divided into four sub-matrices based on the signature genes of each cell type. As anticipated, for each cell type, the signature genes in the LM22 reference have the highest estimate magnitudes in their corresponding cell type categories (indicated by darker red color in the corresponding column). Among the 188 genes selected in the final model, Brilliant correctly identified cell types for 144 signature genes. In other words, the highest proportioned cell type in the LM22 matrix is also estimated by Brilliant to have the highest coefficient magnitude across four cell types for these 144 genes. For another 20 genes, the cell type with the second-largest estimated coefficient magnitude is the highest proportioned cell type in the LM22 reference. These demonstrate that Brilliant is capable of identifying the most informative features associated with the outcomes of interest.

**Fig. 4.** Scatter plots of predicted v.s. referenced cell type fractions. The left panel plots are from CV using the EVAPR data only. The middle panel are from CV using the FHS data only. The right panel are from using EVAPR data as the test set while using the FHS data as the training set.
Fig. 5. Heatmap of coefficient matrix estimated by Brilliant using the FHS data. The matrix has been divided into four blocks based on the signature genes of each cell type. The four blocks contain the signature genes of Neutrophil, Eosinophil, Lymphocyte, and Monocyte cells, respectively.

Conclusion

In this paper, we proposed an analytical tool – Brilliant for variable selection and outcome prediction on multivariate response and high-dimensional predictor data. Brilliant not only utilizes the intrinsic grouping structures of both predictors and responses and performs both group-level and within group-level variable selection, but also explicitly incorporates the covariance structure of the multivariate responses. Brilliant significantly improves the coefficient estimation accuracy, feature selection and outcome prediction performance compared to the competing methods. Brilliant does not require iteratively estimating the precision matrix as in [34]. It only requires a one-step estimation of a working covariance structure according to the user specified grouping structure on the responses.

We applied Brilliant to a methylation–to–gene expression association study and a study using gene expressions to predict cell type fractions. Brilliant can be easily extended to solve many other multivariate-response and multiple-predictor problems, including but not limited to single cell multiomics and expression quantitative trait loci (eQTL) studies, which are beyond the scope of this paper. We will explore its wide applications in the future.

Appendix A: Proofs of technical results

Brilliant objective function

Assume $Y|X \sim N(XB, \Sigma)$, the log-likelihood of data can be written as follows.

$$\ln L(B | Y, X, \Sigma) = -\frac{1}{2} \{C - n \ln(|\Omega|) + \text{tr}[(Y - XB)^\top (Y - XB)\Omega]\}$$

(8)

Given the working precision matrix $\hat{\Omega}$, the penalized log likelihood can be written as below.

$$\hat{B} = \arg\min_B \frac{1}{2n} \text{tr} \left( (Y - XB)^\top (Y - XB)\hat{\Omega} \right) + \lambda_1 \sum_{1 \leq j \leq p, 1 \leq k \leq q} |\beta_{jk}| + \lambda_2 \sum_{g \in G} \|B_g\|_2.$$  

(9)

Derivative of Brilliant objective function

The objective function consists of three terms. We take the derivative of each term with respect to $\beta_{jk}$ separately. Then, we will combine the gradient or sub-gradient together.

Take derivative of the first term with respect to $\beta_{jk}$

$$\frac{\partial Q(B)}{\partial \beta_{jk}} = -2x_j^\top (Y - XB)\hat{\Omega}_{jk}.$$  

(10)
For a coordinate \(\beta_k\), we can get
\[
\frac{\partial Q(B)}{\partial \beta_{jk}} = -2 \times (S_{jk} - |x_j|_2^2 \hat{\beta}_k \hat{\omega}_{kk}).
\]

Take derivative of the second and third term with respect to \(\beta_k\)
For a coordinate \(\beta_k\), when \(L(B)\) is differentiable at \(\beta_k\),
\[
\frac{\partial \ln L(B)}{\partial \beta_{jk}} = -S_{jk}/n + \|x_j\|_2^2 \beta_k \hat{\omega}_{kk}/n + \lambda_1 \text{ sgn} (\beta_{jk}) + \lambda_2 \beta_{jk}/\|B_g\|_2.
\]

For general case, the derivative of the second term or \(\cdot \) penalty can be written as follows.
\[
\lambda_1 \text{ sgn} (\beta_{jk}) = \begin{cases} 
-\lambda_1 & \text{if } \beta_{jk} < 0, \\
-\lambda_1, \lambda_1 & \text{if } \beta_{jk} = 0, \\
\lambda_1 & \text{if } \beta_{jk} > 0.
\end{cases}
\]

Combine the derivative of the three terms
\[
\frac{\partial \ln L(B)}{\partial \beta_{jk}} = \begin{cases} 
-S_{jk}/n + \|x_j\|_2^2 \beta_k \hat{\omega}_{kk}/n - \lambda_1 + \lambda_2 \beta_{jk}/\|B_g\|_2, & \text{if } \beta_{jk} < 0, \\
\lambda_1 \lambda_2 & \text{if } \beta_{jk} = 0, \\
-S_{jk}/n + \|x_j\|_2^2 \beta_k \hat{\omega}_{kk}/n + \lambda_1 + \lambda_2 \beta_{jk}/\|B_g\|_2, & \text{if } \beta_{jk} > 0.
\end{cases}
\]

Solution to the Brilliant objective function
Set derivative to 0 and solve for \(\beta_{jk}\) with the corresponding constraint, we can get
\[
\hat{\beta}_{jk} = \begin{cases} 
\frac{S_{jk} + \lambda_1}{\|x_j\|_2^2 \hat{\omega}_{kk}/n + \lambda_2/\|B_g\|_2} & \text{if } S_{jk} < -\lambda_1, \\
0 & \text{otherwise,} \\
\frac{S_{jk} - \lambda_1}{\|x_j\|_2^2 \hat{\omega}_{kk}/n + \lambda_2/\|B_g\|_2} & \text{if } S_{jk} > \lambda_1.
\end{cases}
\]

Or, in one unified form:
\[
\hat{\beta}_{jk} = \frac{\text{sgn}(S_{jk})(|S_{jk}| - |\lambda_1|)}{\|x_j\|_2^2 \hat{\omega}_{kk} + \lambda_2/\|B_g\|_2}.
\]

Details when \(\beta_{ij} > 0\)
\[
-S_{jk}/n + \|x_j\|_2^2 \beta_{jk} \hat{\omega}_{kk}/n + \lambda_1 + \lambda_2 \beta_{jk}/\|B_g\|_2 = 0
\]
\[
\text{with constraint } \beta_{jk} > 0, \text{ or } \|x_j\|_2^2 \hat{\omega}_{kk} + n \lambda_2 > 0, \text{ we have}
\]
\[
S_{jk} < -\lambda_1.
\]

Details when \(\beta_{ij} < 0\)
\[
-S_{jk}/n + \|x_j\|_2^2 \beta_{jk} \hat{\omega}_{kk}/n + \lambda_1 + \lambda_2 \beta_{jk}/\|B_g\|_2 = 0
\]
\[
\text{with constraint } \beta_{jk} < 0, \text{ or } \|x_j\|_2^2 \hat{\omega}_{kk} + n \lambda_2 < 0, \text{ we have}
\]
\[
S_{jk} > \lambda_1.
\]

Details on \(\beta_{jk} = 0\)
We will see different scenarios involving \(\beta_{ij} = 0\).
S1. The group containing \(\beta_{jk}\) is not zero groups. This suggests that the Lasso (\(L_1\)) penalty is not differentiable, but the \(\cdot \) penalty is differentiable for all groups containing \(\beta_{jk}\). Set the derivative to 0, we get
\[
S_{jk}/n = \lambda_2 \beta_{jk}/\|B_g\|_2 + \|x_j\|_2^2 \beta_{jk} \hat{\omega}_{kk}/n + n \lambda_1 u.
\]

S2. The group containing \(\beta_{jk}\) is a zero group. Let \(G^* = \{g : \beta_{jk} \in B_g \in G, \|B_g\|_2 > 0\}\). Let \(B_{g_0}\) be the zero group containing \(\beta_{jk}\). Let \(|u| < 1, \text{ and } \|v\|_2 < 1\). Set the derivative to 0, we have
\[
-S_{jk}/n + \|x_j\|_2^2 \beta_{jk} \hat{\omega}_{kk}/n + \lambda_1 u + \lambda_2 v_{jk} + \lambda_2 \beta_{jk}/\|B_g\|_2 = 0
\]
\[
\text{with constraint } \beta_{jk} > 0, \text{ or } \|x_j\|_2^2 \hat{\omega}_{kk} + n \lambda_2 > 0, \text{ we can get}
\]
\[
S_{jk} = n \lambda_1.
\]

Appendix B: Simulation Result
For \(p = 500\) and \(q = 100\), the simulation results under compound symmetry (CS) covariance structure (setting 1-5 in Table 1) are presented in the main article. For results under first order autoregressive (AR1) covariance structure (setting 6-10 Table 1), we observe very similar performance of Brilliant as in CS settings. More specifically, Brilliant achieves improved prediction and coefficient estimation compared to MSGLasso and univariate Lasso. Brilliant also shows significant improved feature selection performance, especially in F1 score and precision.
Fig. 7. Simulation result of \( P = 500 \) in 10 groups, \( Q = 100 \) in 5 groups, covariance structure of each predictor group is first-order auto-regressive (AR1) with associated correlation coefficient \( \rho = 0.5 \), covariance structure of each error group is CS with \( \rho = 0.1, 0.3, 0.5, 0.7, \) or 0.9. Training sample size is 200. **A.** Mean squared error (MSE) metric of prediction error, and lower value means better prediction performance; **B.** Mean absolute estimation error (MAEE) of coefficient matrix with lower value suggesting better estimation; **C.** F1 score for feature selection; **D.** Precision for feature selection. **E.** Recall for feature selection; For **C-E** higher value indicates better feature selection performance.

Appendix C: Further explanation of Lasso feature selection performance

We use a simply simulation study to demonstrate that Lasso is susceptible to selecting more FP signals when there exist correlated outcomes. In this example, we simulated a group of 100 predictors and 40 outcomes. The predictors were generated using the multivariate normal distribution with covariance matrix following compound symmetry (cs) structure with \( \rho = 0.2 \). The outcomes were simulated from the multivariate normal distribution with covariance matrix following compound symmetry structure and \( \rho = 0.9 \). Then we replaced \( Y_1 \) with \( Y_1 + X_1 + X_2 \). There were 25 observations in the training dataset. In this case, although both Lasso and Brilliant select the true signals, namely \( Y_1 - X_1 \) and \( Y_2 - X_2 \), Lasso identify an additional 331 false positives (FPs), while Brilliant only have 6 more FPs (Figure 8).

Appendix D: Details on cell-type deconvolution for blood cell samples

In the LM22 reference matrix, all entries are positives with higher value indicating higher expression level of the gene in the corresponding cell type. For each gene, we set it as the signature gene for the cell type with the largest value in the reference matrix. The number of available signature genes in the EVAPR and FHS dataset are summarized in Table 5. Table 6 summarizes the top performers of cell type deconvolution for the EVAPR dataset.

Competing interests

The authors declare no competing interest is declared. This study was partially supported by National Science Foundation (award number 2225775).

Author contributions statement

X.Z., Y.L. and W.C. conceived and designed the study. M.Y. and J.C. was responsible for data management. X.Z. and M.C. were responsible for data cleaning and analysis. X.Z., Y.D., Y.L. and W.C. were responsible for interpretation. X.Z. and Y.L. wrote the paper. All authors read and approved the final article.

Acknowledgments

The authors thank all the participants and their families in the Epigenetic Variation and Childhood Asthma in Puerto Ricans study and Framingham Heart Study for enabling this research and further our understanding of childhood asthma and allergy by their participation in the studies. We appreciate the researchers of the Framingham Heart Study for making the data available on dbGAP. This research is supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483, and the H2P cluster, which is supported by NSF award number OAC-2117681.
Fig. 8. Heatmaps of estimated coefficient to demonstrate Lasso select more false positives when the multivariate outcomes are correlated. Note: Predictors without any nonzero estimated coefficients by both Lasso and Brilliant are omitted in the graph.

Table 5. Signature genes available in EVAPR and FHS dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total number of genes</th>
<th>Neutrophil</th>
<th>Lymphocyte</th>
<th>Monocyte</th>
<th>Eosinophil</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVAPR</td>
<td>200</td>
<td>64</td>
<td>53</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>FHS</td>
<td>196</td>
<td>61</td>
<td>40</td>
<td>40</td>
<td>55</td>
</tr>
<tr>
<td>Overlap</td>
<td>188</td>
<td>59</td>
<td>51</td>
<td>37</td>
<td>37</td>
</tr>
</tbody>
</table>

* The values denote the number of signature genes.
* The overlapping genes in the two datasets are used to build the final model using the FHS data, and predict the cell type fractions for the EVAPR data.

References

6. Russell P Bowler and James D Crapo. Oxidative stress in allergic respiratory diseases. Journal of Allergy and...


22. Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and...


