Exploring the Genetic Patterns of Complex Diseases via the Integrative Genome-Wide Approach

Teng, Ben; YANG, Can; LIU, Jiming; Cai, Zhipeng; WAN, Xiang

Published in:
IEEE/ACM Transactions on Computational Biology and Bioinformatics

DOI:
10.1109/TCBB.2015.2459692

Published: 01/05/2016

Document Version:
Peer reviewed version

Exploring the genetic patterns of complex diseases via the integrative genome-wide approach

Ben Teng, Can Yang, Jiming Liu, Zhipeng Cai and Xiang Wan *

Abstract—Genome-wide association studies (GWASs), which assay more than a million single nucleotide polymorphisms (SNPs) in thousands of individuals, have been widely used to identify genetic risk variants for complex diseases. However, most of the variants that have been identified contribute relatively small increments of risk and only explain a small portion of the genetic variation in complex diseases. This is the so-called missing heritability problem. Evidence has indicated that many complex diseases are genetically related, meaning these diseases share common genetic risk variants. Therefore, exploring the genetic correlations across multiple related studies could be a promising strategy for removing spurious associations and identifying underlying genetic risk variants, and thereby uncovering the mystery of missing heritability in complex diseases. We present a general and robust method to identify genetic patterns from multiple large-scale genomic datasets. We treat the summary statistics as a matrix and demonstrate that genetic patterns will form a low-rank matrix plus a sparse component. Hence, we formulate the problem as a matrix recovering problem, where we aim to discover risk variants shared by multiple diseases/traits and those for each individual disease/trait. We propose a convex formulation for matrix recovery and an efficient algorithm to solve the problem. We demonstrate the advantages of our method using both synthesized datasets and real datasets. The experimental results show that our method can successfully reconstruct both the shared and the individual genetic patterns from summary statistics and achieve comparable performances compared with alternative methods under a wide range of scenarios. The MATLAB code is available at:http://www.comp.hkbu.edu.hk/~xwan/iga.zip.

Index Terms—SNPs, GWASs, biclustering, low-rank and sparse, convex optimization.

1 INTRODUCTION

Many common human diseases, such as type-1 and type-2 diabetes, depression, schizophrenia, and prostate cancer, are influenced by several genetic and environmental factors. Scientists and public health officials have great interests in finding genetic patterns associated with complex diseases, not only to advance our understanding of multi-gene disorders, but also to provide more insights into complex diseases. Disease association studies have provided substantial evidences indicating that complex diseases originate in disorders of multiple genes [1], [2]. Nevertheless, until recently the full-coverage identification of the genetic variants contributing to complex diseases has been unattainable.

After the completion of the Human Genome Project [3], [4] and the initiation of the International HapMap Project [5], interest has focused on genome-wide association studies (GWASs), in which the goal is to identify single-nucleotide polymorphisms (SNPs) that are associated with complex diseases (such as diabetes) or traits (such as human height). As of Dec. 2014, more than $15,000$ SNPs have been reported to be associated with at least one disease/trait at the genome-wide significance level (P-value$\leq 5 \times 10^{-8}$) [6]. However, most of the findings only explain a small portion of the genetic contributions to complex diseases. For example, all of the 18 SNPs identified in type 2 diabetes (T2D) only account for about 6% of the inherited risk [7]. There is still a large portion of disease/trait heritability that remains unexplained. This is the so-called missing heritability problem [7], [8], which is often used to denote the gap between the expected heritability of many common diseases, as estimated by family and twin studies, and the overall additive heritability obtained by accumulating the effects of all of the SNPs that have been found to be significantly associated with these conditions.

A recent study [9] has suggested that most of the heritability is not missing but can be explained by the effects of many genetic variants, with each variant probably contributing a weak effect. However, finding variants with small effects is very challenging in computation because the traditional single-locus based test cannot identify such variants and the number of groups of multiple variants to be investigated in GWAS is astronomical. In addition, in the high-dimensional and low-sample size settings of GWAS, many irrelevant variants tend to have high sample correlations due to randomness, which makes

B. Teng, J. Liu and X. Wan are with Department of Computer Science and Institute of Theoretical and Computational Study, Hong Kong Baptist University, Hong Kong. C. Yang is with Department of Mathematics, Hong Kong Baptist University, Hong Kong. Z. Cai is with Department of Computer Science, Georgia State University, Atlanta, USA.

** The corresponding author. Email: xwan@comp.hkbu.edu.hk*

GWAS prone to false scientific discoveries. To solve the missing heritability problem, a large sample size is required, but such a requirement is usually beyond the capacity of a single GWAS, as the sample recruitment is expensive and time consuming.

Evidence has indicated that many complex diseases are genetically related [10], [11], [12], [13], meaning that these diseases share common genetic risk variants. This suggests that an integrative analysis of related genomic data could be a promising strategy for removing spurious associations and identifying risk genetic variants with small effects, and thus finding the missing heritability of complex diseases. As high-throughput data acquisition becomes popular in biomedical research, new computational methods for large-scale data analysis become more and more important.

When analyzing genomic data from multiple related studies, the ideal scenario is for the individual-level data to be available for all of the included studies, but this may be difficult to achieve due to restrictions on sharing individual-level data. In fact, summary data (mostly P-values or z-scores) are more frequently shared. To identify significant SNPs shared by all of the included studies, the commonly used statistical approach is to combine P-values using Fisher's method [14]. [15] generalized Fisher's method to include weights when combining P-values. [16] suggested using the inverse normal transformation and Mosteller and [17] further generalized Stouffer's method by including weight when combining z-scores. There are two issues in such traditional statistical approaches. First, one small P-value can overwhelm many large P-values and dominate the test statistic. In high-dimensional and low-sample size settings, many irrelevant variants tend to have high significance due to randomness, which may cause wrong statistical inferences. Second, the information about genetic correlations between SNPs in the original data is completely lost after combining P-values. This information is necessary for understanding the genetic architecture of complex diseases because common complex diseases are associated with multiple genetic variants.

To identify shared genetic structures across multiple related studies, one feasible approach is to conduct a biclustering analysis on a matrix of summary statistics, in which the columns represent studies and the rows represent genetic variants, to simultaneously group studies and genetic variants. Many biclustering methods have been proposed and some comprehensive reviews of biclustering methods can be found in [18], [19], and [20]. However, the traditional biclustering methods do not perform well on genomic data because genomic data is high dimensional and most of its genetic variants are irrelevant. To obtain sparse and interpretable biclusters, a novel statistical approach, **sparseBC**, is recently proposed, which adopts an l_1 penalty to the means of the biclusters [21]. A big drawback of **sparseBC** is that it does not allow for overlapping biclusters, which limits its application in genomic data analysis because the shared genetic patterns in GWASs may be very complex. Furthermore, in genomic data, besides the shared genetic structure, each disease/trait owns some distinct genetic variants. The typical biclustering model may treat them as noises and discard them.

In this paper, we introduce a new method to identify genetic patterns in high dimensional genomic data. Our method possesses several advantages over existing works. First, our method admits a single model to detect both shared and individual genetic patterns among multiple studies. Second, our method employs two tuning parameters that control the size of the shared genetic pattern and the numbers of individual signals. The choices of these parameters have solid theoretical support. Third, our method produces the unique global minimizer to a convex problem, which means that the solution is always stable.

To demonstrate the performance of our proposed method, we conduct comparison experiments using both synthesized datasets and real datasets. Simulation results show that the proposed method achieves comparable performances compared with existing methods in many settings. A large dataset containing 32 GWASs is also analyzed to demonstrate the advantage of our method. Specifically, we propose the convex formulation, the algorithm, and the parameter selection in Section 2. Simulation studies and real data analysis are presented in Section 3. We conclude the paper with some discussions in Section 4.

2 METHOD

2.1 Formulation

Mathematically, the summary statistics from multiple related studies can be expressed as a matrix $\mathbf{D} \in \mathbb{R}^{p\times n}$, where each entry d_{ij} is a z-score (if only P-values are available, we can transform them into z-scores), and n and p are the numbers of studies and SNPs, respectively. Our goal is to (1) detect shared genetic patterns across studies, which can be represented as sparse biclusters in this matrix and (2) detect individual genetic variants for each study, which we assume are randomly distributed and sparse. Since the sparsity of biclusters in a matrix indicates a low-rank property (please see examples in simulation studies), the problem of identifying these two types of genetic patterns can be treated as a problem of recovering a low-rank component $\mathbf{X}$ and a sparse component $\mathbf{E}$ from the input data $\mathbf{D}$. Our proposed approach is based on the assumed sparsity of genetic patterns because in large-scale genomic data, most genetic variants are irrelevant.

We propose to use the following decomposition model to detect genetic patterns from noisy input:

$$\mathbf{D} = \mathbf{X} + \mathbf{E} + \epsilon, \tag{1}$$

where $\mathbf{X}$ is a low-rank component, $\mathbf{E}$ is a sparse component, and ϵ is a noise component. In GWAS data analysis, the low-rank component corresponds to the causal SNPs that are shared by several diseases/traits. The sparse component corresponds to the causal SNPs that affect one specific disease/trait. The noise component corresponds to the measurement error, which is often modeled by i.i.d. Gaussian distribution with a zero mean.

Naturally, to achieve the decomposition, the following minimization problem is considered:

$$\begin{aligned} \min_{\mathbf{X},\mathbf{E},\epsilon} \ & \frac{1}{2} \|\epsilon\|_F^2 + \alpha \text{rank}\,(\mathbf{X}) + \beta \|\mathbf{E}\|_0 \\ \text{s.t.} \ & \mathbf{D} = \mathbf{X} + \mathbf{E} + \epsilon, \end{aligned} \tag{2}$$

where $\|\epsilon\|_F = \sqrt{\sum_{i,j} \epsilon_{ij}^2}$ is the Frobenius norm and $\|\mathbf{E}\|_0$ is the ℓ_0-norm that counts the number of nonzero values in $\mathbf{E}$. The solution to Eq.(2) will give a penalized maximum likelihood estimate with respect to the variables $\mathbf{X}, \mathbf{E}, \epsilon$.

However, the proposed model in Eq.(2) is intractable and NP-hard. Thus, in order to effectively recover $\mathbf{X}$ and $\mathbf{E}$, we use the convex relaxation to replace the rank $(\cdot)$ by the nuclear norm and the ℓ_0-norm by the ℓ_1-norm. Here, the nuclear norm is defined as $\|\mathbf{X}\|_* = \sum_{i=1}^r \sigma_i$, where $\sigma_1, \cdots, \sigma_r$ are the singular values of $\mathbf{X}$. It is the tightest convex surrogate to the rank operator [22] and has been widely used for low-rank matrix recovery [23]. The ℓ_1-norm is defined as $\|\mathbf{X}\|_1 = \sum_{i,j} |X_{ij}|$. The ℓ_1 relaxation has proven to be a powerful technique for sparse signal recovery [24].

Finally, instead of directly solving Eq.(2), we solve the following problem,

$$\mathcal{F}(X, E) = \min_{\mathbf{X},\mathbf{E}} \ \frac{1}{2} \|\mathbf{D} - \mathbf{X} - \mathbf{E}\|_F^2 + \alpha \|\mathbf{X}\|_* + \beta \|\mathbf{E}\|_1. \tag{3}$$

It is easy to prove that Eq.(3) is a convex problem and therefore, the global optimal solution is unique. We will introduce an algorithm to solve this optimization problem in the next subsection.

2.2 Algorithm

The optimization problem of Eq.(3) can be solved by alternatively solving the following two sub-problems until convergence:

$$\hat{\mathbf{X}} \leftarrow \arg\min_{\mathbf{X}} \mathcal{F}(\mathbf{X}, \hat{\mathbf{E}}) \tag{4}$$

$$\hat{\mathbf{E}} \leftarrow \arg\min_{\mathbf{E}} \mathcal{F}(\hat{\mathbf{X}}, \mathbf{E}). \tag{5}$$

The theoretical proof for the convergence can be found in [25].

The problem in Eq.(4) can be reduced to

$$\min_{\mathbf{X}} \ \frac{1}{2} \|\mathbf{D} - \hat{\mathbf{E}} - \mathbf{X}\|_F^2 + \alpha \|\mathbf{X}\|_*, \tag{6}$$

which becomes a nuclear-norm regularized least-squares problem and has the following closed-form solution [26],

$$\hat{\mathbf{X}} = \mathcal{D}_\alpha \left(\mathbf{D} - \hat{\mathbf{E}} \right), \tag{7}$$

where $\mathcal{D}_\lambda$ refers to the singular value thresholding (SVT)

$$\mathcal{D}_\lambda(\mathbf{M}) = \sum_{i=1}^r (\sigma_i - \lambda)_+ \mathbf{u}_i \mathbf{v}_i^T. \tag{8}$$

Here, $(x)_+ = \max(x, 0)$. $\{\mathbf{u}_i\}$, $\{\mathbf{v}_i\}$, and $\{\sigma_i\}$ are the left singular vectors, the right singular vectors, and the singular values of $\mathbf{M}$, respectively.

The problem in Eq.(5) can be rewritten as

$$\min_{\mathbf{E}} \frac{1}{2} \|\mathbf{D} - \hat{\mathbf{X}} - \mathbf{E}\|_F^2 + \beta \|\mathbf{E}\|_1. \tag{9}$$

It admits a closed-form solution

$$\hat{\mathbf{E}} = \mathcal{S}_\beta \left(\mathbf{D} - \hat{\mathbf{X}} \right), \tag{10}$$

where $\mathcal{S}_\beta(\mathbf{M})_{ij} = \text{sign}(M_{ij})(M_{ij} - \beta)_+$ refers to the elementwise soft-thresholding operator [25].

Overall, the algorithm to optimize the proposed model in Eq.(3) is summarized in Algorithm 1. It will give a global optimal solution independent of initialization.

Algorithm 1 The algorithm to solve Eq.(3).

1. **Input:** $\mathbf{D}$
2. Initialize all variables to be zero.
3. **repeat**
4. Update $\mathbf{X}$ by solving Eq.(6) via singular value thresholding.
5. Update $\mathbf{E}$ by solving Eq.(9) via soft thresholding.
6. **until** convergence
7. **Output:** $\hat{\mathbf{X}}$ and $\hat{\mathbf{E}}$

2.3 Parameter selection

There are two parameters in our model, which can be estimated properly via the analysis of the size of the input matrix (n, p) and the standard variation of the noise σ [23], [27].

First we choose α following the same way as in [28], [27]. When we fix $\mathbf{E} = 0$ in Eq.(3), the solution $\hat{\mathbf{X}}$ of Eq.(3) is equal to the singular value thresholding version of $\mathbf{D}$ with threshold α. Similarly, when we fix $\mathbf{X} = 0$ in Eq.(3), the solution $\hat{\mathbf{E}}$ is equal to the entry-wise shrinkage version of $\mathbf{D}$ with threshold β. Thus, we can choose α to be the smallest value such that the minimizer of Eq.(3) is likely to be $\hat{\mathbf{X}} = \hat{\mathbf{E}} = 0$. In this

way, α is large enough to threshold away the noise, but not too large to over-shrink the original matrices. Since ϵ is modeled by i.i.d. Gaussian distribution with a zero mean, we estimate $\alpha = (\sqrt{n} + \sqrt{p})\sigma$, which is the expected ℓ_2-norm of a $p \times n$ random matrix with entries sampled from $\mathcal{N}(0, \sigma^2)$.

As causal SNPs are sparse in the data, we can estimate σ from the data by the median-absolute-deviation estimator [29]

$$\hat{\sigma} = 1.48 \text{ median} \{|\mathbf{D} - \text{median}(\mathbf{D})|\}. \quad (11)$$

The relative weight $\lambda = \beta/\alpha$ balances the two terms in $\alpha\|\mathbf{X}\|_* + \beta\|\mathbf{E}\|_1$ and consequently controls the rank of $\mathbf{X}$ and the sparsity of $\mathbf{E}$. It has been proved that $\lambda = 1/\sqrt{m}$ gives a large probability of recovering $\mathbf{X}$ and $\mathbf{E}$ under their assumed conditions [23] and the value of λ can be adjusted slightly to obtain the best results in specific applications. Here, m is the larger dimension of the input matrix. In our problem, $m = p$, i.e. the number of SNPs. However, on real datasets, the shared SNPs rarely form a perfectly low-rank matrix, and we set $\lambda = 2/\sqrt{m}$ to keep sufficient variations in $\mathbf{X}$. We conduct an experiment to see the performance of our model for different values of λ. The results are shown in the supplementary.

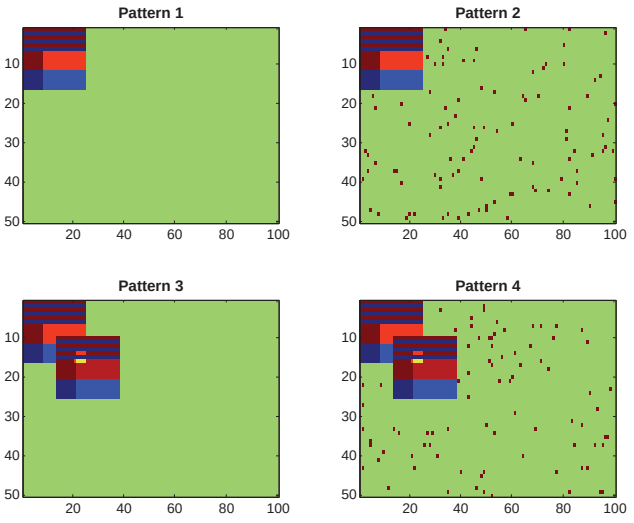

Figure 1. Four scenarios in our simulation study. Pattern 1 contains a rank-1 component representing one bicluster. Pattern 2 adds some sparse signals in Pattern 1. Pattern 3 contains a rank-2 component representing two overlapped biclusters. Pattern 4 contains sparse signal in addition to overlapped biclusters.

3 RESULTS

3.1 Simulation studies

We first compare the performance of our method under four simulation studies, with four existing biclustering methods: sparseBC (sparse biclustering) [21], LAS [30], BBC [31] and SSVD [32]. The results of the biclustering method are a set of sub-matrices of the observed data matrix. In general, if one sub-matrix meets a predefined criterion, then all entries in this sub-matrix will be considered as meaningful signals. Specifically, for sparse biclustering method, we use the parameters that have been mentioned in [21], and the entries in the clusters which satisfy a preselected cutoff are recognized as the final result. For LAS, we use the default settings. For SSVD, which uses a variant of singular value decomposition to find biclusters, we try different setting of parameters and report the best one as its result. LAS and SSVD can detect overlapping biclusters but sometimes they report the entire matrix as one bicluster. Thus, for both LAS and SSVD, the biclusters that contain the entire matrix are discarded. For BBC method, we use its implementation in MTBA [33], a MATLAB Toolbox for biclustering analysis. For our method, the parameters are selected as stated in Section 2. Then we use a threshold T to determine whether the entries (i, j) of the matrix are reported or not by comparing the value of $X(i, j)$ and $E(i, j)$ with T.

We evaluate each method in terms of the $F1$-score, which can be calculated as following:

$$precision = \frac{tp}{tp + fp}, \quad (12)$$

$$recall = \frac{tp}{tp + fn}, \quad (13)$$

$$F1\text{-score} = \frac{2 * precision * recall}{precision + recall}, \quad (14)$$

where tp and fp denote the number of true positives and false positives, respectively, and fn denotes the number of false negatives.

3.1.1 Simulation settings

We adopt four patterns (each in one simulation study) illustrated in Figure 1 to generate synthetic data.

- Pattern 1 adopts a case from [32], which generated a rank-1 true signal matrix. Let $\mathbf{M} = d\mathbf{u_1}\mathbf{v_1^T}$ be a 100×50 matrix with $d = 50$, $\hat{v_1} = [10, 9, 8, 7, 6, 5, 4, 3, r(2, 17), r(0, 75)]$, $\hat{u_1} = [10, -10, 8, \ -8, 5, -5, r(-3, 5), r(0, 34)]^T$, $u_1 = \hat{u_1}/\|\hat{u_1}\|_2$, and $v_1 = \hat{v_1}/\|\hat{v_1}\|_2$, where $r(a, b)$ denotes a vector of length b with all entries equal a. This case simulates the shared causal SNPs among several studies.
- Pattern 2 extends Pattern 1 by adding some sparse signals. That is, we generate a sparse component $\mathbf{E}$, whose entries are independently distributed, each taking on value 0 with probability $1 - p_s$, and value 6 with probability $p_s = 0.01$.

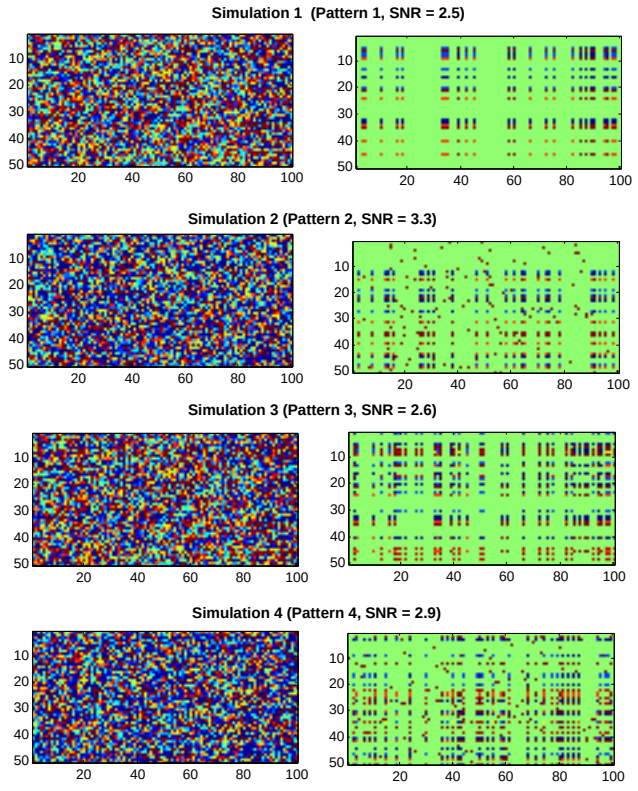

Figure 2. Illustrations of four simulations. For each simulation, the generated matrix with noises is shown in the left panel and the ground truth matrix is shown in the right panel.

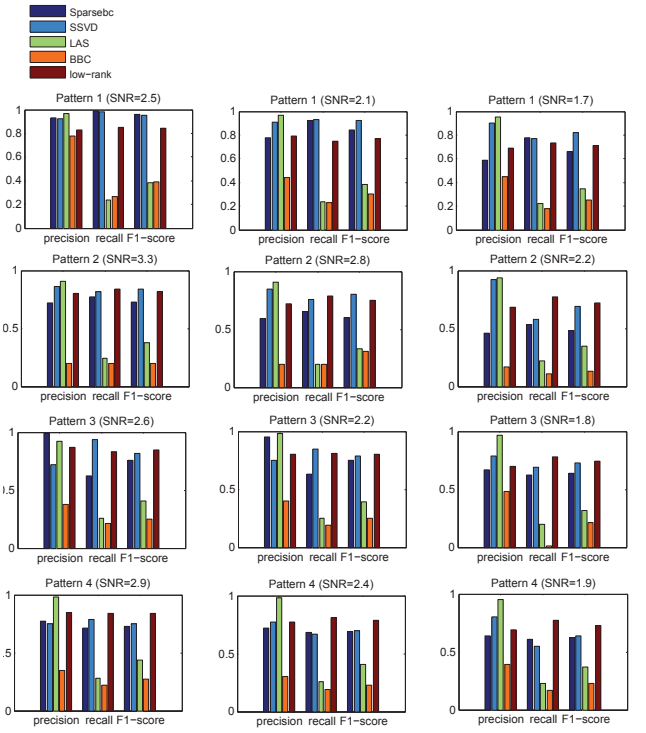

Figure 3. Comparison results of different methods in four simulation studies, each using one pre-defined pattern.

- Pattern 3 adopts the case from [21], which generated two overlapping biclusters. Let $\mathbf{M} = d(\mathbf{u_1 v_1^T} + \mathbf{u_2 v_2^T})$ be a 100×50 matrix with $d = 50$, u_1 and v_1 as defined in simulation 1, $\hat{u_2} = [r(0, 13), 10, 9, 8, 7, 6, 5, 4, 3, r(2, 17), r(0, 62)]$, $\hat{v_2} = [r(0, 9), 10, -9, 8, -7, 6, -5, r(4, 5), r(-3, 5), r(0, 25)]^T$ $u_2 = \hat{u_2}/\|\hat{u_2}\|_2$, and $v_2 = \hat{v_2}/\|\hat{v_2}\|_2$.
- Pattern 4 extends Pattern 3 by adding some sparse signals in the same way as Pattern 2.

3.1.2 Data generation

Given a specific pattern mentioned above, we first generate the data matrix. To simulate the real situation, we randomly shuffle the rows and the columns. Next, we add Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to each item. Figure 2 illustrates the ground truth data and the generated data. For each generated data matrix, we also compute the signal to noise ratio (SNR). To illustrate how the methods perform for the data with different SNRs, we further scale down the ground truth signal by dividing the original values by 1.2 and 1.5, respectively.

3.1.3 Simulation results

The results of four simulation studies are shown in Figure 3. We use 'low-rank' to represent our method as our model is to find biclusters via a low-rank approximation. The details of the simulation results can be found in the supplementary materials. In general, our proposed method achieves comparable performance with SSVD and Sparsebc in the four simulation studies.

Figure 4 shows one result in the fourth pattern. Our proposed method can successfully recover a low-rank component and a sparse component from raw data. In the first simulation, the $F1$-scores of sparse biclustering method and SSVD method almost get to 1. The reason why our method performs worse is that we use the default parameters which are not best fit for this simulation set-up. When adjusting the parameters, our method can also get a high $F1$-score. Since the generated clusters in the simulations are multiplicative biclusters, the LAS method and the BBC method, which are designed to identify constant biclusters and additive biclusters, perform poorly under these simulations.

Furthermore, we can observe from Figure 3 that our method always performs equally well in terms of both precision and recall while the other methods often favor precision against recall. In the large-scale data analysis, the conservative method with high precision and low recall may not be suitable for new discoveries because most signals are irrelevant. For such situations, our method has a clear advantage over competitors.

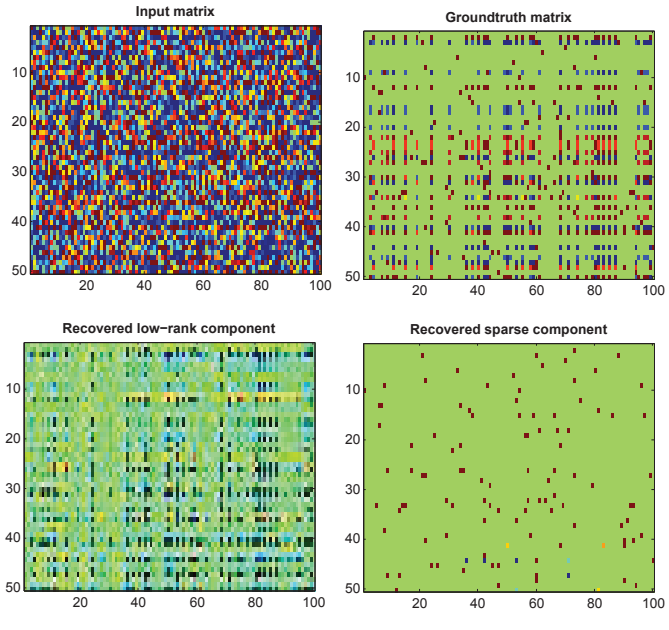

Figure 4. An illustration of the simulation result. The low-rank component and the sparse component are recovered by our method.

We also compare our method with the other methods on the simulations with biological characteristics. The details of the simulation results can be found in the supplementary materials.

3.2 Real application

We applied our method to analyze 32 independent diseases/traits, including

- 3 anthropometrics related data.
- 9 psychiatry related data.
- 8 CAD data.
- 2 social science studies.
- 2 glycaemic traits.
- 6 inflammatory bowel disease data.
- systemic lupus erythematosus.
- Parkinson's disease.

The details of the data sets including the references and the web links for downloading the data can be found in the supplementary materials. Since each study reports different SNPs, we take the SNPs that are reported by at least 28 diseases/traits and obtain their P-values and impute the missing ones. Finally, we get a P-value matrix $P \in R^{466423\times 32}$ for these 32 diseases/traits. Next, we convert the P-value matrix to the z-score matrix $Z \in R^{466423\times 32}$. We analyze this data set using our method on a desktop PC with 2.40GHz CPU and 4GB RAM. The running time of our method on 32 GWASs data sets is only 152.1s. The alterative methods investigated in this work cannot be applied due to the large size of the data.

The experiment results are given in Figure 5. The shared causal SNPs are presented in the low-rank component and individual-specific SNPs are shown in the sparse component. We take the first three right singular vectors of the recovered low-rank matrix and use them as the coordinate of each study in Figure 6. From Figure 6, it is clear to see that 3 clusters are recovered from 32 diseases/traits:

- 2 social science studies (edu_years and college);
- diastolic blood pressure and systolic blood pressure (DBP and SBP);
- total cholesterol and low density lipoprotein (TC and LDL).

The diseases/traits in each cluster are highly related to each other. We compare the identified causal SNPs by our method on 32 GWAS data with some previous findings. For 3 pairs of diseases/traits that are clustered together, we mainly investigate the shared SNPs that are identified by our method. For two social science related data, our method has detected SNP *rs3789044*, SNP *rs12046747*, and SNP *rs12853561*, which are mapped to genes *LRRN2* and *STK24*, respectively. These were reported in the original article [35] because they have significant P-values (the details are provided in the supplementary materials). However, besides those SNPs with significant P-values, our method has also identified some loci with moderate P-values. SNP *rs2532269*, whose original P-values are 1.01×10^{-4} in edu_years data and 1.11×10^{-4} in college data, is detected as a causal SNP by our method. This SNP was previously reported (P-value $= 2 \times 10^{-11}$) [36] and mapped to the gene *KIAA1267*. This gene is highly connected with Koolen-De Vries syndrome. Koolen-De Vries syndrome is characterized by moderate to severe intellectual disability, hypotonia, friendly demeanor, and highly distinctive facial features, including tall, broad forehead, long face, upslanting palpebral fissures, epicanthal folds, tubular nose with bulbous nasal tip, and large ears [37].

For diastolic blood pressure and systolic blood pressure, the identified SNPs in our experiment are also connected with some previously published genes, such as *ULK4, FGF5* and *C10orf107* [38]. Similarly, some additional loci are identified by the low-rank component. SNP *rs4986172* (original P-values in SBP data and DBP data are 3.09×10^{-5} and 0.0172, respectively), located in the gene *ACBD4*, is detected by the low-rank component. This gene has been associated with high blood pressure in [39].

To illustrate the power of our method in identifying the causal SNPs that are not shared by several diseases/traits, we take the result of bipolar disorder as an example. The SNPs in the result of bipolar disorder can be matched to *ANK3, CACNA1C, SYNE1* and *PBRM1*, which have been confirmed to be associated with bipolar disorder [34]. The detailed results of other diseases/traits can be found in the supplementary materials. Clearly, the experiment results show

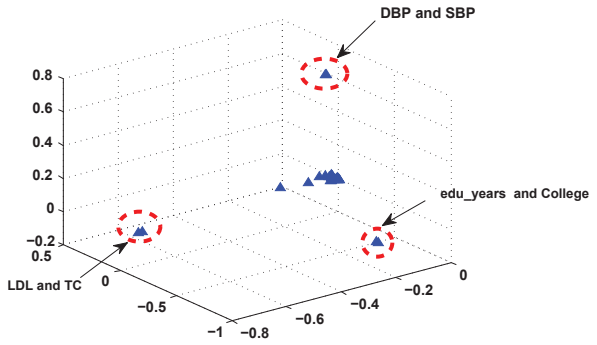

Figure 6. The geometric relationships of all studies using the coordinates derived from the first three right singular vectors of the recovered low-rank matrix.

that not only can our method recognize SNPs with small P-values, but also detect those SNPs with moderate P-values.

4 Conclusion

Finding weak-effect variants to explain the missing heritability of complex diseases is a challenging task and bottlenecked by the available sample size of GWAS. Based on the fact that related diseases/traits tend to co-occur, discovering shared genetic components among related studies becomes a popular way to address this issue. In the last few years, hundreds of GWASs have been carried out. Therefore, it is timely to systematically investigate GWAS data sets to find those shared patterns for comprehensive understanding of the genetic architecture of complex diseases/traits. In this work, we present a novel method for exploring the genetic patterns of complex diseases. We assume that causal SNPs can be divided into two categories: SNPs shared by multiple diseases/traits and SNPs for individual disease/trait. Thus, by modeling the problem as recovering a low-rank component and a sparse component from a noise matrix, we formulate it as a convex optimization problem. To demonstrate the performance of our proposed method, we conducted several simulation studies under different settings. Simulation results show that the proposed method are comparable to the alternative methods in many settings. In the real data studies, we collected 32 large-scale GWAS data sets. We have successively analyzed these data sets via our proposed method and discovered some interesting shared genetic patterns. Many identified variants have been confirmed by other works. To conclude, our proposed method not only possesses a competing power (precision, recall, and F1-score) but also provides easily understood results for better understanding shared genetic architectures of complex diseases/trais.

In this work, we mainly focus on the analysis of summary statistics. The major limitation of our model is that it can only guarantee that the low-rank component does not contain the spurious associations. Depending on the strength of spurious signals, the sparse component may contain some spurious associations. At present, with the development of new technology, more and more supplementary information, such as functional annotation data, structural data, and biochemical data, can be quickly obtained. Therefore, one feasible solution to overcome this limitation is to incorporate these supplementary information into our model. We will tackle this extension in our future work.

Acknowledgments

This work was supported by Georgia State University Deep Grant, Hong Kong Baptist University Strategic Development Fund, Hong Kong Baptist University Research Grant FRG2/14-15/077 and FRG2/14-15/069, and Hong Kong General Research Grant HKBU12202114 and HKBU22302815.

References

[1] J. M. McClellan, E. Susser, and M.-C. King, "Schizophrenia: a common disease caused by multiple rare alleles," *The British Journal of Psychiatry*, vol. 190, no. 3, pp. 194–199, 2007.

[2] A. P. Morris, B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan *et al.*, "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes," *Nature genetics*, vol. 44, no. 9, p. 981, 2012.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[4] E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[5] R. Sachidanandam, D. Weissman, S. Schmidt, J. Kakol, L. Stein, G. Marth, S. Sherry, J. Mullikin, B. Mortimore, D. Willey *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, 2001.

[6] L. Hindorff, H. Junkins, P. Hall, J. Mehta, and T. Manolio, "A catalog of published genome-wide association studies. available at: www.genome.gov/gwastudies." *Accessed January 22, 2015*, 2015.

[7] T. Manolio, F. Collins, N. Cox, D. Goldstein, L. Hindorff, D. Hunter, M. McCarthy, E. Ramos, L. Cardon, A. Chakravarti *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[8] B. Maher, "Personal genomes: The case of the missing heritability." *Nature*, vol. 456, no. 7218, p. 18, 2008.

[9] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery *et al.*, "Common snps explain a large proportion of the heritability for human height," *Nature genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[10] S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell, "Abundant pleiotropy in human complex diseases and traits," *The American Journal of Human Genetics*, vol. 89, no. 5, pp. 607–618, 2011.

[11] S. Vattikuti, J. Guo, and C. C. Chow, "Heritability and genetic correlations explained by common snps for metabolic syndrome traits," *PLoS genetics*, vol. 8, no. 3, p. e1002637, 2012.

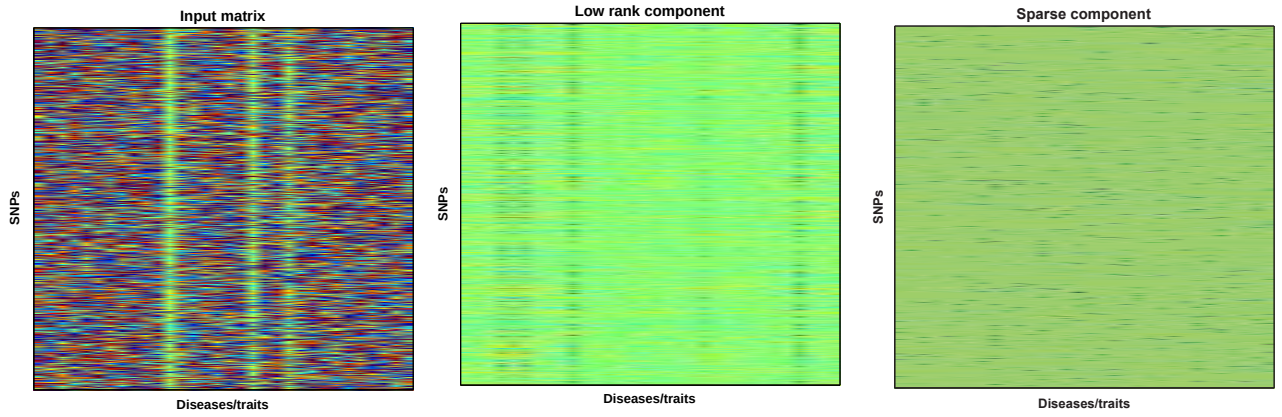

Figure 5. The experiment results on 32 GWASs. The low-rank component (middle panel) and the sparse component (right panel) are recovered by our method.

[12] P. G. Consortium *et al.*, "Genetic relationship between five psychiatric disorders estimated from genome-wide snps," *Nature genetics*, 2013.

[13] Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.*, "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *Lancet*, vol. 381, no. 9875, p. 1371, 2013.

[14] R. A. Fisher, "Statistical methods for research workers," 1934.

[15] I. Goods, "On the weighted combination of significance tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 264–265, 1955.

[16] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr, "The American soldier: adjustment during army life.(Studies in social psychology in World War II, Vol. 1.)." 1949.

[17] F. Mosteller, R. R. Bush, and B. F. Green, *Selected quantitative techniques*. Addison-Wesley, 1970.

[18] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 1, no. 1, pp. 24–45, 2004.

[19] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[20] S. Busygin, O. Prokopyev, and P. M. Pardalos, "Biclustering in data mining," *Computers & Operations Research*, vol. 35, no. 9, pp. 2964–2987, 2008.

[21] K. M. Tan and D. M. Witten, "Sparse biclustering of transposable data," *Journal of Computational and Graphical Statistics*, vol. 23, no. 4, pp. 985–1008, 2014.

[22] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, 2002.

[23] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011.

[24] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[25] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[26] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, p. 1956, 2010.

[27] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *Proceedings of the IEEE International Symposium on Information Theory*, 2010.

[28] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[29] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim, "Robust regression methods for computer vision: A review," *International Journal of Computer Vision*, vol. 6, no. 1, pp. 59–70, 1991.

[30] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, pp. 985–1012, 2009.

[31] J. Gu and J. S. Liu, "Bayesian biclustering of gene expression data," *BMC genomics*, vol. 9, no. Suppl 1, p. S4, 2008.

[32] M. Lee, H. Shen, J. Z. Huang, and J. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010.

[33] N. K. V. J. K. Gupta, S. Singh, "Mtba: Matlab toolbox for biclustering analysis." IEEE, 2013, pp. 94–97.

[34] D. Chung, C. Yang, C. Li, J. Gelernter, and H. Zhao, "GPA: A statistical approach to prioritizing gwas results by integrating pleiotropy and annotation," *PLoS genetics*, vol. 10, no. 11, p. e1004787, 2014.

[35] C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal *et al.*, "Gwas of 126,559 individuals identifies genetic variants associated with educational attainment," *Science*, vol. 340, no. 6139, pp. 1467–1471, 2013.

[36] E. G. G. E. Consortium *et al.*, "Common variants at 6q22 and 17q21 are associated with intracranial volume," *Nature genetics*, vol. 44, no. 5, pp. 539–544, 2012.

[37] D. A. Koolen, J. M. Kramer, K. Neveling, W. M. Nillesen, H. L. Moore-Barton, F. V. Elmslie, A. Toutain, J. Amiel, V. Malan, A. C.-H. Tsai *et al.*, "Mutations in the chromatin modifier gene kansl1 cause the 17q21. 31 microdeletion syndrome," *Nature genetics*, vol. 44, no. 6, pp. 639–641, 2012.

[38] International Consortium for Blood Pressure Genome-Wide Association Studies *et al.*, "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[39] C. Newton-Cheh, T. Johnson, V. Gateva, M. D. Tobin, M. Bochud, L. Coin, S. S. Najjar, J. H. Zhao, S. C. Heath, S. Eyheramendy *et al.*, "Eight blood pressure loci identified by genome-wide association study of 34,433 people of european ancestry," *Nature genetics*, vol. 41, no. 6, p. 666, 2009.