PathGPS: Discover shared genetic architecture using GWAS summary data

Zijun Gao

Marshall Business School, University of Southern California, CA, U.S.A.

email: zijungao@marshall.usc.edu

and

Qingyuan Zhao

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, U.K.

email: qyzhao@statslab.cam.ac.uk

and

Trevor Hastie

Department of Statistics and Department of Biomedical Data Science, Stanford University, CA, U.S.A.

email: hastie@stanford.edu

SUMMARY: The increasing availability and scale of biobanks and "omic" datasets bring new horizons for understanding biological mechanisms. PathGPS is an exploratory data analysis tool to discover genetic architectures using Genome Wide Association Studies (GWAS) summary data. PathGPS is based on a linear structural equation model where traits are regulated by both genetic and environmental pathways. PathGPS decouples the genetic and environmental components by contrasting the GWAS associations of "signal" genes with those of "noise" genes. From the estimated genetic component, PathGPS then extracts genetic pathways via principal component and factor analysis leveraging the low-rank and sparse properties. In addition, we provide a bootstrap aggregating ("bagging") algorithm to improve stability under data perturbation and hyper-parameter tuning. When applied

Biometrics 000, 000–000 DOI: 000
000 0000

to a metabolomics dataset and the UK Biobank, PathGPS confirms several known gene-trait clusters and suggests multiple new hypotheses for future investigations.

Key words: GWAS; Pathway analysis; Structural equation model; Summary data.

1. Introduction

Understanding the biological mechanisms by which genetic variation influences phenotypes is one of the primary challenges in human genetics [Lappalainen and MacArthur, 2021]. Genome-wide association studies (GWAS) have successfully mapped thousands of genetic loci associated with complex human traits. However, it is extremely time-consuming and inefficient to investigate every identified association and validate the function [Visscher et al., 2017]. Moreover, complex traits are usually highly polygenic and are associated with a large number of variants across the genome, each explaining only a small fraction of the genetic variance [Manolio et al., 2009, Shi et al., 2016]. These difficulties have hindered the translation of GWAS findings into drug development and clinical applications [Cano-Gamez and Trynka, 2020].

Recently, studies have revealed that many complex traits are associated with the same genomic loci [Pickrell et al., 2016] and identified many pairs of traits with strong genetic correlation [Solovieff et al., 2013, Bulik-Sullivan et al., 2015, Ning et al., 2020]. This phenomenon indicates that disease-causing variants may cluster into key pathways that drive several diseases at the same time [Boyle et al., 2017]. Motivated by the underlying connection among traits, we aim to use large-scale biobank data containing thousands of phenotypes to aggregate information from correlated complex traits and infer their shared genetic architectures. In this article, we aim to use large-scale biobank data to aggregate the information from various traits and infer their shared genetic architectures.

Our motivating dataset is the UK Biobank, a rich database of genetic and phenotypic information from hundreds of thousands of participants across the UK. Participants are genotyped to capture genome-wide genetic variation at millions of single nucleotide polymorphisms (SNPs). A wide variety of phenotypes are recorded, including biological

measurements, lifestyle indicators, bio-markers in blood, and disease diagnosis. The UK Biobank data provide plenty of opportunities for identifying genetic associations with complex traits.

There are several non-trivial hurdles to recover shared genetic architectures from GWAS data. First, any trait is a product of genetic and environmental influences. The environmental factors can both lead to spurious associations or shadow true genetic signals. Unfortunately, environmental factors are not directly observed in most datasets, making it difficult to isolate the genetic contribution from the environmental influences. Second, the biobank data are usually gathered in multiple batches and are regularly augmented with newly collected data. Therefore, the database includes observations from several cohorts and the summary statistics are derived from multiple population. Third, the set of measured traits in large biobanks is evergrowing and many traits are repeatedly measured in slightly different ways. It is desirable to develop a statistical method that is insensitive to data perturbation and yields consistent statistical conclusions as the dataset continues to be enriched.

[Figure 1 about here.]

In this paper, we develop a new statistical method—PATHway discovery through Genome-Phenome Summary data (PathGPS)—based on a model that assumes most human traits are regulated by one or several genetic or environmental pathways (Figure 1a). PathGPS can generate clusters of genes and traits associated with the same biological pathways (Figure 1b) and addresses the aforementioned challenges. First, by subtracting the empirical covariance of traits computed using "noise" genes (genetic variants showing no or weak associations with the traits) from that using "signal" genes (genetic variants showing strong associations with some traits), PathGPS disentangles genetic mechanisms

from environmental factors. PathGPS then applies principal component analysis (PCA) to the disentangled covariance matrix and provides a low-rank representation of genetic associations. Second, PathGPS can be applied to summary statistics derived from several cohorts as long as the underlying genetic pathways are shared across cohorts. Third, to stabilize PathGPS, we design a novel implementation of the bootstrap aggregation ("bagging") applied to unsupervised learning (Figure 1c). In particular, by resampling the genes, PathGPS obtains a weighted graph which estimates the likelihood that a pair of variables (could be genes or traits) appear in the same pathway. Dimension reduction techniques and clustering algorithms are then applied to visualize this graph and produce clusters.

PathGPS only requires GWAS summary statistics, which can be more easily accessed compared to individual genetic data. An additional benefit is that the computational complexity of our method does not depend on the sample size of the GWAS, once the summary statistics are already produced. Our proposal is motivated by the literature investigating summary statistics for heritability and latent factor characterization [Finucane et al., 2015, Tanigawa et al., 2019, Bulik-Sullivan et al., 2015]. In addition, we draw upon statistical methods with both sparsity and low-rank structures [Zou et al., 2006, Witten et al., 2009, Kaiser, 1958, Jennrich, 2001]. PathGPS also builds on the idea of using bootstrap resamples to reduce the variance of statistical learning methods [Breiman, 2001].

The paper is organized as follows. In Section 2, we introduce the statistical model and the PathGPS algorithm. In Section 3, we investigate the performance of PathGPS in simulated and real datasets and discuss findings using the UK BioBank data. We conclude the paper with more discussion in Section 4.

2. Method

In Section 2.1, we lay out the model characterizing latent pathways. We discuss column space estimation in Section 2.2.1, a preparation step for the gene-trait clustering in Section 2.2.2. In Section 2.3, we propose to use bootstrap aggregation to boost the stability of the proposed clustering algorithm.

2.1 *Structural Equation Model of Latent Pathways*

We describe latent genetic and environmental pathways connecting SNPs and traits using a linear structural equation model (SEM). We start with an individual level model of the SNPs and traits, and then derive the summary statistics from the individual level model.

Suppose there are p SNPs $\boldsymbol{X} = (X_1, \ldots, X_p)$ and q traits $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$. We assume the traits are influenced by the SNPs through r latent genetic mediators $\boldsymbol{M} = (M_1, \ldots, M_r)$. Meanwhile, we assume the traits are also affected by s unobserved environmental mediators $\boldsymbol{m} = (m_1, \ldots, m_s)$. Mathematically, we adopt the linear SEM,

$$\boldsymbol{M} = \boldsymbol{X}\mathbf{U} + \boldsymbol{\varepsilon}^M, \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{M}\mathbf{V}^\top + \boldsymbol{m}\mathbf{W}^\top + \boldsymbol{\varepsilon}^Y, \tag{2}$$

where $\boldsymbol{\varepsilon}^M \in \mathbb{R}^r$, $\boldsymbol{\varepsilon}^Y \in \mathbb{R}^q$ denote zero-mean errors in the mediators $\boldsymbol{M}$ and traits $\boldsymbol{Y}$, respectively, and $\mathbf{U} \in \mathbb{R}^{p\times r}$, $\mathbf{V} \in \mathbb{R}^{q\times r}$, $\mathbf{W} \in \mathbb{R}^{q\times s}$ are coefficient matrices. We assume the errors $\boldsymbol{\varepsilon}^M$, $\boldsymbol{\varepsilon}^Y$ are zero-mean and independent of the SNPs as well as the environmental mediators.

Our goal is to discover genetic pathways: SNPs $\rightarrow$ genetic mediator (latent) $\rightarrow$ traits. Since the genetic mediators are unobserved, we look for clusters of SNPs and traits related to the same underlying genetic pathway. Figure 1 displays an example: there are two genetic pathways: $X_1, X_2 \rightarrow M_1 \rightarrow Y_1, Y_2$ in red and $X_2, X_3 \rightarrow M_2 \rightarrow Y_3$ in blue, and we aim to

uncover the corresponding gene-trait clusters $\{X_1, X_2, Y_1, Y_2\}$ and $\{X_2, X_3, Y_3\}$. In terms of the SEM, let $\mathbf{U}_{\cdot k}$, $\mathbf{V}_{\cdot k}$ be the k-th column of $\mathbf{U}$, $\mathbf{V}$, and Eq. (1) and (2) are equivalent to

$$M_k = \boldsymbol{X}\mathbf{U}_{\cdot k} + \boldsymbol{\varepsilon}_k^M, \quad 1 \le k \le r,$$
$$\boldsymbol{Y} = \sum_{k=1}^{r} M_k \mathbf{V}_{\cdot k}^\top + \sum_{k=1}^{s} m_k \mathbf{W}_{\cdot k}^\top + \boldsymbol{\varepsilon}^Y.$$

The k-th genetic pathway refers to the SNPs' effects on the k-th mediator M_k, denoted by $\boldsymbol{X}\mathbf{U}_{\cdot k}$, and the effect of M_k on the traits $\boldsymbol{Y}$, denoted by $M_k\mathbf{V}_{\cdot k}^\top$. The k-th gene-trait cluster comprises the SNPs and traits with non-zero loadings in $\mathbf{U}_{\cdot k}$ and $\mathbf{V}_{\cdot k}$, respectively.

Our analysis does not operate with the individual level data but instead handles the more readily available summary statistics—gene-trait effect (marginal association) estimates. The estimated marginal associations of gene-trait pairs, denoted by $\hat{\beta}$, are obtained from running simple linear regressions with one trait as the response and one SNP as the predictor. For pre-processing, we use the PLINK software to select (approximately) independent index SNPs. We normalize the SNP vectors $\boldsymbol{X}$ to zero mean and unit variance, and the matrix $\mathbf{X}^\top\mathbf{X}$ is close to an identity matrix. Under the SEM model and the above normalization, the estimated marginal associations of the index SNPs ideally take the form

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}. \tag{3}$$

By plugging Eq. (1), (2) into Eq. (3) and collecting zero-mean environmental mediators $\boldsymbol{m}$, two sources of errors $\boldsymbol{\varepsilon}^M$, $\boldsymbol{\varepsilon}^Y$ in $\mathbf{E} = \mathbf{X}^\top\mathbf{m}\mathbf{W}^\top + \mathbf{X}^\top\mathbf{E}^M\mathbf{V}^\top + \mathbf{X}^\top\mathbf{E}^Y$, the estimated marginal association matrix $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times q}$ satisfies

$$\hat{\boldsymbol{\beta}} = \mathbf{U}\mathbf{V}^\top + \mathbf{E}. \tag{4}$$

In the following, we explain how to use $\hat{\boldsymbol{\beta}}$ to uncover the gene-trait clusters—non-zero loadings in matching column pairs of $\mathbf{U}$ and $\mathbf{V}$.

2.2 *Estimation of Gene-Trait Clusters*

Two biological phenomena facilitate the learning of gene-trait clusters from the summary statistics $\hat{\boldsymbol{\beta}}$. First, the ubiquity of pleiotropy—a single mutation may affect multiple traits—is supported by increasing evidence. Correspondingly, in model (1) and (2), the number of strong genetic pathways is expected to be significantly smaller than the numbers of traits and SNPs collected, i.e., $r \ll p, q$, and thus the product matrix $\mathbf{U}\mathbf{V}^\top \in \mathbb{R}^{p\times q}$ is low-rank (of rank at most r). Second, though the total number of SNPs is colossal, only a small proportion of SNPs are expected to get involved in a certain genetic pathway. In the statistical terminology, the underlying true coefficient matrix $\mathbf{U}$ should consist of sparse columns. In addition, a genetic pathway may only influence a limited number of the traits collected. Therefore, most of the elements in the columns $\mathbf{V}_{\cdot k}$ are anticipated to be zero. The low-rank and sparse structures together imply that the traits and SNPs can be grouped into a few clusters (low-rank property), each containing a relatively small number of SNPs and traits (sparse property).

We discuss our proposal PathGPS leveraging the low-rank and sparse structures. We start with estimating the low-dimensional column spaces of $\mathbf{U}$ and $\mathbf{V}$ in Section 2.2.1. In Section 2.2.2, we discuss methods to find $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$ with sparse columns in the estimated column spaces $\widehat{\mathsf{colsp}}(\mathbf{U})$, $\widehat{\mathsf{colsp}}(\mathbf{V})$, respectively. Finally, we construct a gene-trait cluster for each column pair $(\hat{\mathbf{U}}_{\cdot k}, \hat{\mathbf{V}}_{\cdot k})$, $1 \le k \le r$, corresponding to the k-th genetic pathway. The whole procedure is summarized in Algorithm 1.

Algorithm 1: PathGPS

Input: Estimated marginal association matrices $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times q}$ (signal SNPs) and $\hat{\boldsymbol{\beta}}^0 \in \mathbb{R}^{p^0\times q}$ (noise SNPs), number of latent mediators r.

Initialization: List $\mathcal{L} = \emptyset$ of gene-trait clusters.

1. Remove environmental confounders: compute the truncated eigen-decomposition with r components

$$\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}} - \frac{p}{p^0}\hat{\boldsymbol{\beta}^0}^\top \hat{\boldsymbol{\beta}}^0 = \tilde{\mathbf{V}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top,$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{q\times r}$ is orthonormal, $\tilde{\mathbf{D}} \in \mathbb{R}^{r\times r}$ is diagonal. Let $\tilde{\mathbf{U}} = \hat{\boldsymbol{\beta}}\tilde{\mathbf{V}}$.

2. Apply Varimax or Promax to $\tilde{\mathbf{V}}$ and find the transformation matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$ such that the transformed columns have high loadings on a few coordinates but near-zero values on the rest. Let $\hat{\mathbf{V}} = \tilde{\mathbf{V}}\mathbf{R}$, $\hat{\mathbf{U}} = \tilde{\mathbf{U}}(\mathbf{R}^{-1})^\top$. We further truncate the near-zero elements in $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ to zero.

3. **for** k *from* 1 *to* r **do**

   Define the k-th gene-trait cluster as

$$C_k := \left\{X_i : \hat{\mathbf{U}}_{ik} \neq 0\right\} \cup \left\{Y_j : \hat{\mathbf{V}}_{jk} \neq 0\right\}.$$

   Add the k-th cluster into the cluster list $\mathcal{L} \leftarrow \mathcal{L} \cup \{C_k\}$.

end

Output: List $\mathcal{L}$ of gene-trait clusters.

2.2.1 *Column Space Estimation.* Matrices $\mathbf{U}$, $\mathbf{V}$ in Eq. (4) are not identifiable without further assumptions. In fact, for any invertible matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$, define $\mathbf{V}' = \mathbf{V}\mathbf{R}$, $\mathbf{U}' = \mathbf{U}\,(\mathbf{R}^{-1})^\top$, then $\mathbf{U}'\mathbf{V}'^\top = \mathbf{U}\mathbf{V}^\top$. However, the column spaces $\mathsf{colsp}(\mathbf{U})$ and $\mathsf{colsp}(\mathbf{V})$ are uniquely defined. Therefore, we start with estimating $\mathsf{colsp}(\mathbf{U})$ and $\mathsf{colsp}(\mathbf{V})$.

We use a baseline method to demonstrate the challenge posed by the presence of environmental influences in estimating the column spaces. Provided with the true number of latent mediators r, arguably the most straightforward column space estimator, which we call "simple SVD"in the following, consists of two steps,

(1) Compute the singular value decomposition (SVD) of $\hat{\boldsymbol{\beta}}$ and take the top r singular vectors $\hat{\mathbf{U}}_{\text{SVD}}$, $\hat{\mathbf{V}}_{\text{SVD}}$;

(2) Let $\widehat{\mathsf{colsp}}(\mathbf{U}) = \mathsf{colsp}(\hat{\mathbf{U}}_{\text{SVD}})$, $\widehat{\mathsf{colsp}}(\mathbf{V}) = \mathsf{colsp}(\hat{\mathbf{V}}_{\text{SVD}})$.

However, by using Eq. (4) and taking expectation over errors $\boldsymbol{\varepsilon}^M$ and $\boldsymbol{\varepsilon}^Y$, the estimated marginal associations satisfy

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}\right] = \mathbf{V}\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top + \frac{p}{n}\left(\mathbf{W}\boldsymbol{\Sigma}_m\mathbf{W}^\top + \mathbf{V}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^M}\mathbf{V}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^Y}\right). \tag{5}$$

The decomposition suggests that the column space of $\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}$, concentrated around its expectation $\mathbb{E}\left[\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}\right]$, is contaminated by the environmental variation $\mathbf{W}\boldsymbol{\Sigma}_m\mathbf{W}^\top$ and the response error covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^Y}$. The contamination is serious when the ratio p/n is not ignorable. As a consequence, the "simple SVD"will mistake environmental influences for genetic components.

To separate the genetic and environmental components in Eq. (5), we propose a method using noise SNPs. The idea is to use the estimated marginal associations $\hat{\boldsymbol{\beta}}^0$ of p^0 noise SNPs, which are not (or only weakly) associated with traits, to learn the non-genetic structure and remove it from the estimated marginal associations $\hat{\boldsymbol{\beta}}$ of signal SNPs. Explicitly, the marginal associations of the noise SNPs, satisfy

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}^0}^\top\hat{\boldsymbol{\beta}}^0\right] = \frac{p^0}{n}\left(\mathbf{W}\boldsymbol{\Sigma}_m\mathbf{W}^\top + \mathbf{V}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^M}\mathbf{V}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^Y}\right). \tag{6}$$

Compared to Eq. (5), the expectation of $\hat{\boldsymbol{\beta}^0}^\top\hat{\boldsymbol{\beta}}^0$ of the noise SNPs does not contain any genetic component and is a scalar multiple of the environmental effects. As a direct corollary of Eq. (5) and (6),

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}} - \frac{p}{p^0}\hat{\boldsymbol{\beta}^0}^\top\hat{\boldsymbol{\beta}}^0\right] = \mathbf{V}\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top. \tag{7}$$

Eq. (7) demonstrates the environmental influences can be removed by subtracting a scalar multiple of $\hat{\boldsymbol{\beta}^0}^\top\hat{\boldsymbol{\beta}}^0$ from $\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}$. Motivated by this cancellation of the non-genetic influences, we introduce the differencing estimator of $\mathsf{colsp}(\mathbf{V})$:

(1) Compute the truncated eigen-decomposition of $\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}} - \frac{p}{p^0}\hat{\boldsymbol{\beta}^0}^\top\hat{\boldsymbol{\beta}^0}$ with r components. Denote the matrix of r eigenvectors by $\tilde{\mathbf{V}}$. Let $\widehat{\mathsf{colsp}}(\mathbf{V}) = \mathsf{colsp}(\tilde{\mathbf{V}})$.

(2) As for $\mathsf{colsp}(\mathbf{U})$, we suggest

$$\tilde{\mathbf{U}} = \hat{\boldsymbol{\beta}}\tilde{\mathbf{V}},\ \widehat{\mathsf{colsp}}(\mathbf{U}) = \mathsf{colsp}(\tilde{\mathbf{U}}). \tag{8}$$

2.2.2 *Gene-trait clustering.* In this section, we find sparse matrices in the estimated column spaces in Section 2.2.1 and construct gene-trait clusters from the non-zero loadings.

Let $\tilde{\mathbf{U}}$, $\tilde{\mathbf{V}}$ be two candidate matrices from $\widehat{\mathsf{colsp}}(\mathbf{U})$, $\widehat{\mathsf{colsp}}(\mathbf{V})$, respectively. We aim to find a transformation matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$ such that the columns of $\hat{\mathbf{U}} := \tilde{\mathbf{U}}(\mathbf{R}^{-1})^\top$, $\hat{\mathbf{V}} := \tilde{\mathbf{V}}\mathbf{R}$ are sparse. The task aligns with a number of readily available methods from factor analysis with an focus on sparsity. In particular, we adopt two commonly used approaches summarized below.

- *Varmiax* [Kaiser, 1958]. We start from an orthonormal matrix $\tilde{\mathbf{V}}$ and solve for a rotation matrix $\mathbf{R}$ to maximize the variances of squared loadings of $\tilde{\mathbf{V}}^\top\mathbf{R}$'s columns. The resulting $\hat{\mathbf{V}}$ tend to have many small loadings and we set those values to zero. Finally, we let $\hat{\mathbf{U}} = \tilde{\mathbf{U}}(\mathbf{R}^{-1})^\top = \tilde{\mathbf{U}}\mathbf{R}$ and again set small values in $\hat{\mathbf{U}}$ to zero.
- *Promax* [Hendrickson and White, 1964]. Promax first applies the above Varimax to $\tilde{\mathbf{V}}$, and then rotates the orthogonal columns of Varimax to a least squares fit. The approach relaxes the orthonormal restriction of $\mathbf{R}$ in Varimax and thus the loadings in $\hat{\mathbf{V}}$ are pushed further apart. We let $\hat{\mathbf{U}} = \tilde{\mathbf{U}}(\mathbf{R}^{-1})^\top$ and truncate small values in $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$ to zero.

Finally, provided with sparse column estimators $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$, we loop over r column pairs $(\hat{\mathbf{U}}_{\cdot k}, \hat{\mathbf{V}}_{\cdot k})$ and assign the traits and genes with non-zero loadings into a cluster. Details are summarized in Algorithm 1. In the upcoming section, we will build upon Algorithm 1 and enhance its stability by bootstrap aggregation.

2.3 *Bootstrap Aggregation of PathGPS*

Several issues may undermine the reliability of Algorithm 1. First, in modern biobanks, the set of measured traits is evergrowing. Many traits are repeatedly measured in slightly different ways. It is desirable to obtain stable results if the traits are slightly perturbed. Second, in the data preprocessing procedures, we use an external SNP dataset to select signal and noise index SNPs. We expect to arrive at similar gene-trait clusters if we perturb the index SNP sets, especially the signal SNPs, by a small amount. Third, Algorithm 1 relies on a set of hyper-parameters, such as the number of latent mediators r. The cluster list $\mathcal{L}$ should be robust to the selection of hyper-parameters.

We propose a bootstrap aggregation (bagging) approach to stabilize the pipeline and make the results more replicable by perturbing the entire procedure many times and then aggregating the results. In the following, we discuss the two components of the bagging procedure: SNP bootstrapping (Section 2.3.1) and the aggregation method via a co-appearance graph (Section 2.3.2). The full bagging procedure is summarized in Algorithm 2.

2.3.1 *SNP Bootstrapping.*

Motivated by [Breiman, 2001], we bootstrap the SNPs used by Algorithm 1. In each trial, we resample the same number of signal SNPs with replacement and obtain bootstrapped signal estimated marginal associations $\hat{\boldsymbol{\beta}}^b$. We then apply Algorithm 1 to $\hat{\boldsymbol{\beta}}^b$ and $\hat{\boldsymbol{\beta}}^0$ and arrive at a cluster list $\mathcal{L}_b$. We repeat from the resampling B times and obtain a collection of cluster lists $\{\mathcal{L}_b\}$. The left panel of Figure 2 describes an example of the bootstrap process.

[Figure 2 about here.]

2.3.2 *Co-appearance Graph Aggregation.*

Based on the multiple gene-trait clusters $\{\mathcal{L}_b\}$ generated by the SNP bootstrapping, we propose to aggregate the cluster lists using a co-

appearance graph. Consider a graph whose nodes denote SNPs and traits. For two nodes v_i and v_j, we define the weight for the edge connecting v_i and v_j (called the co-appearance frequency in the following)

$$w_{ij} := \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{L}_b|}\sum_{C_k\in\mathcal{L}_b}\mathbb{1}_{\{v_i\in C_k\}}\mathbb{1}_{\{v_j\in C_k\}}, \tag{9}$$

where C_k denotes the gene-trait cluster in the list $\mathcal{L}_b$ obtained from the b-th bootstrap sample. The right panel of Figure 2 displays an example of the co-appearance graph. If two nodes always show up in the same cluster, the pair will have a high co-appearance frequency (9), and we are more confident about the connection of the pair.

The co-appearance graph is convenient for downstream clustering and visualization. One option is using t-SNE [Hinton and Roweis, 2002] or UMAP [McInnes et al., 2018] to find low-dimensional embeddings. The embeddings can be further used to visualize genes and traits that are closely connected in the co-appearance graph. The representations can also be fed to various clustering methods based on feature vectors like k-means. Alternatively, we can directly use graph clustering methods, such as spectral clustering and label propagation, to obtain gene-trait clusters.

3. Results

In Section 3.1, we generate simulated datasets following the SEM (1), (2). We demonstrate the differencing estimator's performance in estimating the column space in Section 3.1.1. We showcase that bagging enhances the stability of PathGPS in Section 3.1.2. In Section 3.2, we report the findings from applying PathGPS to the metabolomics data and the UK Biobank data.

3.1 *Simulated Datasets*

Algorithm 2: Bagged PathGPS

Input: Marginal association estimate matrix of signal SNPs $\hat{\boldsymbol{\beta}}$, marginal association estimate matrix of noise SNPs $\hat{\boldsymbol{\beta}}^0$, number of latent mediators r, number of bootstrap trials B.
Initialization: A set of gene-trait cluster list $\mathcal{S} = \emptyset$.
1.**for** b *from* 1 *to* B **do**
  a. Resample the same number of signal genes with replacement and obtain $\hat{\boldsymbol{\beta}}^b$.
  b. Apply Algorithm 1 with inputs $\hat{\boldsymbol{\beta}}^b$, $\hat{\boldsymbol{\beta}}^0$, r, and get a cluster list $\mathcal{L}_b$. Update the set of lists $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{L}_b\}$.
end
2. Compute the co-appearance frequency (9) using $\mathcal{S}$ for all signal SNPs and traits.
3. Use t-SNE/UMAP to find low-dimensional embeddings and cluster. Alternatively, run graph clustering algorithms on the co-appearance graph to get gene-trait clusters. Denote the final list of clusters by $\mathcal{L}$.
Output: Gene-trait cluster list $\mathcal{L}$.

3.1.1 *Column Space Estimation.* We construct simulation datasets following the SEM (1), (2). We consider $n = 2000$ individuals, $p = 100$ signal SNPs, $p^0 = 400$ noise SNPs, $q = 100$ traits, $r = 4$ latent genetic mediators, and $s = 2$ latent environmental mediators. The number of latent genetic mediators are significantly smaller than the number of index SNPs and the number of traits. The signal/noise SNPs, latent environmental mediators, and the random errors $\boldsymbol{\varepsilon}^M$, $\boldsymbol{\varepsilon}^Y$ are independent standard Gaussian random variables. As for coefficient matrices, we first generate elements of the coefficient matrices $\mathbf{U}$, $\mathbf{V}$ uniformly from $[-1, 1]$, and then randomly set 80% of the entries to zero to create sparse matrices $\mathbf{U}$, $\mathbf{V}$. We also generate elements of the environmental mediators' coefficient matrix $\mathbf{W}$ uniformly from $[-4, 4]$. We adjust the magnitudes of the environmental influence, the errors in the mediators $\boldsymbol{\varepsilon}^M$, and the errors in the traits $\boldsymbol{\varepsilon}^Y$ so that the proportion of the environmental mediators' variance $\mathsf{Var}(\boldsymbol{m}\mathbf{W})/\mathsf{Var}(\boldsymbol{\varepsilon}')$ (environmental factor strength) varies from 10% to 90%, while the total variance of the non-genetic component stays the same. We compare two column space estimators: the differencing estimator in Algorithm 1

and the "simple SVD"estimator. We measure the performance of column space estimators by the column space distance: let $\mathbf{U}_1$ and $\mathbf{U}_2$ be two arbitrary matrices of dimension $p \times r$, and let $\mathsf{colsp}(\mathbf{U}_1)$, $\mathsf{colsp}(\mathbf{U}_2)$ be the corresponding column spaces, respectively, define the column space distance as

$$\mathrm{dist}(\mathsf{colsp}(\mathbf{U}_1), \mathsf{colsp}(\mathbf{U}_2)) := \max_{\boldsymbol{\xi} \in \mathbb{R}^p, \|\boldsymbol{\xi}\|_2 = 1} \|\boldsymbol{P}_{\mathbf{U}_1}\boldsymbol{\xi} - \boldsymbol{P}_{\mathbf{U}_2}\boldsymbol{\xi}\|_2, \tag{10}$$

where $\boldsymbol{P}_{\mathbf{U}_1}$, $\boldsymbol{P}_{\mathbf{U}_2}$ are the projection operators onto the column spaces of $\mathbf{U}_1$, $\mathbf{U}_2$.

In Figure 3, we report the column space distances (10) of the two methods under different levels of environmental influences. The performance of the "simple SVD"approach deteriorates as the environmental influence increases. The "simple SVD"approach mistakenly counts the leading environmental factor as a genetic influence. In contrast, the proposed column space estimator is robust to the environmental factors. This is because, despite of the magnitude of the environmental influence, the environmental factors are nearly entirely captured by its estimator $\hat{\boldsymbol{\beta}}^{0\top}\hat{\boldsymbol{\beta}}^{0}$ and subtracted from $\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\beta}}$. When the environmental factor strength (the proportion of the environmental mediators' variance $\mathsf{Var}(\boldsymbol{m}\mathbf{W})/\mathsf{Var}(\boldsymbol{e})$ in Eq. (4)) exceeds 75%, the one standard deviation intervals of the column space distance based on the proposed estimator fall strictly below those produced by the "simple SVD"approach. In the supplementary materials, we also compare the clustering results based on the proposed estimator against the "simple SVD". In the supplementary materials, we have included additional simulations for a more comprehensive analysis. In particular, we demonstrate that our proposed method exhibits robustness in the presence of the many weak effects commonly associated with polygenic traits.

[Figure 3 about here.]

3.1.2 *Gene-Trait Clustering.* As in Section 3.1.1, we follow the SEM (1), (2) and consider $n = 500$ individuals, $p = 50$ signal SNPs, $p^0 = 150$ noise SNPs, $q = 30$ traits, $r = 3$ latent genetic mediators, and $s = 2$ latent environmental mediators. In the default setting (*default*), we design sparse coefficient matrices $\mathbf{U}$, $\mathbf{V}$ to have a total of $n_\emptyset = 16$ genes and traits in each cluster. We also enforce each gene and trait to belong to at most one cluster. In addition to the default setting, we consider two variations: a sparse setting (*sparse*) with only $n_\emptyset = 12$ genes and traits per cluster; an overlap setting (*overlap*) where a gene/trait may belong to multiple clusters (multiple membership). Details of the three simulation settings are summarized in Table 1.

We compare several versions of the PathGPS: (a) the one-shot pipeline (baseline) following Algorithm 1 (no further clustering is required, clustering method denoted by "NA"); (b) the clustering without bootstrapping (*one-shot*) following Algorithm 1; (c) the co-appearance clustering with bootstrapping (*bootstrap*) following Algorithm 1 with 200 bootstrap resamples. As for the clustering methods used by the approach *one-shot* and *bootstrap* based on co-appearance graphs, we consider: (a) first learn the low-dimensional embeddings via t-SNE or UMAP, and then cluster the embeddings by k-means; (b) directly apply graph clustering methods: spectral clustering and hierarchical clustering.

We evaluate the performance of the clustering by the minimal clustering error across label permutation below. In particular, let $\mathcal{L}$ be the true list of r clusters defined on set $\mathcal{A}$, and $\hat{\mathcal{L}}$ be an estimate with the same number of clusters , then we define the clustering error

$$\min_{\pi} \frac{1}{r} \sum_{k=1}^{r} \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\mathbb{1}_{\{a \in C_{\pi(k)}, a \notin \hat{C}_k\}} + \mathbb{1}_{\{a \notin C_{\pi(k)}, a \in \hat{C}_k\}} \right), \tag{11}$$

where $\pi : \{1, 2, \ldots, r\} \to \{1, 2, \ldots, r\}$ denotes a permutation over cluster labels. To test

the robustness to the misspecification of hyper-parameters, we input a sequence of hyper-parameters around the true values. For the number of latent genetic pathways r, we provide r' such that $r' - r \in \{-1, 0, 1, 2, 3\}$, which includes the correct specification $r' = 3$. For the number of genes and traits in each channel $n'_\emptyset$, we input hyper-parameters $n'_\emptyset - n_\emptyset \in \{-4, 0, 4, 8, 12\}$, which also includes the correct specification.

In Figure 4, across different simulation settings, the *baseline* achieves the best accuracy at the true hyper-parameters, while the *bootstrap* approach is in general more robust.

[Table 1 about here.]

[Figure 4 about here.]

3.2 *Real datasets*

[Figure 5 about here.]

3.2.1 *Metabolomics Data.* We use the genome-metabolome wide association study in [Kettunen et al., 2016] as the main dataset. The summary statistics are derived from 24925 participants and contain 123 metabolites, 1.3×10^7 passing quality control SNPs. We remove 18 traits with less than 1.5% variance explained according to the Supplementary Table 1 of [Kettunen et al., 2016]. To select approximately independent index SNPs, we apply PLINK to an external dataset [Davis et al., 2017] of 72 metabolites including a large proportion of the traits in the main dataset. We regard 50 index SNPs with at least one significant marginal association as signal SNPs, and 250 index SNPs with no significant marginal associations as noise SNPs. We extract the marginal associations corresponding to the signal and noise SNPs from the main dataset as the input for PathGPS.

We apply PathGPS to a metabolomics dataset (Section 3.2.1) and the UK Biobank (Section 3.2.2), and discuss the gene-trait clusters produced. Preprocessing procedures can

be found in the supplementary materials, including the details of choosing "signal" and "noise" genes. The results, including the lists and visualizations of the gene-trait clusters, are summarized in Figure 5.

For the metabolomics dataset, PathGPS produces 7 clusters which roughly correspond to large high-density lipoprotein (HDL), small HDL, low-density lipoprotein (LDL), intermediate-density lipoprotein (IDL), large very-low-density lipoprotein (VLDL), small VLDL, and non-lipoprotein measurements (Figure 5a). Thus, using genetic data only, PathGPS is able to recover the known taxonomy of circulating metabolites. In the supplementary materials, we provide a comparison of the clusters produced by PathGPS and those of the "simple SVD" method. PathGPS confirms several known causal genes, such as *PLTP* as a regulator of HDL size [Huuskonen et al., 2001] and *PCSK9* as a regulator of LDL cholesterol [Maxwell and Breslow, 2004]. PathGPS also proposes several biological hypotheses that are not as well established, including *RNF111* in relation to HDL [Holmen et al., 2014] and *TM4SF5* in relation to lipid measurements [Choi et al., 2021]. In fact, few gene-trait pairs suggested by PathGPS directly reach the genome-wide significant level after correcting for multiple testing, but the majority of the gene-trait pairs are at least moderately associated. This demonstrates the ability of PathGPS to associate a group of genes with a group of traits, when any single association is not sufficiently strong.

In each trial, we subsample around half of the traits without replacement, and apply PathGPS to the selected subset of traits. We compare the co-appearance weights (co-appearing probabilities) obtained with (Figure 5a4 UMAP) and without (Figure 5a4 Baseline) bootstrap aggregation. Close-to-one co-appearance weights (co-appearing probabilities) indicate the associated pairs always fall into the same cluster and close-to-zero values imply the associated pairs always end up in different clusters. We observe the

histograms of co-appearance weights (co-appearing probabilities) of PathGPS with bagging have sharper spikes around 0 and 1. The bowl-shaped histograms indicate the bagging procedure increases the stability of PathGPS towards trait inclusion.

3.2.2 *UK Biobank Data.* We use the GWAS summary statistics from the UK Biobank data generated by the Neale Lab. The summary statistics are derived from 3.6×10^5 participants of white-British ancestry and contain 1.3×10^7 passing quality control SNPs.

For data preprocessing, we first remove traits with missing female or male summary statistics. The female summary statistics are used for SNP and trait selection, and the male summary statistics are used for downstream gene-trait cluster exploration. We select approximately independent index SNPs based on the female summary statistics using the PLINK software [Purcell et al., 2007]. To eliminate unreliable estimates of genetic associations, we focus on the 175 traits with at least one significant index SNPs at a 5% confidence level. The resulting traits are a combination of lab measurements, disease diagnoses, medication, and a small number of lifestyle habits. Among the index SNPs, we regard the SNPs with at least one significant marginal association test as signal SNPs (1200 in total). We regard the SNPs with no significant marginal association tests as noise SNPs (250 in total). Finally, the estimated marginal associations between the selected traits and signal (noise) SNPs from the male population are used as the input for PathGPS.

PathGPS produces 10 clusters (Figure 5b3), among which 3 are closely related to some diseases (venous thromboembolism, cardiovascular diseases, and type 2 diabetes), and the other 7 contain biometric measurements, such as bone mineral density, immune system, fat-free mass, and skin or hair colors. In the UMAP visualization (Figure 5b2), the edges reflect high (top 350) co-appearances between vertex pairs and may offer insights into disease mechanisms. For instance, our analysis finds the medication simvastatin is closely

related to high cholesterol and cardiovascular diseases (CVD), which is not surprising given that it is widely prescribed to reduce CVD risk [Bibbins-Domingo et al., 2016]. We also find atorvastatin—another drug in the statin family—is highly related to bone mineral density (BMD) and associated traits. This finding is consistent with existing evidence that statin increases BMD [Li et al., 2020, Lupattelli et al., 2004]. In addition, edges connecting monocyte, neutrophil, and lymphocyte to diabetes and asthma diagnosis have high weights, suggesting connections between the immune system and the two common diseases. In particular, diabetes may be related to the immune system through multiple mechanisms: for example, hyperglycemia in diabetes may cause dysfunction of the immune response [Berbudi et al., 2020]. As for asthma, T lymphocytes are critical to the development of asthma [Larché et al., 2003]. The cooccurrence of diabetes and asthma may be attributed to the shared immunological pathways [Torres et al., 2021]. In the supplementary materials, we provide a comparison of the clusters produced by PathGPS and those of the "simple SVD" method.

Regarding the genetic architecture, our analysis confirms many existing discoveries, such as the association between *HERC2* and hair color [Branicki et al., 2011], *PELO* and red blood cells [Mills et al., 2016], and *NME7* and venous thromboembolism [Heit et al., 2012]. We also find less well established biological hypotheses, such as *BCL2* and Atrial fibrillation [Li et al., 2018], *GFI1* and lymphocyte cells [Van der Meer et al., 2010]. The UMAP embedding provides further information beyond the cluster membership. For example, the cluster containing smoking, alcohol, and diabetes is adjacent to the cluster containing CVD, indicating a multifaceted health effect of alcohol consumption and tobacco usage.

4. Discussion

In this article, we propose PathGPS—a promising statistical tool to discover genes and traits sharing latent biological pathways. When applied to the UK Biobank and a metabolomics data, PathGPS not only confirms many established genetic associations but also generates novel biological hypotheses. By grouping diseases with shared biological pathways, PathGPS can enhance the understanding of comorbidities and contribute to the development of comprehensive clinical practices.

We highlight that PathGPS only requires summary statistics and thus can be readily applied to a number of biobank datasets [Chen et al., 2011, Christensen et al., 2012, Avlund et al., 2014]. It is possible that, for certain traits, the underlying genetic pathways differ across sub-populations around the globe. The heterogeneity of genetic pathways can potentially lead to individualized treatments. Therefore, it is of value to compare the output gene-trait clusters of PathGPS applied to various biobank datasets.

Our proposal of using PathGPS with bootstrap aggregation addresses the call of reproducibility research. When scientific findings rely on statistical analysis, the statistical results should be stable under "reasonable" data perturbations [Yu, 2013]. In particular, biobanks and other databses are often regularly augmented by additional measurements and samples, and it would be desirable to obtain consistent conclusions when more data become available.

There are several avenues for future work. First, research shows that the interactions between genes and environment shape human development, and childhood experiences can alter gene expression. So it may be useful to extend the current model to include the interaction of genetic and environmental factors. Second, PathGPS outputs groups of traits associated with the same pathway and it would be of great interest to further investigate

the causal mechanism. Third, given that it is difficult to develop rigorous uncertainty quantification and inference for clustering and unsupervised learning tasks, it would be useful to consider how experiments can be designed to validate or disprove the potential pathways generated by PathGPS. Finally, the PathGPS deliberately selects index SNPs to be distant from each other to ensure their independence, which results in a limited number of such SNPs. To expand the scope and involve a larger number of SNPs, our method would need to be extended to handle dependent SNPs. This would require us to model the covariance matrix of the index SNPs in order to establish connections between marginal associations and the coefficients derived from a full regression.

Acknowledgement

Qingyuan Zhao is supported by the Isaac Newton Trust and EPSRC grant EP/V049968/1. Trevor Hastie are partially supported by grants DMS 2013736 and IIS 1837931 from the (US) National Science Foundation and grant 5R01 EB 001988-21 from the (US) National Institutes of Health.

Supplementary Materials

Web Appendices, Tables, and Figures, referenced in Section 2 and Section 3 are available with this paper at the Biometrics website on Oxford Academic.

Data Availability

The codes used for the analysis in this paper are available at https://github.com/ZijunGao/PathGPS. The datasets http://www.nealelab.is/uk-biobank, [Kettunen et al., 2016], [Davis et al., 2017] used in this paper are open for access.

References

Avlund, K., Osler, M., Mortensen, E. L., Christensen, U., Bruunsgaard, H., Holm-Pedersen, P., Fiehn, N.-E., Hansen, A. M., Bachkati, S. H., Meincke, R. H., et al. (2014). Copenhagen aging and midlife biobank (camb): an introduction.

Berbudi, A., Rahmadika, N., Tjahjadi, A. I., and Ruslami, R. (2020). Type 2 diabetes and its impact on the immune system. *Current diabetes reviews*, 16(5):442–449.

Bibbins-Domingo, K., Grossman, D. C., Curry, S. J., Davidson, K. W., Epling, J. W., García, F. A., Gillman, M. W., Kemper, A. R., Krist, A. H., Kurth, A. E., et al. (2016). Statin use for the primary prevention of cardiovascular disease in adults: Us preventive services task force recommendation statement. *JAMA*, 316(19):1997–2007.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186.

Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pośpiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A., and Kayser, M. (2011). Model-based prediction of human hair color using dna variants. *Human genetics*, 129(4):443–454.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., Neale, B. M., and and (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241.

Cano-Gamez, E. and Trynka, G. (2020). From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11(nil):nil.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. (2011). China kadoorie

biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International journal of epidemiology*, 40(6):1652–1666.

Choi, C., Son, Y., Kim, J., Cho, Y. K., Saha, A., Kim, M., Im, H., Kim, K., Han, J., Lee, J. W., et al. (2021). Tm4sf5 knockout protects mice from diet-induced obesity partly by regulating autophagy in adipose tissue. *Diabetes*, 70(9):2000–2013.

Christensen, H., Nielsen, J. S., Sørensen, K. M., Melbye, M., and Brandslund, I. (2012). New national biobank of the danish center for strategic research on type 2 diabetes (dd2). *Clinical epidemiology*, 4:37–42.

Davis, J. P., Huyghe, J. R., Locke, A. E., Jackson, A. U., Sim, X., Stringham, H. M., Teslovich, T. M., Welch, R. P., Fuchsberger, C., Narisu, N., et al. (2017). Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in finnish men from the metsim study. *PLoS genetics*, 13(10):e1007079.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228.

Heit, J. A., Armasu, S. M., Asmann, Y. W., Cunningham, J. M., Matsumoto, M. E., Petterson, T. M., and De Andrade, M. (2012). A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24. 2 and 9q. *Journal of thrombosis and haemostasis*, 10(8):1521–1531.

Hendrickson, A. E. and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British journal of statistical psychology*, 17(1):65–70.

Hinton, G. and Roweis, S. T. (2002). Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer.

Holmen, O. L., Zhang, H., Fan, Y., Hovelson, D. H., Schmidt, E. M., Zhou, W., Guo, Y., Zhang, J., Langhammer, A., Løchen, M.-L., et al. (2014). Systematic evaluation of coding variation identifies a candidate causal variant in tm6sf2 influencing total cholesterol and myocardial infarction risk. *Nature genetics*, 46(4):345–351.

Huuskonen, J., Olkkonen, V. M., Jauhiainen, M., and Ehnholm, C. (2001). The impact of phospholipid transfer protein (pltp) on hdl metabolism. *Atherosclerosis*, 155(2):269–281.

Jennrich, R. I. (2001). A simple general procedure for orthogonal rotation. *Psychometrika*, 66(2):289–306.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kettunen, J., Demirkan, A., Würtz, P., Draisma, H. H., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A. J., Lyytikäinen, L.-P., Pirinen, M., et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. *Nature communications*, 7(1):1–9.

Lappalainen, T. and MacArthur, D. G. (2021). From variant to function in human disease genetics. *Science*, 373(6562):1464–1468.

Larché, M., Robinson, D. S., and Kay, A. B. (2003). The role of t lymphocytes in the pathogenesis of asthma. *Journal of Allergy and Clinical Immunology*, 111(3):450–463.

Li, G. H.-Y., Cheung, C.-L., Au, P. C.-M., Tan, K. C.-B., Wong, I. C.-K., and Sham, P.-C. (2020). Positive effects of low ldl-c and statins on bone mineral density: an integrated epidemiological observation analysis and mendelian randomization study. *International journal of epidemiology*, 49(4):1221–1235.

Li, Y., Song, B., and Xu, C. (2018). Effects of guanfu total base on bcl-2 and bax expression

and correlation with atrial fibrillation. *Hellenic Journal of Cardiology*, 59(5):274–278.

Lupattelli, G., Scarponi, A. M., Vaudo, G., Siepi, D., Roscini, A. R., Gemelli, F., Pirro, M., Latini, R. A., Sinzinger, H., Marchesi, S., et al. (2004). Simvastatin increases bone mineral density in hypercholesterolemic postmenopausal women. *Metabolism*, 53(6):744–748.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Maxwell, K. N. and Breslow, J. L. (2004). Adenoviral-mediated expression of pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype. *Proceedings of the National Academy of Sciences*, 101(18):7100–7105.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mills, E. W., Wangen, J., Green, R., and Ingolia, N. T. (2016). Dynamic regulation of a ribosome rescue pathway in erythroid cells and platelets. *Cell reports*, 17(1):1–10.

Ning, Z., Pawitan, Y., and Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics*, 52(8):859–864.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7):709–717.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller,

J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495.

Tanigawa, Y., Li, J., Justesen, J. M., Horn, H., Aguirre, M., DeBoever, C., Chang, C., Narasimhan, B., Lage, K., Hastie, T., et al. (2019). Components of genetic associations across 2,138 phenotypes in the uk biobank highlight adipocyte biology. *Nature communications*, 10(1):1–14.

Torres, R. M., Souza, M. D. S., Coelho, A. C. C., de Mello, L. M., and Souza-Machado, C. (2021). Association between asthma and type 2 diabetes mellitus: Mechanisms and impact on asthma controla literature review. *Canadian respiratory journal*, 2021.

Van der Meer, L., Jansen, J., and Van Der Reijden, B. (2010). Gfi1 and gfi1b: key regulators of hematopoiesis. *Leukemia*, 24(11):1834–1843.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Yu, B. (2013). Stability. *Bernoulli*, 19(4):1484–1500.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

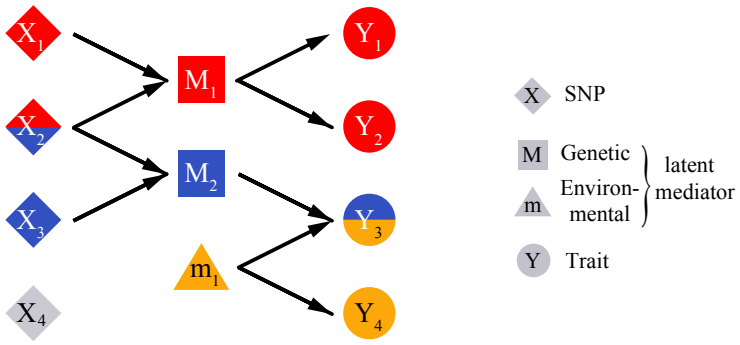

Figure 1: An example of the structural equation model. The example contains 4 SNPs and 4 traits. There are 2 latent genetic mediators: the genetic mediator M_1 is influenced by SNPs X_1, X_2 and affects traits Y_1, Y_2; the genetic mediator M_2 is influenced by SNPs X_2, X_3 and affects trait Y_3. There is one latent environmental mediator m_1 which is independent of the SNPs and affects traits Y_3, Y_4.

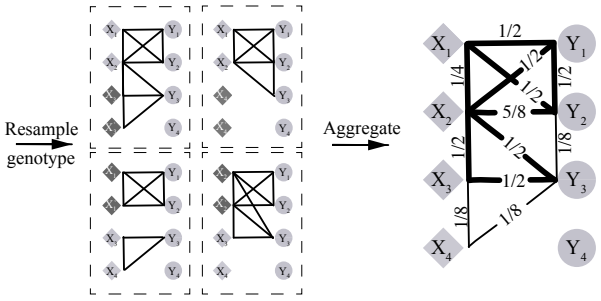

Figure 2: SNP bootstrapping and co-appearance graph. In the left panel, we obtain multiple clusterings based on different sets of bootstrapped SNPs. In the right panel, we display the co-appearance graph obtained based on the four bootstrap samples in the left panel. The weight computation can be found in the supplementary materials.

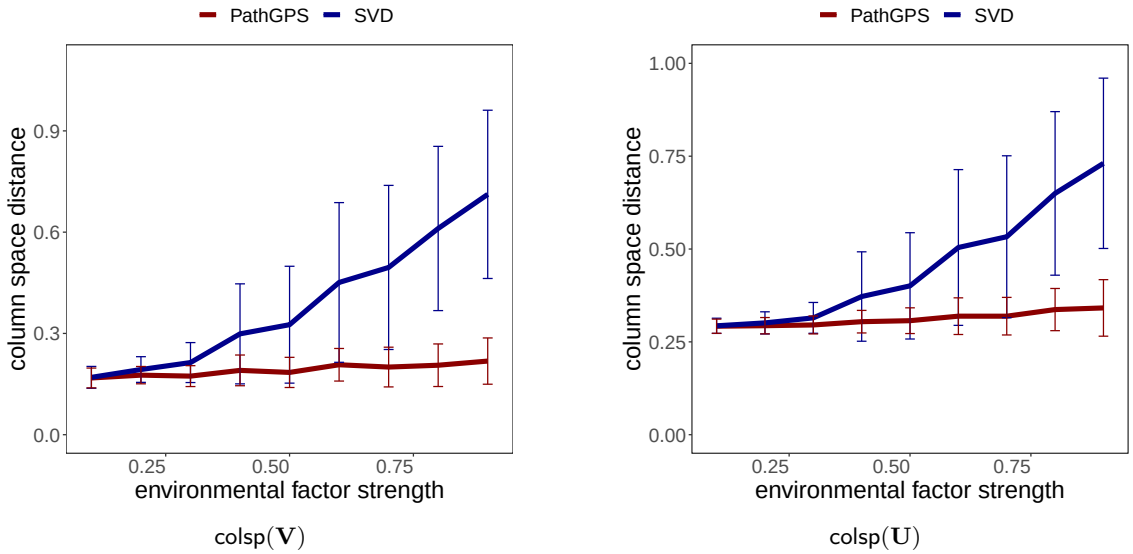

Figure 3: Column space estimation. We compare the "simple SVD"approach (blue) and the differencing estimator in PathGPS (red). We evaluate the estimation performance by the column space distance (10). We vary the proportion of the environmental mediators' variance $\mathsf{Var}(\boldsymbol{m}\mathbf{W})/\mathsf{Var}(\boldsymbol{e})$ (environmental factor strength) in Eq. (4) from 10% to 90%, and display the average column space distances plus and minus one standard deviation for $\mathbf{V}$ in left figure and $\mathbf{U}$ in the right figure. All results are aggregated over 100 trials.

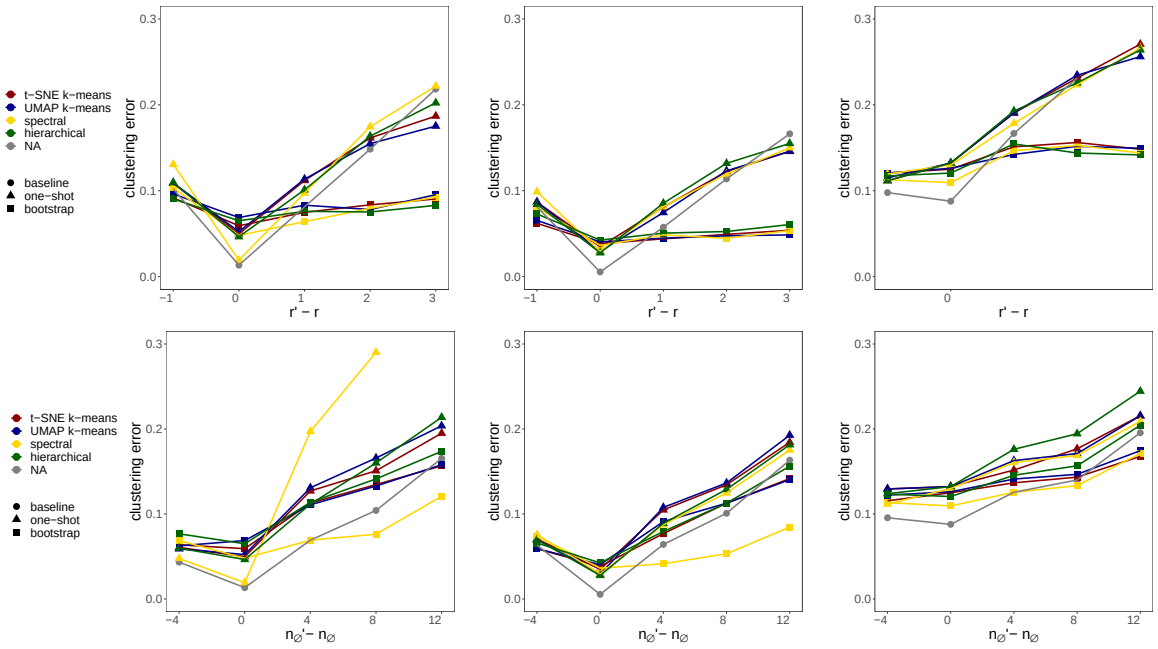

Figure 4: Gene-trait cluster discovery. We compare three variants of PathGPS the one-shot pipeline (baseline in circle) following Algorithm 1 (no further clustering is required, clustering method denoted by "NA"), the clustering without bootstrapping (one-shot in triangle) following Algorithm 1, and the co-appearance clustering with bootstrapping (bootstrap) following Algorithm 1 with 200 bootstrap resamples. We evaluate the estimation performance by the clustering error (11). We input a sequence of r' (first row) and $n'_\emptyset$ (second row) around the true values, and plot the average clustering errors aggregated over 100 trials.

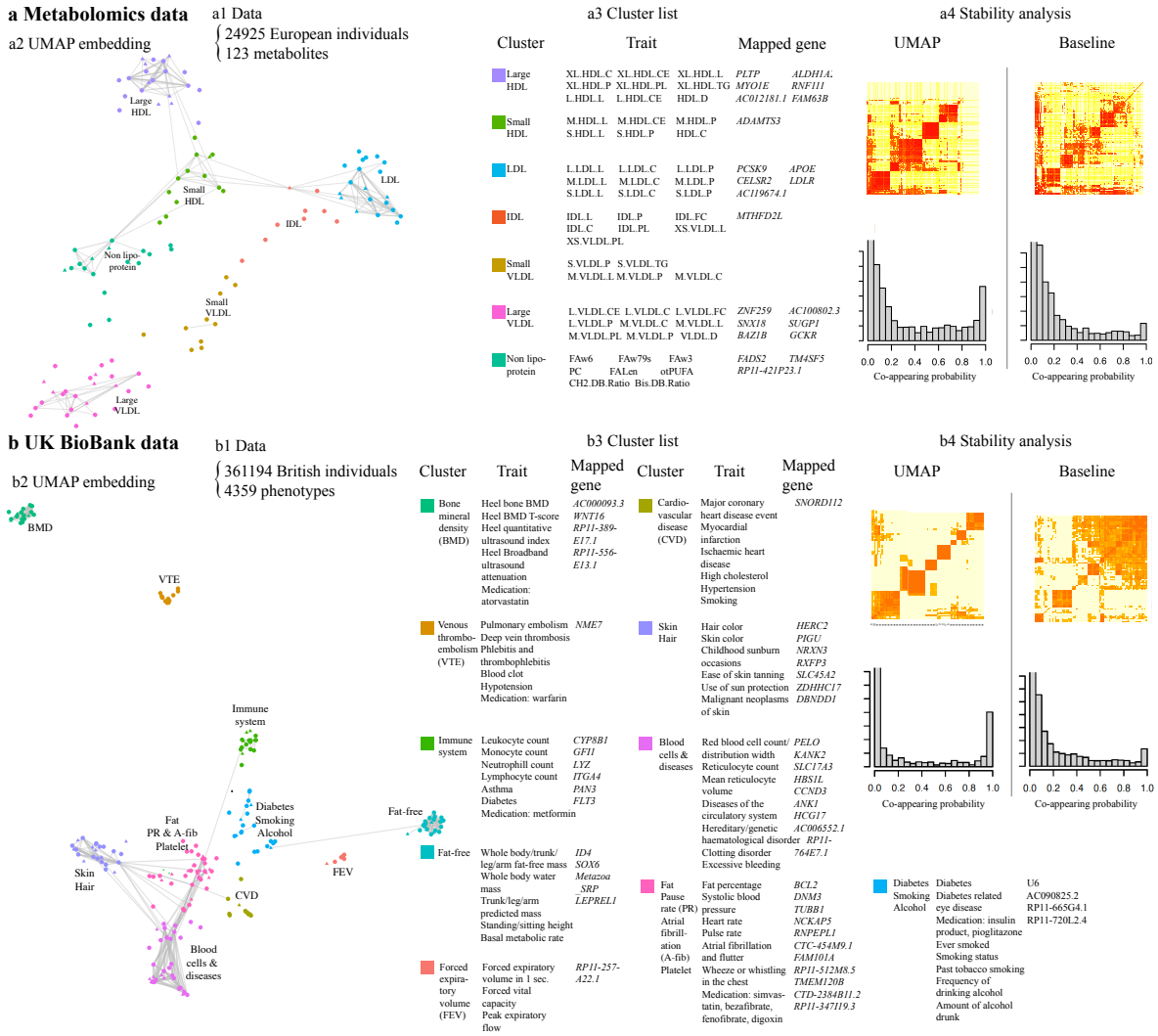

Figure 5: Applications of PathGPS. Panel **A** displays the summary of the metabolomics data (**a1**), the UMAP embeddings of 7 gene-trait clusters produced by PathGPS with co-appearance edge weights (**a2**), and representative traits and mapped genes in each cluster (**a3**). In (**a4**), we subsample traits without replacement, and PathGPS (UMAP) produces more consistent cluster memberships than the baseline method (Figure 1b5). Panel **B** displays the summary of the UK Biobank data (**b1**), the UMAP visualization (**b2**), and representative genes and traits of the 10 clusters produced by PathGPS (**b3**). PathGPS (UMAP) again produces more stable clusters (**b4**). The representative traits and mapped genes in (**a3**) and (**b3**) are selected manually.

setting	number of genes & traits/cluster	multiple membership
default	16	×
sparse	12	×
overlap	16	✓

Table 1: Summary of simulation settings of gene-trait cluster discovery.