Statistical Analysis of Big Data on Pharmacogenomics

Jianqing Fan and Han Liu

Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544, USA

Abstract

This paper discusses statistical methods for estimating complex correlation structure from large pharmacogenomic datasets. We selectively overview several prominent statistical methods for estimating large covariance matrix for understanding correlation structure, inverse covariance matrix for network modeling, large-scale simultaneous tests for selecting significantly differently expressed genes and proteins and genetic markers for complex diseases, and high dimensional variable selection for identify important molecules for understanding molecule mechanisms in pharmacogenomics. Their applications to gene network estimation and biomarker selection are used to illustrate the methodological power. Several new challenges of Big data analysis, including complex data distribution, missing data, measurement error, spurious correlation, endogeneity, and the need for robust statistical methods, are also discussed.

Keywords: Big data, High dimensional statistics, Approximate factor model, Graphical model, Multiple testing, Variable selection, Marginal screening, Robust statistics.

1. Introduction

The ultimate goal of pharmacogenomics is to improve health care based on individual genomic profiles [1, 2, 3, 4, 5, 6, 7]. Together with other factors that may affect drug response — such as diet, age, diseases, lifestyle, environment, and state of health — pharmacogenomics has the potential to facilitate the creation of individualized treatment plan for patients and leads to the overarching goal of personalized medicine. Similar to other areas of human genomics, pharmacogenomics is experiencing an explosion of data, specifically by overloads of omics information (genomes, transcriptomes and other data from cells, tissues and drug effects)[8]. The pharamacogenomics research is entering the era of "Big data" — a term that refers to the explosion of available information. The vast amount of data brings both opportunities and new challenges. Statistical analysis for Big data is becoming increasingly important [9, 10, 11, 12].

This paper selectively overviews several state of the art statistical methods for analyzing large pharmacogenomics data. In particular, we emphasize on the fundamental problems of estimating two types of correlation structures: *marginal* correlation and *conditional* correlation. The former represents the correlation between two variables by ignoring the remaining variables, while the latter represents the correlation between two variables by conditioning on the remaining ones. The marginal correlation can be estimated using the covariance matrix. The conditional correlation can be estimated using the inverse covariance matrix. We introduce several recently developed statistical methods for estimating high dimensional covariance and inverse covariance matrices. We also introduce cutting-edge methods to estimate false discovery proportions for large-scale simultaneous tests, and to select important molecules or SNPs in high dimensional regression models. The former corresponds to finding molecules or SNPs that have significant marginal correlations with biological outcomes, while the latter finds conditional correlations with biomedical responses in presence of many other molecules or SNPs.

The rationale behind this paper is that we believe Big data provides new opportunities for estimating complex correlation structures among a large number of variables. These methods are new to the pharmacogenomics community and have the potential to play important roles in analyzing the next-generation Big data within the pharmacogenomics area.

A notable feature of most methods introduced in this paper is the exploitation of sparsity assumption, which is an essential concept for modern statistical methods applied to high dimensional data. For covariance estimation, we briefly introduce the thresholding approach [13] and its extension called POET (Principal Orthogonal complEment Thresholding) [14, 15] which provides a unified view of most previous methods. For inverse covariance estimation, we mainly focus on introducing two inverse covariance estimation methods named CLIME [16] and TIGER [17], which stand respectively for "Constrained L_1-Minimization for Inverse Matrix Estimation" and "Tuning-Insensitive Graph Estimation and Regression". Both methods estimate the inverse covariance matrix in a column-by-column fashion and achieve the minimax optimal rates of convergence under different norms. For applications, we discuss how the estimated covariance matrix and inverse covariance matrix can be exploited on gene network estimation and large-scale simultaneous tests. We introduce the principal factor approximation (PFA) in [18] for estimating and

controlling false discovery proportions in large-scale simultaneous tests that are dependent. In addition, we introduce penalized likelihood [19], variable screening [20], and their iterated version of screening and selection for high dimensional variable selection. All these problems are important and widely applicable in different subareas of pharmacogenomics. Besides surveying existing methods, we also discuss new challenges and possible solutions of Big data analysis. Topics include modeling nonGaussianity, handling missing data, dealing with measurement error, spurious correlation, endogeneity, and developing robust methods.

The rest of this paper is organized as follows. In §2, we summarize some notations. In §3, we introduce statistical methods for estimating large covariance matrices. In §4, we introduce statistical methods for estimating large inverse covariance matrices. In §5, we apply the results of §3 and §4 to large-scale multiple tests. In §6, we introduce high dimensional variable selection. In §7, we provide case studies of several described methods. In §8, we discuss more on the future directions.

2. Notation

In this section, we summarize some notations. Let $\mathbf{A} = (a_{jk}) \in \mathbb{R}^{d\times d}$, $\mathbf{B} = (b_{jk}) \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$. Denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of $\mathbf{A}$. The inner product of $\mathbf{A}$ and $\mathbf{B}$ is defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$. Define the vector norms: $\|\boldsymbol{v}\|_1 = \sum_j |v_j|$, $\|\boldsymbol{v}\|_2^2 = \sum_j v_j^2$, $\|\boldsymbol{v}\|_\infty = \max_j |v_i|$. We also define matrix operator and elementwise norms: $\|\mathbf{A}\|_2^2 = \lambda_{\max}(\mathbf{A}^T\mathbf{A}), \|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{j,k} a_{jk}^2$. Notation $\mathbf{A} \succ \mathbf{0}$ means that $\mathbf{A}$ is positive definite and $a_n \asymp b_n$ implies there are positive constants c_1 and c_2 independent of n such that $c_1 b_n \le a_n \le c_2 b_n$.

3. Estimating Large Covariance Matrix

Estimating a large covariance or correlation matrix under a small sample size is a fundamental problem which has many applications in phamacogenomics.

For example, in functional genomics, an important problem is to cluster genes into different groups based on the similarities of their microarray expression profiles. One popular measure of the similarity between a pair of genes is the correlation of their expression profiles. Thus, if d genes are being analyzed (with d ranges from the order of ∼1,000 to ∼10,000), a correlation matrix of size $d \times d$ needs to be estimated. Note that $1,000 \times 1,000$ covariance matrices involve already over half a million elements. Yet, the sample size n is of order ∼100, which is significantly smaller than the dimensionality d. Thus, the sample covariance matrix degenerates and needs regularization.

As another example, many multivariate statistical methods that are useful in pharmacogenomics, including principal component analysis (PCA), linear discriminant analysis (LDA), multivariate regression analysis, the Hotelling T^2-statistic, and multivariate likelihood ratio tests as in finding quantitative trait loci based on longitudinal data [21] request the estimation of the covariance matrix as their first step. Thus a reliable estimate of the covariance matrix is of paramount importance.

Moreover, the correlation matrix itself can be used to construct co-expression gene network, which is an undirected graph whose edges indicate marginal correlations among the d genes. The network is built by drawing edges between those pairs of genes whose magnitude of pairwise correlation coefficients exceed a certain threshold. More applications of large covariance matrix estimation will be discussed in §7.

A common key problem underlying all these examples is as follows: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be n independent observations of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$. Without loss of generality, we assume $\mathbb{E}\boldsymbol{X} = \mathbf{0}$. We want to find a reliable estimate of the population covariance matrix $\boldsymbol{\Sigma}^* = \mathbb{E}\boldsymbol{X}\boldsymbol{X}^T$. At the first sight, this problem does not seem to be challenging. In the literature, $\boldsymbol{\Sigma}^*$ was traditionally estimated by the sample covariance matrix

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i^T - \bar{\boldsymbol{x}})^T \quad \text{with} \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i, \tag{3.1}$$

which has many good theoretical properties when the dimensionality d is small. However, in the more realistic settings where the dimensionality d is comparable or even larger than n (i.e., d/n goes to a nonzero constant or infinity), the sample covariance matrix in (3.1) is no longer a good estimate of the population covariance matrix $\boldsymbol{\Sigma}^*$. More details will be explained as follows.

3.1. Inconsistency of Sample Covariance in High Dimensions

We use a simple simulation to illustrate the inconsistency of the sample covariance matrix in high dimensions. Specifically, we sample n data points from a d-dimensional spherical Gaussian distribution $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{\Sigma}^* = \mathbf{I}_d$. Here $\mathbf{I}_d$ is a d-dimensional identity matrix. We consider different setups including $d/n = 2$, $d/n = 1$, $d/n = 0.2$, and $d/n = 0.1$. The results are summarized in Figure 1.

Figure 1 shows the sorted eigenvalues of the sample covariance matrix $\mathbf{S}$ (black curve) with the true eigenvalues (dashed red curve) for $d = 1,000$. By examining these plots, we see that when the dimensionality d is large, the eigenvalues of $\mathbf{S}$ significantly deviate from their true values. In fact, even when n is reasonably large compared with d ($d/n = 0.1$), the result is still not accurate.

This phenomenon can be characterized by random matrix theory. Let $d/n \to \gamma$ with $\gamma \in (0, 1)$ and $\lambda_{\max}(\mathbf{S})$ be the largest eigenvalue of $\mathbf{S}$. It has been shown that $\lambda_{\max}(\mathbf{S})$ converges to $(1 + \sqrt{\gamma})^2$ almost surely [22, 23], i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} \lambda_{\max}(\mathbf{S}) = (1 + \sqrt{\gamma})^2\right) = 1.$$

This result illustrates that, for large d, the largest eigenvalue of the sample covariance matrix $\mathbf{S}$ does not converge to that of the population covariance matrix $\boldsymbol{\Sigma}^*$.

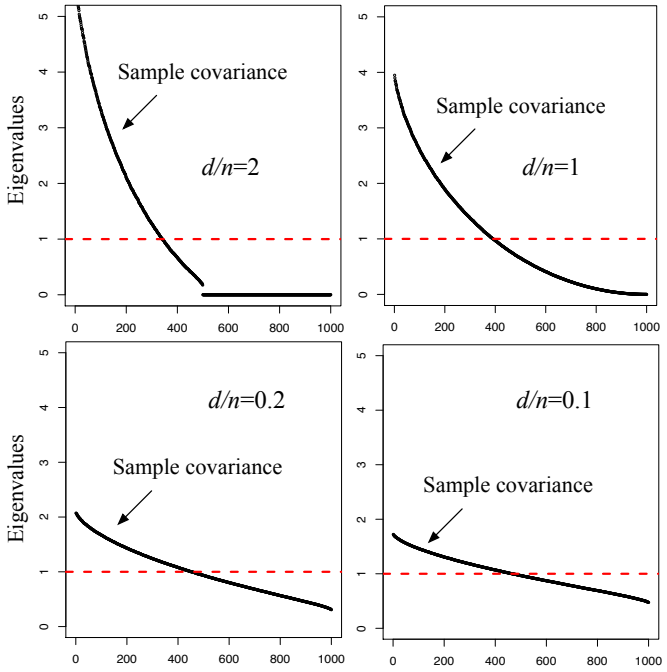


Figure 1: Sorted eigenvalues of the sample covariance matrix $\mathbf{S}$ (black curve) and that of the population covariance matrix $\mathbf{\Sigma}^*$ (dashed red curve). In the simulation, we always use $d = 1,000$ but different ratios of d/n. Subfigures correspond to $d/n = 2$, $d/n = 1$, $d/n = 0.2$ and $d/n = 0.1$.

3.2. Sparse Covariance Matrix Estimation Methods

To handle the inconsistency issue of high dimensional sample covariance matrix, most existing methods assume the population covariance matrix is sparse, i.e., many off-diagonal entries of $\mathbf{\Sigma}^*$ are zero. Such a sparsity assumption is reasonable in many pharmacogenomics applications. For example, in a longitudinal study where variables have a natural order, it is natural to assume that variables are weakly correlated when they are far apart [24]. Under the sparsity assumption, many regularization based statistical methods have been proposed to estimate large covariance matrix [25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. These methods usually only require elementwise thresholding procedures and are computationally efficient. The simplest thresholding estimator is the hard-thresholding estimator [35, 13]

$$\widehat{\mathbf{\Sigma}} = (s_{ij} I(|s_{ij}| \geq \tau_{ij})), \tag{3.2}$$

where $I(\cdot)$ is the indicator function and s_{ij} is the sample covariance between X_i and X_j. Here τ_{ij} is a thresholding parameter. Another example is the adaptive thresholding [29] which takes $\tau_{ij} = \mathrm{SD}(s_{ij})\tau$. Here

$$\mathrm{SD}(s_{ij}) = \sqrt{\frac{\sum_{i=1}^n (x_{ki}x_{kj} - s_{ij})^2}{n}}$$

is the estimated standard error of s_{ij} and τ is a user-specified parameter (e.g., $\sqrt{(2\log d)/n}$). [15] suggests a simplified version $\tau_{ij} = \sqrt{s_{ii}s_{jj}}\tau$ so that the correlation is thresholded at level τ (e.g., $\tau = 0.2$). Estimator (3.2) is not necessarily positive definite. However, when τ is sufficiently large, it is positive definite with high probability.

One additional example is the soft-thresholding estimator [15]:

$$\widehat{\mathbf{\Sigma}} = \Big(\mathrm{sgn}(s_{ij})(|s_{ij}| - \tau_{ij})_+\Big), \tag{3.3}$$

where $(a)_+ = \max\{0, a\}$ and $\mathrm{sgn}(s_{ij}) = I(s_{ij} > 0) - I(s_{ij} < 0)$. It makes the matrix (3.3) positive definite for a wider range of τ_{ij} than the hard-thresholding estimator (3.2) [15]. Although estimators (3.2) and (3.3) suffice for many applications, more general sparse covariance estimation can be found via the following penalized least-squares [36]:

$$\widehat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma}=(\sigma_{jk})}{\mathrm{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2 + \sum_{j\neq k} P_{\lambda,\gamma}(\sigma_{jk}) \right\}, \tag{3.4}$$

where the term $\|\boldsymbol{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2$ measures the goodness of fit and $\sum_{j\neq k} P_{\lambda,\gamma}(\sigma_{jk})$ is a sparsity-inducing penalty that encourages sparsity [19]. The tuning parameter λ controls the bias-variance tradeoff and γ is a possible fine-tune parameter which controls the concavity of the penalty. Popular choices of the penalty function $P_{\lambda,\gamma}(\cdot)$ include the hard-thresholding, soft-thresholding [37, 38], smoothly clipped absolution deviation (SCAD, [19]), and minimax concavity penalties (MCP, [39]). In the following we explain these penalties in more detail.

Penalized least-squares or more generally penalized likelihood is a generally applicable method for variable selection [19, 40]. We introduce it in the context of sparse covariance matrix estimation. First, it is easy to see that the optimization problem in (3.4) decomposes into many univariate subproblems:

$$\widehat{\sigma}_{jk} = \underset{u}{\mathrm{argmin}} \left\{ \frac{1}{2}|s_{jk} - u|^2 + P_{\lambda,\gamma}(u) \right\}, \tag{3.5}$$

whose solution is carefully studied in [41]. Each can be analytically solved for the following commonly used penalty functions:

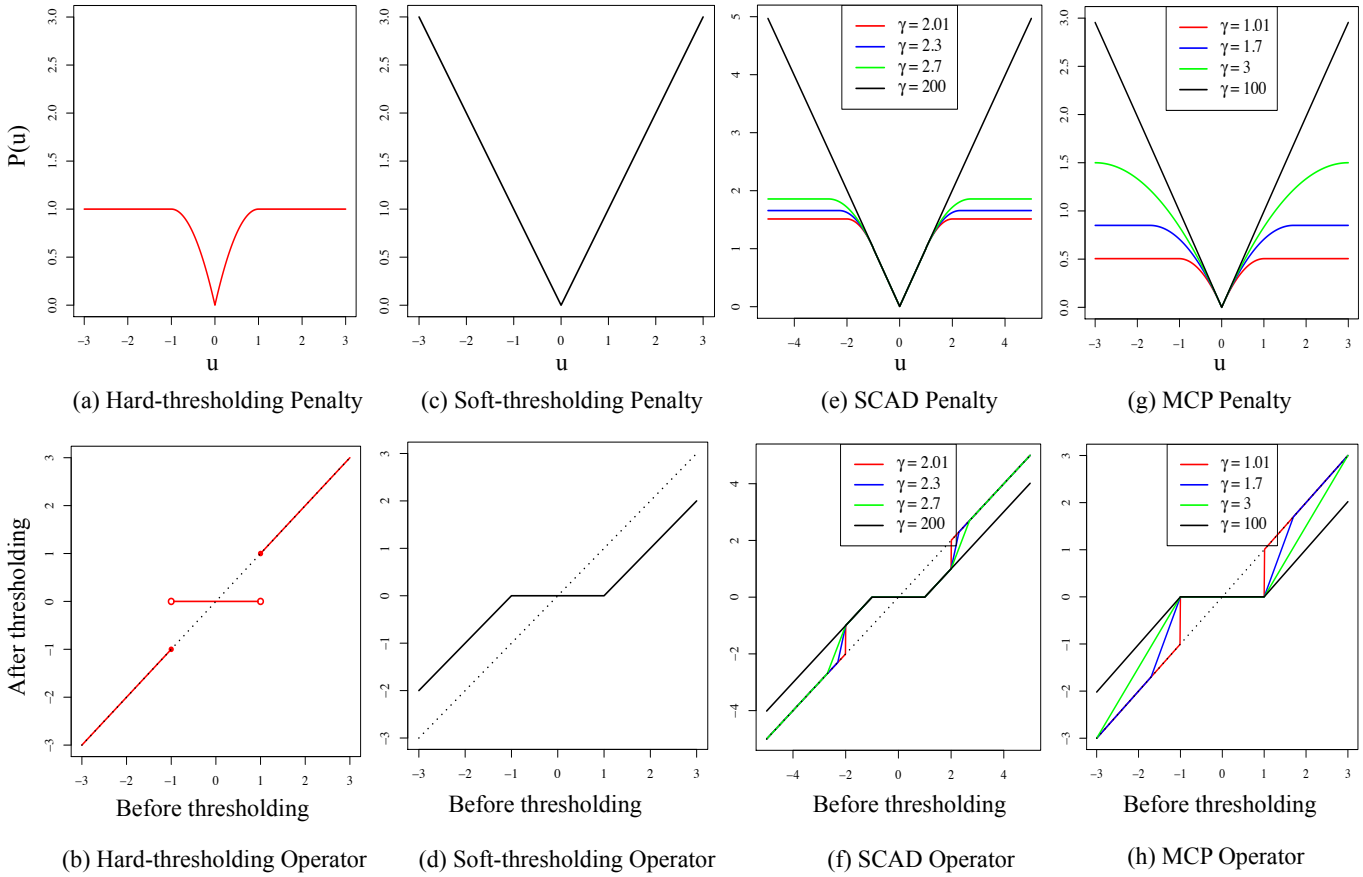
(1) Complexity penalty: $P_{\lambda,\gamma}(u) = \lambda I(u \neq 0)$.
(2) Hard-thresholding penalty: $p_\lambda(u) = \lambda^2 - (\lambda - |u|)_+^2$.
(3) L_1-penalty: $P_{\lambda,\gamma}(u) = \lambda|u|$.
(4) Smoothly clipped absolution deviation penalty (SCAD): with $\gamma > 2$,

$$P_{\lambda,\gamma}(u) = \begin{cases} \lambda|u| & \text{if } |u| \leq \lambda \\ \frac{u^2 - 2\gamma\lambda|u| + \lambda^2}{2(1-\gamma)} & \text{if } \lambda < |u| \leq \gamma\lambda \\ \frac{(\gamma+1)\lambda^2}{2} & \text{if } |u| > \gamma\lambda \end{cases}.$$

(5) minimax concavity penalty (MCP): With $\gamma > 1$,

$$P_{\lambda,\gamma}(u) = \begin{cases} |u| - \frac{u^2}{2\lambda\gamma} & \text{if } |u| \leq \lambda\gamma \\ \frac{\lambda^2\gamma}{2} & \text{if } |u| > \lambda\gamma \end{cases}.$$

Note that for problem (3.5) both the complexity penalty and the hard-thresholding penalty produce the same solution, which is the hard-thresholding rule. Figure 2 visualizes these penalty functions for $\lambda = 1$. All penalty functions are folded concave, but L_1-penalty is also convex. The parameter γ in SCAD and MCP controls the degree of concavity. From Figure 2(e) and Figure 2(g), we see that a smaller value of γ results in more concave penalties. When γ becomes larger, the SCAD and



(a) Hard-thresholding Penalty (c) Soft-thresholding Penalty (e) SCAD Penalty (g) MCP Penalty

(b) Hard-thresholding Operator (d) Soft-thresholding Operator (f) SCAD Operator (h) MCP Operator

Figure 2: Visualization of the penalty functions (above) and the corresponding thresholding operators (below) of Hard-thresholding, Soft-thresholding, SCAD and MCP penalties. In all cases, $\lambda = 1$. For SCAD and MCP, different values of γ are chosen as shown in graphs (Best viewed in color).

MCP converge to the L_1-penalty. MCP is a generalization of the hard-thresholding penalty which corresponds to $\gamma = 1$.

With the aforementioned thresholding penalties, the closed-form solution to the penalized least-squares (3.5) can be found. The complexity and hard-thresholding penalties give hard-thresholding rule (3.2), and the L_1-penalty yields the soft-thresholding rule (3.3). The SCAD thresholding operator is given by

$$\widehat{\sigma}_{jk} = \begin{cases} 0 & \text{if } |s_{jk}| \le \lambda \\ \mathrm{sgn}(s_{jk})(|s_{jk}| - \lambda) & \text{if } \lambda < |s_{jk}| \le 2\lambda \\ \dfrac{(\gamma-1)s_{jk} - \mathrm{sgn}(s_{jk})\gamma\lambda}{\gamma - 2} & \text{if } 2\lambda < |s_{jk}| \le \gamma\lambda \\ s_{jk} & \text{if } |s_{jk}| > \gamma\lambda \end{cases},$$

and the MCP thresholding rule is

$$\widehat{\sigma}_{jk} = \begin{cases} 0 & \text{if } |s_{jk}| \le \lambda \\ \mathrm{sgn}(s_{jk})\left(\frac{|s_{jk}|-\lambda}{1-1/\gamma}\right) & \text{if } \lambda < |s_{jk}| \le \gamma\lambda \\ s_{jk} & \text{if } |s_{jk}| > \gamma\lambda \end{cases}.$$

When $\gamma \to \infty$, the last two operators become the soft-thresholding operator.

How shall we choose among these thresholding operators in applications? We provide rough insights and suggestions on this. From Figure 2(b), we see that the hard-thresholding operator does not introduce extra bias for the large nonzero entries. However, it is highly discontinuous. Unlike the hard-thresholding operator, the soft-thresholding operator in Figure 2(d) is continuous. However, it introduces biases for the nonzero entries. From Figure 2(f) and Figure 2(h), we see that the SCAD and MCP thresholding operators are both continuous and do not introduce extra bias for nonzero entries with large absolute values. In applications, we always recommend to use either SCAD or MCP thresholding since they combine the advantages of both hard- and soft-thresholding operators.

3.3. Positive Definite Sparse Covariance Matrix Estimation

The above thresholding methods, though simple, do not guarantee the positive definiteness of the estimated covariance matrix. This may cause trouble in downstream analysis such as evaluating the predictive likelihood or linear discriminant analysis. To handle this problem, we introduce a covariance estimation method named EC2 (Estimation of Covariance with Eigenvalue Constraints) which explicitly enforces the positive definiteness constraint [42].

EC2 solves the following constrained optimization

$$\widehat{\mathbf{\Sigma}} = \underset{\lambda_{\min}(\mathbf{\Sigma}) \ge \tau}{\mathrm{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{S} - \mathbf{\Sigma}\|_{\mathrm{F}}^2 + \sum_{j \ne k} P_{\lambda,\gamma}(\sigma_{jk}) \right\}, \tag{3.6}$$

where $\tau > 0$ is a pre-specified tuning parameter that controls the smallest eigenvalue of the estimated covariance matrix $\widehat{\mathbf{\Sigma}}$. Compared with (3.4), the optimization problem in (3.6) is not decomposable due to the constraint $\lambda_{\min}(\mathbf{\Sigma}) \geq \tau$. [42] proves that the optimization problem in (3.6) is convex when the penalty function is convex and develops an efficient ISP (Iterative Soft-thresholding and Projection) algorithm to solve it. More details about ISP can be found in [42]. A similar algorithm is also proposed in [43].

Other sparse covariance matrix estimation methods with positive definiteness constraints include [28] and [44]. Both use a log-determinant function to enforce the positive definiteness of the estimated covariance matrix. Their main difference is that [44] adopts a convex least square formulation, while [28] adopts the penalized Gaussian log-likelihood approach.

3.4. Theory of Sparse Covariance Matrix Estimation

Under the sparsity assumption, the above sparse covariance estimators have good theoretical properties. In particular, for the positive definite sparse covariance matrix estimator $\widehat{\mathbf{\Sigma}}$ defined in (3.6), it has been proven by various authors (For more details, see [42]) that

$$\sup_{\mathbf{\Sigma}^* \in \mathcal{M}_k} \mathbb{E}\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_2 \asymp k\sqrt{\frac{\log d}{n}},$$

where $\mathcal{M}_k$ represents the model class

$$\mathcal{M}_k = \left\{\mathbf{\Sigma} : \ \mathbf{\Sigma} \succ \mathbf{0} \text{ and } \max_{1\leq j\leq d}\sum_{\ell=1}^{d} I(\mathbf{\Sigma}_{j\ell} \neq 0) \leq k\right\}. \tag{3.7}$$

This rate of convergence is minimax optimal. Similar results also hold for different types of thresholding estimators [36] defined in (3.4). Using the triangle inequality, we get

$$\left|\mathbb{E}\lambda_{\max}(\widehat{\mathbf{\Sigma}}) - \lambda_{\max}(\mathbf{\Sigma}^*)\right| \leq \mathbb{E}\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_2 \asymp k\sqrt{\frac{\log d}{n}}.$$

This result implies that $\lim_{n\to\infty}\mathbb{E}\lambda_{\max}(\widehat{\mathbf{\Sigma}}) = \lambda_{\max}(\mathbf{\Sigma}^*)$. Thus the inconsistency issue of the sample covariance matrix discussed in §3.1 is avoided.

3.5. POET: New Insight on Large Covariance Matrix Estimation

All the aforementioned methods assume that $\mathbf{\Sigma}^*$ is sparse. Though this assumption is reasonable for many pharmacogenomics applications, it is not always appropriate. For example, all the genes from the same pathway may be co-regulated by a small amount of regulatory factors, which makes the gene expression data highly correlated; when genes are stimulated by cytokines, their expressions are also highly correlated. The sparsity assumption is obviously unrealistic in these situations. To solve this problem, we introduce the POET method in recent paper with discussion [15], which provides an integrated framework for combining latent factor analysis and sparse covariance matrix estimation.

3.5.1. The POET Method

The POET estimator is formed by directly running the singular value decomposition on the sample covariance matrix $\mathbf{S}$. It keeps the covariance matrix formed by the first K principal components and applies the thresholding procedure to the residual covariance matrix. The final covariance matrix estimate is then obtained by combing these two components.

Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_d$ be the ordered eigenvalues of $\mathbf{S}$ and $\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_d$ be their corresponding eigenvectors. Then $\mathbf{S}$ has the following spectral decomposition:

$$\mathbf{S} = \sum_{m=1}^{K}\widehat{\lambda}_m\widehat{\boldsymbol{\xi}}_m\widehat{\boldsymbol{\xi}}_m^T + \widehat{\mathbf{R}}_K \tag{3.8}$$

where the first K principal components $\{\boldsymbol{\xi}_m\}_{m=1}^K$ estimate the latent factors that drive the common dependence. We now apply sparse thresholding on $\widehat{\mathbf{R}}_K$ by solving

$$\widehat{\mathbf{R}}_K^{\lambda} = \operatorname*{argmin}_{\mathbf{\Sigma}}\left\{\frac{1}{2}\|\widehat{\mathbf{R}}_K - \mathbf{\Sigma}\|_{\mathrm{F}}^2 + \sum_{j\neq k} P_{\lambda,\gamma}(\sigma_{jk})\right\}. \tag{3.9}$$

For example, one can apply the soft-thresholding (3.3) to $\widehat{\mathbf{R}}_K$.

The POET estimator of $\mathbf{\Sigma}^*$ is defined as:

$$\widehat{\mathbf{\Sigma}}_K = \sum_{m=1}^{K}\widehat{\lambda}_m\widehat{\boldsymbol{\xi}}_m\widehat{\boldsymbol{\xi}}_m^T + \widehat{\mathbf{R}}_K^{\lambda}. \tag{3.10}$$

It is a nonparametric estimator and can be positive definite when λ is large [15]. When $K > 0$ and we use the thresholding method in §3.2 with $\tau_{ij} = \sqrt{\widehat{r}_{K,ii}\widehat{r}_{K,jj}}$, namely, $\tau = 1$, where $\widehat{r}_{K,ii}$ is a diagonal element of $\widehat{\mathbf{R}}_K$. In this case, $\widehat{\mathbf{R}}_K^{\lambda} = \mathrm{diag}(\widehat{\mathbf{R}}_K)$, and our nonparametric estimator $\widehat{\mathbf{\Sigma}}_K$ is positive definite even when $d \gg n$. It is called the estimator based on the strict factor model [26]. The POET method is available in the R-package named `POET`. When $K = 0$, the POET estimator reduces to the thresholding estimator in §3.2.

3.5.2. Approximate Factor Model

POET exploits an approximate factor structure [14, 15] and works the best under such a model. The approximate factor model assumes

$$\boldsymbol{X} = \mathbf{B}\boldsymbol{U} + \boldsymbol{\epsilon} \tag{3.11}$$

where $\mathbf{B}$ is a $d \times K$ loading matrix, $\boldsymbol{U}$ is a $K \times 1$ latent factor vector, and $\boldsymbol{\epsilon}$ is a d-dimensional random error term that is uncorrelated with $\boldsymbol{U}$. The model implied covariance structure is

$$\mathbf{\Sigma}^* = \mathbf{B}\mathbf{\Sigma}_U\mathbf{B}^T + \mathbf{\Sigma}_{\boldsymbol{\epsilon}}, \tag{3.12}$$

where $\mathbf{\Sigma}_U = \mathrm{Var}(\boldsymbol{U})$ and $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \mathrm{Var}(\boldsymbol{\epsilon})$. We assume $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$ is sparse. This can be interpreted as the conditional sparse covariance model: Given the latent factor $\boldsymbol{U}$, the conditional (after taking the linear projection out) covariance matrix of $\boldsymbol{X}$ is sparse.

Model (3.11) is not identifiable. The linear space spanned by the first K principal components of $\mathbf{B}\mathbf{\Sigma}_U\mathbf{B}^T$ is the same as those spanned by the columns of $\mathbf{B}$. Without loss of generality, we impose the identifiability condition that the columns of $\mathbf{B}$ are orthogonal and $\mathbf{\Sigma}_U = \mathbf{I}_K$.

3.5.3. Modeling Assumption

POET assumes that the factors $\boldsymbol{U}$ are pervasive and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is sparse. An example of pervasiveness is that the factor loadings $\{\boldsymbol{b}_i\}_{i=1}^d$ (the transpose of the i^{th} row of $\mathbf{B}$) for $\{X_i\}_{i=1}^d$ are an independent realization from a certain population quantity $\boldsymbol{b}$. In this case,

$$\frac{1}{d}\mathbf{B}\mathbf{B}^T = \frac{1}{d}\sum_{i=1}^{d} \boldsymbol{b}_i\boldsymbol{b}_i^T$$

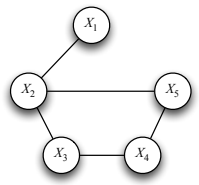
converges to a non-degenerate matrix $\mathbb{E}\boldsymbol{b}\boldsymbol{b}^T$ (as $d \to \infty$). Therefore, the matrix $\mathbf{B}\mathbf{B}^T$ (recalling $\boldsymbol{\Sigma}_U = \mathbf{I}_K$) in (3.12) has spiked eigenvalues. The sparseness is made so that $d^{-1}\|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\|_2 \to 0$. The rationale of this assumption is that, after taking out the common factors, many residual pairs become weakly correlated. Therefore, the first term in (3.12) dominates and the principal component analysis is approximately the same as the factor analysis when d is large.

The spiked eigenvalues assumption still needs to be verified on more applications. Since POET works very well in the presence of spiked principal eigenvalues where most of the aforementioned sparse covariance matrix estimation methods may fail, it has many potential applications in pharmacogenomics.

4. Large Inverse Covariance Matrix Estimation

Estimating a large inverse covariance matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\Sigma}^*)^{-1}$ is another fundamental problem in Big data analysis. Unlike the covariance matrix $\boldsymbol{\Sigma}^*$ which only captures the marginal correlations among $X_1, \ldots, X_d$, the inverse covariance matrix $\boldsymbol{\Theta}^*$ captures the conditional correlations among these variables and is closely related to undirected graphs under a Gaussian model.

More specifically, we define an undirected graph $G = (V, E)$, where V contains nodes corresponding to the d variables in $\boldsymbol{X}$ and the edge $(j,k) \in E$ if and only if $\boldsymbol{\Theta}^*_{jk} \neq 0$. Let $\boldsymbol{X}_{-\{j,k\}} = \{X_\ell : \ell \neq j,k\}$. Under a Gaussian model $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, X_j is independent of X_k given $\boldsymbol{X}_{-\{j,k\}}$ for all $(j,k) \notin E$. Therefore, the graph estimation problem is reduced to estimating the inverse covariance matrix $\boldsymbol{\Theta}^*$ [45].



(a) Undirected Graph

$$\boldsymbol{\Theta}^* = \begin{pmatrix} 1 & 0.2 & 0 & 0 & 0 \\ 0.2 & 1 & 0.2 & 0 & 0.2 \\ 0 & 0.2 & 1 & 0.2 & 0 \\ 0 & 0 & 0.2 & 1 & 0.2 \\ 0 & 0.2 & 0 & 0.2 & 1 \end{pmatrix}$$

(b) Inverse Covariance Matrix

Figure 3: A sparse inverse covariance matrix $\boldsymbol{\Theta}^*$ and the corresponding undirected graph defined by $\boldsymbol{\Theta}^*$

Figure 3 illustrates the difference between the marginal and conditional uncorrelatedness. From Figure 3(b), we see the inverse covariance matrix $\boldsymbol{\Theta}^*$ has many zero entries. Thus the undirected graph defined by $\boldsymbol{\Theta}^*$ is sparse. The covariance matrix $\boldsymbol{\Sigma}^* = (\boldsymbol{\Theta}^*)^{-1}$ is

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 1.05 & -0.23 & 0.05 & -0.02 & 0.05 \\ -0.23 & 1.45 & -0.25 & 0.10 & -0.25 \\ 0.05 & -0.25 & 1.10 & -0.24 & 0.10 \\ -0.02 & 0.10 & -0.24 & 1.10 & -0.24 \\ 0.05 & -0.25 & 0.10 & -0.24 & 1.10 \end{pmatrix}.$$

It is a dense matrix, which implies that every pair of variables are marginally correlated. Thus the covariance matrix and inverse covariance matrix encode different relationships. For example, in Figure 3(a), even though X_1 and X_5 are conditionally uncorrelated given the other variables, they are marginally correlated since both of them are correlated with X_2.

In the following subsections, we first introduce the Gaussian graphical model which provides a general framework for inverse covariance matrix estimation. We then introduce several estimation methods which are computationally simple and suitable for Big data analysis. In particular, we highlight two methods named CLIME [16] and TIGER [17], which have potential to be widely applied for different pharmacogenomics applications. We also briefly explain how to combine the ideas of POET with TIGER.

We first introduce some additional notations. Let $\mathbf{A}$ be a symmetric matrix and I and J be index sets. We denote $\mathbf{A}_{I,J}$ to be the submatrix of $\mathbf{A}$ with rows and columns indexed by I and J. Let $\mathbf{A}_{*j}$ be the j^{th} column of $\mathbf{A}$ and $\mathbf{A}_{*-j}$ be the submatrix of $\mathbf{A}$ with the j^{th} column removed. We define the matrix norm:

$$\|\mathbf{A}\|_q = \max_{\|\boldsymbol{v}\|_q=1} \|\mathbf{A}\boldsymbol{v}\|_q.$$

It is easy to see that when $q = \infty$, $\|\mathbf{A}\|_\infty = \|\mathbf{A}\|_1$.

4.1. Gaussian Graphical Model

Let $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, $\boldsymbol{\alpha}_j = (\boldsymbol{\Sigma}^*_{-j,-j})^{-1}\boldsymbol{\Sigma}^*_{-j,j}$ and $\sigma_j^2 = \boldsymbol{\Sigma}^*_{jj} - \boldsymbol{\Sigma}^*_{-j,j}(\boldsymbol{\Sigma}^*_{-j,-j})^{-1}\boldsymbol{\Sigma}^*_{-j,j}$. Then, we have

$$X_j = \boldsymbol{\alpha}_j^T \boldsymbol{X}_{-j} + \epsilon_j \tag{4.1}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\boldsymbol{X}_{-j} = \{X_\ell : \ell \neq j\}$. Using the block matrix inversion formula, we have

$$\boldsymbol{\Theta}^*_{jj} = \sigma_j^{-2}, \tag{4.2}$$

$$\boldsymbol{\Theta}^*_{-j,j} = -\sigma_j^{-2}\boldsymbol{\alpha}_j. \tag{4.3}$$

Therefore, we can recover $\boldsymbol{\Theta}^*$ by estimating the regression coefficient $\boldsymbol{\alpha}_j$ and the residual variance σ_j^2. Indeed, [46] estimates $\boldsymbol{\alpha}_j$ by solving the Lasso problem

$$\widehat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \frac{1}{2n}\left\|\mathbf{X}_{*j} - \mathbf{X}_{*-j}\boldsymbol{\alpha}_j\right\|_2^2 + \lambda_j\left\|\boldsymbol{\alpha}_j\right\|_1, \tag{4.4}$$

where λ_j is a tuning parameter. Once $\widehat{\boldsymbol{\alpha}}_j$ is given, we get the neighborhood edges by reading out its nonzero coefficients. The final graph estimate $\widehat{G}$ is obtained by combining the neighborhoods for all the d nodes. However, this method only estimates the graph G but not the inverse covariance matrix $\boldsymbol{\Theta}^*$.

4.2. Penalized Likelihood Estimation

When the penalized likelihood approach is applied to estimate $\mathbf{\Theta}^*$, it becomes

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}=(\theta_{ij})}{\operatorname{argmin}}\Big\{ \underbrace{\log|\mathbf{\Theta}| + \operatorname{tr}(\boldsymbol{S}\mathbf{\Theta})}_{\text{negative Gaussian log-likelihood}} + \sum_{i\neq j} P_{\lambda,\gamma}(\theta_{ij})\Big\}, \quad (4.5)$$

where the first part is the negative Gaussian log-likelihood and $P_{\lambda,\gamma}(\cdot)$ is defined the same way as in §3.2. It is a penalty term that encourages sparse solutions. The optimization in (4.5) can be computed by the graphical lasso [47] and the first-order method [48] and have also been studied in [27, 28].

4.3. The CLIME Estimator

To estimate both the inverse covariance matrix $\mathbf{\Theta}^*$ and graph G, [16] proposes the CLIME estimator, which directly estimates the j^{th} column of $\mathbf{\Theta}^*$ by solving

$$\widehat{\mathbf{\Theta}}_{*j} = \underset{\mathbf{\Theta}_{*j}}{\operatorname{argmin}}\big\|\mathbf{\Theta}_{*j}\big\|_1 \text{ subject to } \big\|\mathbf{S}\mathbf{\Theta}_{*j} - \boldsymbol{e}_j\big\|_\infty \le \delta_j, \quad (4.6)$$

where $\boldsymbol{e}_j$ is the j^{th} canonical vector (i.e., the vector with the j^{th} element being 1, while the remaining elements being 0) and δ_j is a tuning parameter. The method borrows heavily the idea from the Dantzig selector [49]. [16] shows that this convex optimization can be formulated into a linear program and has the potential to scale to large problems. Once $\widehat{\mathbf{\Theta}}$ is obtained, we use another tuning parameter τ to threshold it to estimate the graph G.

4.4. The TIGER Estimator

There are several ways to implement model (4.1). Examples include the graphical Dantzig selector [50], the scaled-Lasso estimator [51], and the SICO estimator [52]. Each of these involves d tuning parameters such as $\{\lambda_j\}_{j=1}^d$ in (4.4) and $\{\delta_j\}_{j=1}^d$ in (4.6). Yet, model (4.1) is heteroscedastic, with noise variance depending on unknown σ_j^2. Therefore, these tuning parameters should depend on j, which is difficult to implement in practice.

TIGER method [17] overcomes these scale problems. Let $\widehat{\mathbf{D}} = \operatorname{diag}(\mathbf{S})$ be a d-dimensional diagonal matrix with the diagonal elements being the same as those in $\mathbf{S}$. Let

$$\begin{aligned} \boldsymbol{Z} &= (Z_1,\ldots,Z_d)^T = \boldsymbol{X}\widehat{\mathbf{D}}^{-1/2}, \\ \boldsymbol{\beta}_j &= \widehat{s}_{jj}^{-1/2}\widehat{\mathbf{D}}_{-j,-j}^{1/2}\boldsymbol{\alpha}_j \text{ and } \tau_j^2 = \sigma_j^2\widehat{s}_{jj}^{-1}, \end{aligned}$$

where $\widehat{s}_{jj}$ is the j^{th} diagonal element of $\widehat{\mathbf{S}}$. Therefore, model (4.1) can be standardized as

$$Z_j = \boldsymbol{\beta}_j^T\mathbf{Z}_{-j} + \widehat{s}_{jj}^{-1/2}\epsilon_j. \quad (4.7)$$

We define $\widehat{\mathbf{R}} = (\widehat{\mathbf{D}})^{-1/2}\mathbf{S}(\widehat{\mathbf{D}})^{-1/2}$ to be the sample correlation matrix and let $\mathbf{Z} \in \mathbb{R}^{n\times d}$ be the normalized data matrix, i.e., $\mathbf{Z}_{*j} = s_{jj}^{-1/2}\mathbf{X}_{*j}$ for $j = 1,\ldots,d$. Motivated by the model in (4.7), we propose the following inverse covariance matrix estimator. More specifically,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_j &= \underset{\boldsymbol{\beta}_j\in\mathbb{R}^{d-1}}{\operatorname{argmin}}\Big\{\frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*-j}\boldsymbol{\beta}_j\big\|_2 + \lambda\big\|\boldsymbol{\beta}_j\big\|_1\Big\}, \quad (4.8)\\ \widehat{\tau}_j &= \frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*-j}\widehat{\boldsymbol{\beta}}_j\big\|_2, \\ \widehat{\mathbf{\Theta}}_{jj} &= \widehat{\tau}_j^{-2}\widehat{s}_{jj}^{-1} \text{ and } \widehat{\mathbf{\Theta}}_{-j,j} = -\widehat{\tau}_j^{-2}\widehat{s}_{jj}^{-1/2}\widehat{\mathbf{D}}_{-j,-j}^{-1/2}\widehat{\boldsymbol{\beta}}_j. \end{aligned}$$

Once we have $\widehat{\mathbf{\Theta}}$, the estimated graph $\widehat{G} = (V,\widehat{E})$ is $(j,k) \in \widehat{E}$ if and only $\widehat{\mathbf{\Theta}}_{jk} \neq 0$. The formulation in (4.8) is called a SQRT-Lasso problem. [53] has shown that the SQRT-Lasso is tuning-insensitive. This explains the tuning-insensitive property of the TIGER method. The TIGER method is available in the R package `flare`.

An equivalent form to estimating the j^{th} column of $\mathbf{\Theta}^*$ is to solve

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_j &= \underset{\boldsymbol{\beta}_j\in\mathbb{R}^{d-1}}{\operatorname{argmin}}\Big\{\sqrt{1 - 2\boldsymbol{\beta}_j^T\widehat{\mathbf{R}}_{-j,j} + \boldsymbol{\beta}_j^T\widehat{\mathbf{R}}_{-j,-j}\boldsymbol{\beta}_j} + \lambda\big\|\boldsymbol{\beta}_j\big\|_1\Big\}, \quad (4.9)\\ \widehat{\tau}_j &= \sqrt{1 - 2\widehat{\boldsymbol{\beta}}_j^T\widehat{\mathbf{R}}_{-j,j} + \widehat{\boldsymbol{\beta}}_j^T\widehat{\mathbf{R}}_{-j,-j}\widehat{\boldsymbol{\beta}}_j}, \end{aligned}$$

In (4.9), λ is a tuning parameter. [17] shows that, by choosing $\lambda = \pi\sqrt{\dfrac{\log d}{2n}}$, the obtained estimator achieves the minimax optimal rates of convergence, thus this procedure is asymptotically tuning-free. For finite samples, [17] suggests to set

$$\lambda = \zeta\pi\sqrt{\frac{\log d}{2n}}, \quad (4.10)$$

and ζ can be chosen from a grid in $[\sqrt{2}/\pi, 1]$. Since the choice of ζ does not depend on any unknown quantities, we call the TIGER procedure *tuning-insensitive*. Empirically, we can simply set $\zeta = \sqrt{2}/\pi$ and the resulting estimator works well in many applications.

If a symmetric inverse covariance matrix estimate is preferred, we make the correction: $\widetilde{\mathbf{\Theta}}_{jk} \leftarrow \min\{\widehat{\mathbf{\Theta}}_{jk}, \widehat{\mathbf{\Theta}}_{kj}\}$ for all $k \neq j$. Another correction method is

$$\widetilde{\mathbf{\Theta}} \leftarrow \frac{\widehat{\mathbf{\Theta}} + \widehat{\mathbf{\Theta}}^T}{2}. \quad (4.11)$$

As has been shown by [16], $\widetilde{\mathbf{\Theta}}$ achieves the same rate of convergence as $\widehat{\mathbf{\Theta}}$.

4.5. Combining POET with TIGER: Conditional Sparse Graphical Model

The TIGER estimator can be integrated into the POET framework. Recall that the main idea of the POET method is to exploit conditional sparsity. The common factors $\mathbf{B}\boldsymbol{U}$ are extracted out from (3.11) via a principal component analysis, resulting in the residual matrix $\widehat{\mathbf{R}}_K$ in (3.8). The thresholding rules are applied directly to $\widehat{\mathbf{R}}_K$ due to the assumption of the conditional sparsity of the covariance matrix. If we assume $(\mathbf{\Sigma}_\epsilon)^{-1}$ is sparse, namely, the genomics network is sparse conditioned on unobservable factors, it is inappropriate to apply

POET. In this case, it is more appropriate to apply the TIGER method on $\widehat{\mathbf{R}}_K$ to obtain the final estimate in (3.10). The graph induced by the TIGER method is a conditional graph after taking out the common dependence on latent factors.

5. Large-Scale Simultaneous Tests and False Discovery Control

Selection of differently expressed genes and proteins as well as finding SNPs that are associated with phenotypes or gene expressions give rise to large-scale simultaneous tests in pharmacogenomics [54, 55, 56]. They examine the marginal effects of the treatments and gain substantial popularity in the last decades [57, 58, 59, 60]. These procedures are designed predominately based on the assumption that test statistics are independent or weakly dependent [61, 62, 63]. Yet, biological outcomes such as gene expressions are often correlated, so do the statistical tests in the genomewide association studies (GWAS). It is very important to incorporate the dependence information in controlling the false discovery proportion (FDP) and to improve the power of the tests [64, 65, 66, 67]. In the recent paper [18] with discussions, it is convincingly demonstrated that FDP varies substantially from data to data, but can still be well estimated and controlled for each given data when the dependence information is properly used.

5.1. Rises of Large-Scale Hypothesis Tests

Suppose that gene or protein expression profiles $\{\boldsymbol{X}_i\}_{i=1}^m$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ are collected for m and n individuals respectively from the treatment and control groups. The problem of selecting significantly expressed genes is the following large-scale two-sample testing:

$$H_{0j} : \mu_{X,j} = \mu_{Y,j} \quad \text{vs} \quad H_{1j} : \mu_{X,j} \neq \mu_{Y,j}, \qquad j = 1, \cdots, d$$

based on the assumption that

$$\boldsymbol{X}_i \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}^*) \quad \text{and} \quad \boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}^*), \tag{5.1}$$

where $\mu_{X,j}$ and $\mu_{Y,j}$ are the mean expressions of the j^{th} gene in the treatment and control groups, respectively. Letting $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ be the sample means, the two-sample Z-statistic is

$$\boldsymbol{Z}_1 = \sqrt{mn/(m+n)}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}^*)$$

with $\boldsymbol{\mu}_1 = \sqrt{mn/(m+n)}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)$. Let $\mathbf{D}$ be the diagonal matrix with the pooled estimates of the marginal standard deviations on its diagonal. The vector of the two-sample t-test statistics has the approximate distribution as follows:

$$\boldsymbol{Z} = \mathbf{D}^{-1}\boldsymbol{Z}_1 \overset{a}{\sim} N(\boldsymbol{\mu}, \mathbf{R}) \tag{5.2}$$

where $\boldsymbol{\mu} = \mathbf{D}^{-1}\boldsymbol{\mu}_1$, $\mathbf{R}$ is the correlation matrix of $\boldsymbol{\Sigma}^*$, and $\overset{a}{\sim}$ means "distributed approximately". Our testing problem then reduces to

$$H_{0j} : \mu_j = 0 \quad \text{vs} \quad H_{1j} : \mu_j \neq 0, \qquad j = 1, \cdots, d \tag{5.3}$$

based on the vector of test statistics $\boldsymbol{Z}$. In pharmacogenomics applications, the correlation matrix $\mathbf{R}$ represents the co-expressions of molecules and is in general unknown.

Another concrete example is that in GWAS, we wish to associate the SNPs with phenotypes or gene expressions. For individual i, let X_{ij} be the genotype of the j^{th} SNP and Y_i be its associated outcome. The association between the j^{th} SNP and the response is measured through the marginal regression [18]:

$$(\alpha_j, \beta_j) = \underset{a_j, b_j}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^n \mathbb{E}(Y_i - a_j - b_j X_{ij})^2.$$

Our simultaneous test in GWAS becomes testing

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad H_{1j} : \beta_j \neq 0, \qquad j = 1, \cdots, d. \tag{5.4}$$

Let $\widehat{\boldsymbol{\beta}}$ be the vector of the least-squares estimates of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$, which are obtained by the marginal regressions based on the data $\{(x_{ij}, y_i)\}_{i=1}^n$. Denoting by $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T$ the t-test statistics for the problem (5.4), it is easy to show that $\boldsymbol{Z} \overset{a}{\sim} N(\boldsymbol{\beta}, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix of $\{\boldsymbol{X}_i\}_{i=1}^n$. In this case, the correlation matrix is known.

The above problem can be summarized as follows. Let $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unit standard deviations ($\sigma_{11}^2 = \ldots \sigma_{dd}^2 = 1$) be a vector of dependent test statistics. We wish to simultaneously test

$$H_{0j} : \mu_j = 0 \quad \text{vs} \quad H_{1j} : \mu_j \neq 0, \qquad j = 1, \cdots, d. \tag{5.5}$$

In pharmacogenomics applications, the number of hypotheses d is large and the number of interested genes or proteins or SNPs are small so that most of $\{\mu_j\}_{j=1}^d$ are zero.

5.2. False Discovery Rate and Proportion

Let $\mathcal{S}_0 = \{j : \mu_j = 0\}$ be the set of true nulls and $\Phi(\cdot)$ be the cumulative distribution function of the standard Gaussian random variable. We denote $P_j = 1 - 2\Phi(|Z_j|)$ to be the P-value for the j^{th} test of problem (5.5) based on the dependent test statistics (5.2). Using a threshold t, we define

$$\text{FDP}(t) = V(t)/R(t) \qquad \text{and} \quad \text{FDR}(t) = \mathbb{E}\{\text{FDP}(t)\}$$

to be the false discovery proportion and false discovery rate, where $V(t) = \#\{j \in \mathcal{S}_0 : P_j \leq t\}$ and $R(t) = \#\{j : P_j \leq t\}$ are the number of false rejections and total rejections, respectively. Our aim is to accurately estimate FDP(t) for a large class of $\boldsymbol{\Sigma}^*$. Clearly, FDP is more relevant, since it is related to the data at hand whereas FDR only controls the average FDP. To control FDP at a prescribed level α, we choose the smallest t_0 such that estimated FDP(t_0) $\leq \alpha$ [56, 18].

Note that $R(t)$ is the number of rejected hypotheses or the number of discoveries. It is an observable random variable. Yet, $V(t)$ is unknown to us and needs to be estimated. When test statistics are independent, $V(t)$ is the number of success in a sequence of independent Bernoulli trials, with the number of trials $d_0 = |\mathcal{S}_0|$ (the number of true nulls) and probability of success t. Therefore, by the law of large numbers, FDP and FDR are close, both approximately $d_0 t/R(t) \approx dt/R(t)$ due to sparsity.

For dependent test statistics, however, FDP varies substantially from one data to another, even when FDR is the same. We use a simple example in [18] to illustrate this point.

Let $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^d$. For equally correlated matrix $\mathbf{\Sigma}^* = (1-\rho)\mathbf{I}_d + \rho \mathbf{1}_d \mathbf{1}_d^T$, we consider the test statistic

$$Y_j = \mu_j + \rho U + \sqrt{1-\rho^2}\epsilon_j,$$

where $U \sim N(0,1)$ is independent of $\epsilon_j \sim N(0,1)$. Without loss of generality, we assume that $\mathcal{S}_0$ is the first d_0 genes. Let $z_{t/2}$ be the upper $t/2$-quantile of the standard normal distribution and $|Y_j| \geq z_{t/2}$ be a critical region. Then, by the law of large numbers,

$$V(t) = \sum_{j=1}^{d_0} I(|Y_j| \geq z_{t/2}) \approx d_0 P(t, \eta, \rho) \tag{5.6}$$

where $P(t, \eta, \rho) = \mathbb{P}(|Y_j| \geq z_{t/2} \,|\, U, \mu_j = 0)$ is given by

$$P(t, \eta, \rho) = \Phi\left(\frac{-z_{t/2} - \eta}{\sqrt{1-\rho^2}}\right) + 1 - \Phi\left(\frac{z_{t/2} - \eta}{\sqrt{1-\rho^2}}\right), \tag{5.7}$$

with $\eta = \rho U$. It is clear that the result depends critically on the realization of U. For example, when $\rho = 0.8$, $d_0 = 1,000$ and $z_{t/2} = 2.5$,

$$V(t) \approx 1000 \times \left\{\Phi\left(\frac{-2.5 - 0.8U}{0.6}\right) + 1 - \Phi\left(\frac{2.5 - 0.8U}{0.6}\right)\right\}$$

which are 0, 2.3, 66.8, 433.6 respectively for $U = 0, 1, 2$, and 3. Clearly, it depends on the realization of U. Yet, the realized factor U is estimable. For example, using sparsity, $\bar{Y} \approx \rho U$ and thus U can be estimated by $\bar{Y}/\rho$. On the other hand, $\mathbb{E}V(t) = d_0 t = 12.4$. Clearly, $V(t)$ or FDP(t) is far more relevant than $\mathbb{E}V(t)$ or FDR(t).

In summary, FDP is a quantity of the primary interest for tests that are dependent.

5.3. Principal Factor Approximation

When the correlation matrix $\mathbf{\Sigma}^*$ is known as in GWAS, [18] gives a principal factor approach to approximate FDP(t) for arbitrary $\mathbf{\Sigma}^*$. It generalizes formula (5.6). Let $\{\lambda_j\}_{j=1}^d$ be ordered eigenvalues of $\mathbf{\Sigma}^*$ and $\{\boldsymbol{\xi}_j\}_{j=1}^d$ be their corresponding eigenvectors. We can decompose $\mathbf{\Sigma}^*$ as

$$\mathbf{\Sigma}^* = \sum_{j=1}^d \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T = \mathbf{B}\mathbf{B}^T + \mathbf{A}$$

where $\mathbf{B} = (\sqrt{\lambda_1}\boldsymbol{\xi}_1, \cdots, \sqrt{\lambda_K}\boldsymbol{\xi}_K)$ consists of the first K unnormalized principal components and $\mathbf{A} = \Sigma_{j=K+1}^d \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T$. Then, as in (3.11), we can write

$$\boldsymbol{Y} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{U} + \boldsymbol{\epsilon}, \qquad \boldsymbol{U} \sim N(0, \mathbf{I}_K), \qquad \boldsymbol{\epsilon} \sim N(0, \mathbf{A}). \tag{5.8}$$

Since principal factors $\boldsymbol{U}$ that drive the dependence are taken out, $\boldsymbol{\epsilon}$ can be assumed to be weakly correlated and FDP(t) can be approximated by

$$\text{FDP}_A(t) = \textstyle\sum_{j=1}^d P(t, \eta_j, \|\boldsymbol{b}_j\|_2)/R(t), \qquad \eta_j = \boldsymbol{b}_j^T \boldsymbol{U} \tag{5.9}$$

under some mild conditions [18], where $\boldsymbol{b}_j^T$ is the j^{th} row of $\mathbf{B}$ and the function $P(\cdot)$ is given in (5.7). Note that the covariance matrix is used to calculate η_j and $\|\boldsymbol{b}_j\|_2$ and that formula (5.9) is a generalization of (5.6). Thus, we need only to estimate the realized but unobserved factors $\boldsymbol{U}$ in order to use FDP$_A(t)$. [18] provides a penalized estimator as described below.

Since $\boldsymbol{\mu}$ is sparse, by (5.8), a natural estimator is the minimizer of

$$\min_{\boldsymbol{\mu}, \boldsymbol{U}} \|\boldsymbol{Y} - \boldsymbol{\mu} - \mathbf{B}\boldsymbol{U}\|_2^2 + \lambda \|\boldsymbol{\mu}\|_1 \tag{5.10}$$

which is equivalent [68] to minimizing the Huber's ψ-loss with respect to $\boldsymbol{U}$: $\min_{\boldsymbol{U}} \psi(\boldsymbol{Y} - \mathbf{B}\boldsymbol{U})$. An alternative is to find the minimizer of the L_1-loss:

$$\min_{\boldsymbol{U}} \|\boldsymbol{Y} - \mathbf{B}\boldsymbol{U}\|_1. \tag{5.11}$$

The sampling property of the estimator based on (5.10) has been established in [68].

5.4. Factor Adjustment: an Alternative Ranking of Molecules

With $\widehat{\boldsymbol{U}}$ estimated from (5.10), the common factors can be subtracted out from model (5.8), resulting in

$$\boldsymbol{Z} = \boldsymbol{Y} - \mathbf{B}\widehat{\boldsymbol{U}} \approx \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{5.12}$$

The test statistics $\boldsymbol{Z}$ are more powerful than those based on $\boldsymbol{Y}$ since $\boldsymbol{\epsilon}$ has marginal variances no larger than 1. In fact, $Z_i \sim N(\mu_i, 1 - \|\boldsymbol{b}_i\|_2^2)$, after ignoring the approximation error. Hence, Z_i has a larger signal-to-noise ratio than $Y_i \sim N(\mu_i, 1)$. In addition, ranking of important genes/proteins based on the standardized test statistics $\{|Z_i|/(1 - \|\boldsymbol{b}_i\|_2^2)^{1/2}\}_{i=1}^d$ can be very different from that based on $\{|Y_i|\}_{i=1}^d$. This method is called dependent-adjusted procedure in [18]. It provides pharmacoscientists a chance to discover important genes, proteins, and SNPs that are not highly ranked by the conventional methods, based on ranking of $\{|Y_i|\}_{i=1}^d$.

After common factors being taken out, the elements of $\boldsymbol{\epsilon}$ are weakly correlated. There is no need to apply PFA to the test statistics $\boldsymbol{Z}$ in (5.12). In other words, FDP should be approximately the same as FDR. But application of PFA to $\{|Z_i|/(1 - \|\boldsymbol{b}_i\|_2^2)^{1/2}\}_{i=1}^d$ will not result in very different results from the case with $K = 0$.

5.5. Estimating FDP with Unknown Dependence

In many genomic applications such as selecting significant genes from microarray data, the covariance matrix $\mathbf{\Sigma}^*$ is typically unknown. A natural approach is to estimate it from the data by POET as described in §3.5.1. Then proceed as if the correlation matrix is given. The resulting method is named POET-PFA. More details on the development of this method can be found in [69]. An R package, called `PFA`, implements this procedure.

6. High Dimensional Variable Selection

With proper coding of the treatment and control groups, it is easy to see that the two-sample test statistics in (5.2) are equivalent to the marginal correlations between the expressions of molecules and response [20]. To examine the conditional contribution of a molecule to the response Y given the rest of molecules, one often employs a sparse generalized linear model under canonical link: Up to a normalization constant, the conditional density of Y given $\boldsymbol{X}$ is

$$\log p(y|\boldsymbol{x}, \boldsymbol{\beta}_0) = (\boldsymbol{x}^T\boldsymbol{\beta}_0)y - b(\boldsymbol{x}^T\boldsymbol{\beta}_0). \tag{6.1}$$

Here, the sparsity means that a majority of regression coefficients in $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,d})^T$ are zero. The regression coefficient $\beta_{0,j}$ is frequently regarded as a measure of conditional contribution of the molecule X_j to the response Y given the rest of molecules, similar to the interpretation of the sparse inverse covariance matrix in §4.

When the response Y is a continuous measurement such as in eQTL studies, one takes $b(\theta) = \theta^2/2$. For binary response, one often employs a sparse logistic regression in which $b(\theta) = -\log(1 + \exp(\theta))$. The latter is closely related to classification problems as will be described in §6.4.

Over the last decade, there is a surged interest in sparse regression. For an overview, see [40, 70, 71]. The basic principles are screening and penalization. We give only a brief summary due to the space limitation.

6.1. Penalized Likelihood Method

As has been shown in [19], sparse vector $\boldsymbol{\beta}_0$ in (6.1) can be estimated by penalized likelihood which minimizes

$$n^{-1}\sum_{i=1}^{n} b(\boldsymbol{x}_i^T\boldsymbol{\beta}) - (\boldsymbol{x}_i^T\boldsymbol{\beta})y_i + \sum_{j=1}^{d} P_{\lambda,\gamma}(|\beta_j|) \tag{6.2}$$

for a folded concave penalty function $P_{\lambda,\gamma}(\cdot)$. It has been systematically studied by [72, 73].

There are many algorithms to compute the minimizer of (6.2). Examples include local quadratic approximation [19], local linear approximation [74], coordinate decent algorithm [75], iterative shrinkage-thresholding algorithm [76, 77]. For example, given estimate $\widehat{\boldsymbol{\beta}}^{(k)} = (\beta_1^{(k)}, \cdots, \beta_d^{(k)})^T$ at the k^{th} iteration, by Taylor's expansion,

$$P_{\lambda,\gamma}(|\beta_j|) \approx P_{\lambda,\gamma}(|\beta_j^{(k)}|) + P'_{\lambda,\gamma}(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|). \tag{6.3}$$

Thus, at the $(k+1)^{th}$ iteration, we minimize

$$n^{-1}\sum_{i=1}^{n} b(\boldsymbol{x}_i^T\boldsymbol{\beta}) - (\boldsymbol{x}_i^T\boldsymbol{\beta})y_i + \sum_{j=1}^{d} w_{k,j}|\beta_j|, \tag{6.4}$$

where $w_{k,j} = P'_{\lambda,\gamma}(|\beta_j^{(k)}|)$. Note that problem (6.4) is convex so that a convex solver can be used. If one further approximates the likelihood part in (6.4) by a quadratic function via Taylor expansion, then the LARS algorithm [78] can be used.

6.2. Screening Method

The sure independent screening [20, 79] is another effective method to reduce the dimensionality in sparse regression problems. It utilizes the marginal contribution of a covariate to probe its importance in the joint regression model. For example, assuming each covariate has been standardized, the marginal contribution can be measured by the magnitude of $\widehat{\beta}_j^M$, which, along with $\widehat{\alpha}_j^M$, minimizes the negative marginal likelihood:

$$n^{-1}\sum_{i=1}^{n} b(\alpha_j + \beta_j x_{ij}) - (\alpha_j + \beta_j x_{ij})y_i. \tag{6.5}$$

The set of covariates that survive the marginal screening is

$$\widehat{\mathcal{S}} = \{j : |\widehat{\beta}_j^M| \geq \delta\}, \tag{6.6}$$

for a given threshold δ. One can also measure the importance of a covariate X_j by using its deviance reduction. For the least-squares problem, both methods reduce to ranking importance of predictors by using the magnitudes of their marginal correlations with the response Y. [20, 79] give conditions under which sure screening property can be established and false selection rate are controlled.

The idea of sure screening is very effective in dramatically reducing the computational burden of Big data analysis. It has been extended in various directions. For example, generalized correlation screening was used in [80] and nonparametric screening was proposed by [81]. In addition, [82] utilizes the distance correlation to conduct screening and [83] employs rank correlation.

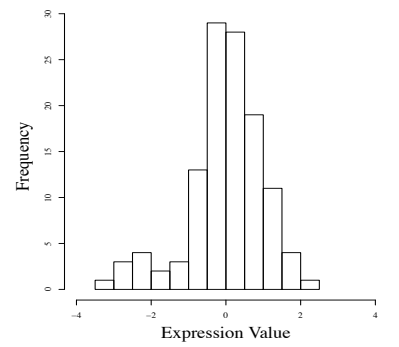
6.3. Iterative Sure Independence Screening

An iterative sure independence screening method was developed in [84]. The basic idea is to iteratively use the large-scale screening (6.6) and the moderate-scale selection, which applies the penalized likelihood method (6.2) to the survived variables (6.6). The software, `ISIS`, is available in R-package, which implements this idea. It also allows us to compute penalized likelihood (6.2) or (6.6) alone.

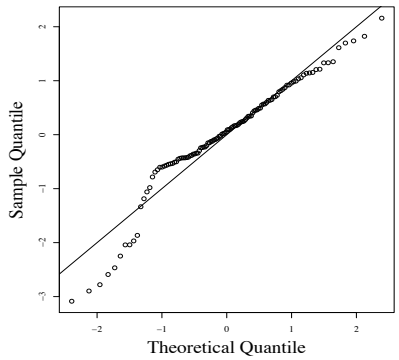
6.4. Sparse Classification

Classification, also known as supervised learning, is to predict the class label for a given covariate. In pharmacogenetic studies, one often wishes to not only predict well the class labels, but also understand the molecule mechanisms for the effectiveness of drugs. Therefore, it is desirable to select a small group of molecules that have a high classification power.

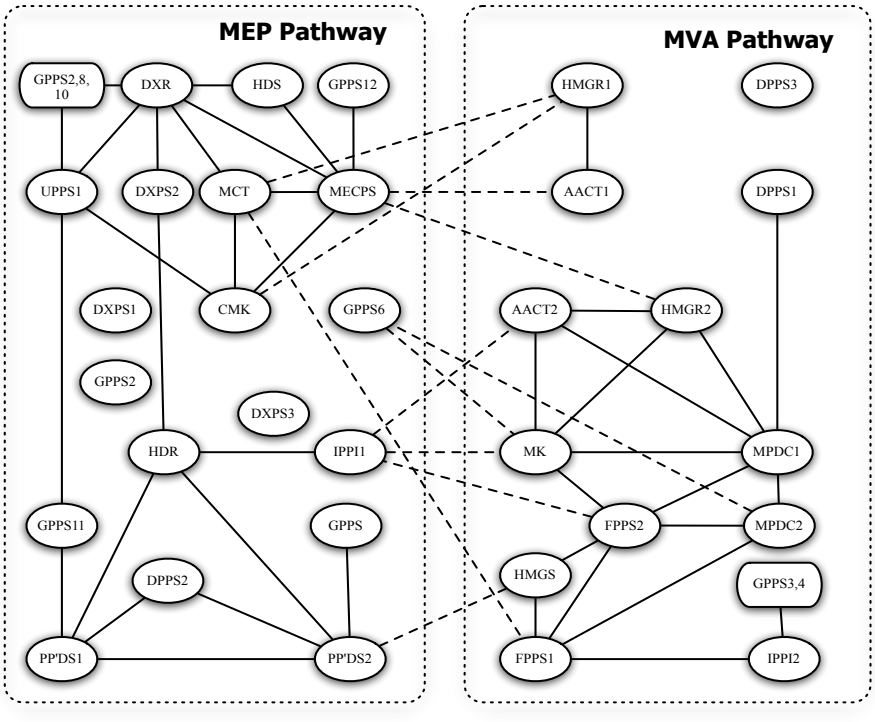
Sparse logistic regression (6.2) provides an effective approach to high dimensional classification. Other approaches include the support vector machine [85] and the AdaBoost algorithm [86, 87]. The former replaces the log-likelihood by the hinge loss $L(\theta, y) = (1 - \theta y)_+$ and the latter replaces the log-likelihood by the exponential loss $L(\theta, y) = \exp(-\theta y)$. In both cases, the response Y is recoded as ± 1.



(a) Histogram of expression of MECPS



(b) Normal QQ-Plot of MECPS



(c) Estimated Gene Network

Figure 4: (a) and (b): The histogram and normal QQ plots of the marginal expression levels of the gene MECPS. We see that the data are not exactly Gaussian distributed. (c) The estimated gene networks of the Arabadopsis data. The within-pathway links are denoted by solid lines and between-pathway links are denoted by dashed lines. Adapted from [17].

When data from both classes follow normal distributions (5.1), the optimal classifier is the Fisher's discriminant rule, which classifies a new data point $\boldsymbol{x}$ to class "X" if

$$(\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)^T (\boldsymbol{\Sigma}^*)^{-1} (\boldsymbol{x} - (\boldsymbol{\mu}_X + \boldsymbol{\mu}_Y)/2) > 0, \tag{6.7}$$

where $\boldsymbol{\Sigma}^*$ is the common covariance matrix in (5.1). When the dimensionality d is large, we could estimate $\boldsymbol{\Sigma}^*$ or its inverse $(\boldsymbol{\Sigma}^*)^{-1}$ by the regularized estimators introduced in §3 or §4, depending on the sparsity assumption. See, for example, [88, 89, 90, 91].

7. Applications

In this section, we apply the aforementioned methods to estimate large gene networks and select significantly differently expressed genes when test statistics are dependent.

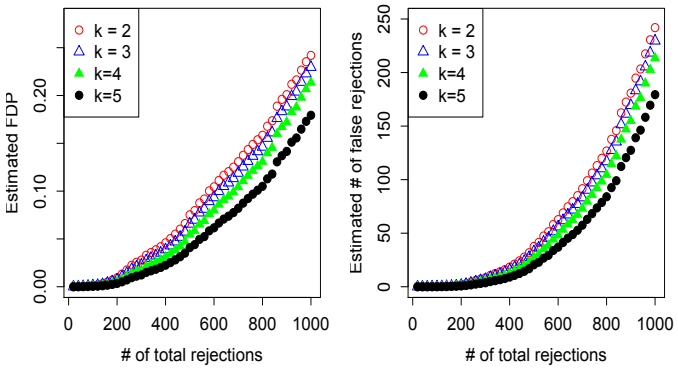
7.1. Network Estimation

As has been explained in §4, an important application of large inverse covariance matrix estimation is to reconstruct the undirected graph of a high dimensional distribution based on observational data. In this section, we apply the TIGER method on a microarray data to illustrate the main idea.

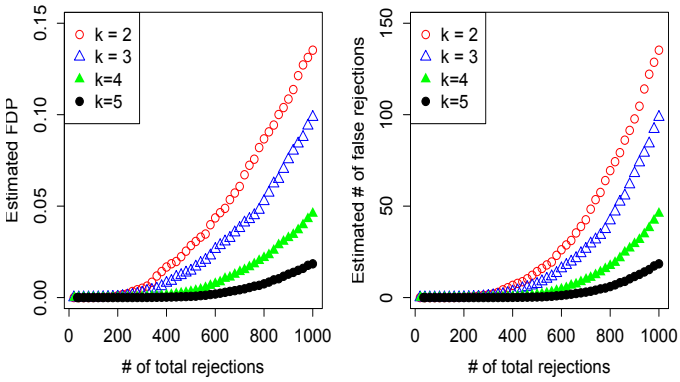
This dataset includes 118 gene expression arrays from Arabidopsis thaliana originally appeared in [92]. Our analysis focuses on expression profiles from 39 genes involved in two isoprenoid metabolic pathways: 16 from the mevalonate (MVA) pathway are located in the cytoplasm, 18 from the plastidial (MEP) pathway are located in the chloroplast, and 5 are located in the mitochondria. While the MVA and MEP pathways generally operate independently, crosstalk is known to happen [92]. Our goal is to reconstruct the gene network by estimating the undirected graph, with special interest in crosstalk.

We first examine whether the data actually satisfies the Gaussian distribution assumption. In Figure 4(a) and Figure 4(b), we plot the histogram and normal QQ plot of the expression levels of a gene named MECPS. From the histogram, we see that the distribution is left-skewed compared to the Gaussian distribution. From the normal QQ plot, we see that the empirical distribution has a heavier tail compared to Gaussian. To apply the TIGER method to analyze this data, we need to first transform the data so that its distribution is closer to Gaussian. For this, we Gaussianize the expression values of each gene by converting them to the corresponding normal-scores. This is automatically done by the `huge.npn` function in the `R` package `huge` [93].

We apply the TIGER method on the transformed data using the default tuning parameter $\zeta = \sqrt{2}/\pi$. The estimated network is shown in Figure 4(c). We see the estimated network is very sparse with only 44 edges. We draw the within-pathway



(a) Results from the POET-PFA method



(b) Results from the dependence-adjusted method

Figure 5: The estimated false discovery proportion and the estimated number of false discoveries as functions of the number of total discoveries for $d = 3,226$ genes, where the number of factors are chosen to be $k = 2, 3, 4, 5$. (a) Results from the POET-PFA method; (b) Results from the dependence-adjusted method. Adapted from [69].

connections using solid lines and the between-pathway connections using dashed lines. Our result is consistent with previous investigations, which suggest that the connections from genes AACT1 and HMGR2 to gene MECPS indicate a primary sources of the crosstalk between the MEP and MVA pathways and these edges are presented in the estimated network. MECPS is clearly a hub gene for this pathway.

For the MEP pathway, the genes DXPS2, DXR, MCT, CMK, HDR, and MECPS are connected as in the true metabolic pathway. Similarly, for the MVA pathway, the genes AACT2, HMGR2, MK, MPDC1, MPDC2, FPPS1 and FPP2 are closely connected. Our analysis suggests 11 cross-pathway links, which is consistent to previous investigation in [92]. This result suggests that there might exist rich inter-pathway crosstalks.

7.2. Select Significantly Expressed Genes under Dependency

We now apply the POET-PFA method in §5.5 and the dependence-adjusted method in §5.4 to analyze a microarray data. The data contains expression profiles from a well-known breast cancer study [94, 95]. This study involves 15 woman subjects with two genetic mutations: BRCA1 (7 subjects) and BRCA2 (8 subjects). These two genetic mutations are known to increase the lifetime risk of hereditary breast cancer. We want to find a set of genes that are associated with these genetic mutations. This allows us to identify cases of hereditary breast cancer on the basis of gene-expression profiles.

We make two assumptions: (i) A large proportion of the genes are not differently expressed; (ii) The gene expression follows an approximate k-factor model. Both assumptions have gained increasing popularity among biologists in the past decade, since it has been widely acknowledged that gene activities are usually driven by a small number of latent variables and the genetic mutations are only caused by a small amount of genes. See [65, 96] for more details.

We first apply the POET-PFA method to obtain a consistent FDP estimator $\widehat{\text{FDP}}(t)$ for a given threshold value t and a fixed number of factors k.

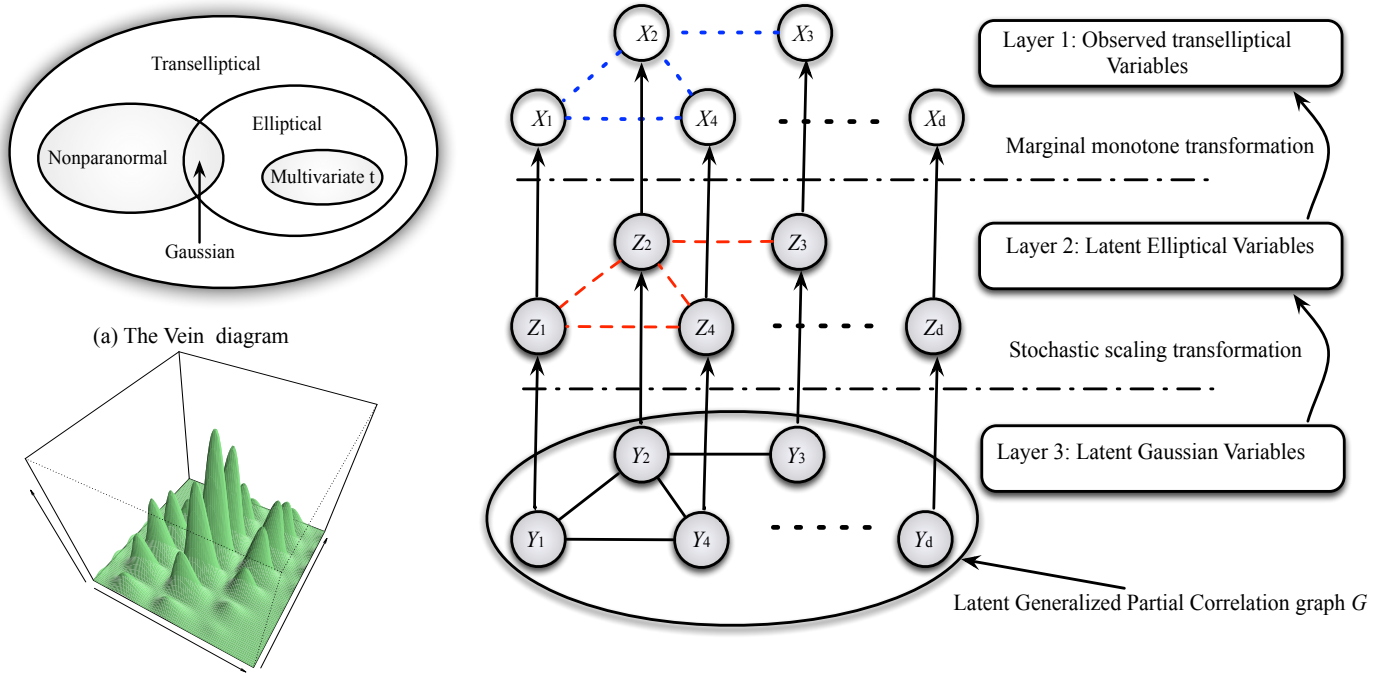
Let $\widehat{V}(t)$ be the estimated number of false rejections, which is the numerator of (5.9). The results of our analysis are summarized in Figure 5(a), which has two subfigures that plot $(R(t), \widehat{\text{FDP}}(t))$ and $(R(t), \widehat{V}(t))$ for $k = 2, 3, 4, 5$. We see that $\widehat{\text{FDP}}(t)$ is close to zero when $R(t)$ is below 200, suggesting that the rejected hypotheses in this range have high accuracy to be the true discoveries. In addition, when 1,000 hypotheses, almost 1/3 of the total number, have been rejected, the estimated FDPs are still as low as 25%. Finally, it is worth noting that although our procedure seems robust under different choices of number of factors, the estimated FDP tends to be relatively small with larger number of factors.

We also apply the dependence-adjusted procedure to the data. The results are shown in Figure 5(b). Comparing with Figure 5(a), both the estimated FDP and the estimated number of false rejections become smaller.

The selected gene sets based on the conventional two-sample t tests and the factor-adjusted tests are different (not presented here). This provides pharmaco-scientists a chance to discover important genes, proteins, and SNPs that are not highly ranked by the conventional methods. In fact, the factor-adjusted method is more powerful than the conventional two-sample t-tests when test statistics are dependent.

8. Discussion and Future Directions

The paper is written in the context that massive amounts of pharmacogenomic data combined with quantitative statistical approaches are beginning to shed lights towards personalized medicine. While the so-called Big data movement has received a lot of attention, and the pharmacogenomics applications are prime examples, we highlight three characteristics of modern pharmacogenomic data that worth equal attention. These include: (1) Complex Data - the data distribution can not be characterized by simple parametric models; (2) Noisy Data - the data are usually aggregated from numerous sources and we must deal with large biological and measurement heterogeneity with possible outliers, measurement errors, and missing data; (3) Dependent Data - the data exhibit complicated correlations with relatively weak signals. These problems are now so ubiquitous, with new variations accruing at such an alarming rate, one might refer to them as simply the **Modern data** problem. The large amount of data are usually obtained by using high-throughput technologies and they contain inevitably



(a) The Vein diagram

(b) The perspective plot of a transelliptical density

(c) The hierarchical diagram illustrating the transelliptical graphical model

Figure 6: Transelliptical family. (a) The Vein diagram illustrating the relationships of the distribution families. (b) The perspective plot of a transelliptical density. (c) The hierarchical latent variable representation of the transelliptical graphical model with the latent variables grey-colored. Here the first layer is composed of observed X_j, and the second and third layers are composed of latent variables Z_j and Y_j. The solid undirected lines in the third layer encode the conditional independence graph of $Y_1, \ldots, Y_d$ (Adapted from a manuscript that is under review).

measurement errors. Measurement errors distort statistical conclusion, reducing correlation with clinical outcomes. Spurious correlations arise inevitably in high dimensional data. They induce the so-called endogenous covariates. When some collected variables are correlated with the residual noise of a response, namely, when the part of the response that can not be explained by relevant molecules are correlated with irrelevant molecules, endogeneity arises. Like measurement errors, endogeneity causes model selection inconsistency, leading to erroneous scientific conclusions.

To handle the challenges of modern pharmacogenomic data, we need statistical methods that are simultaneously robust to the main issues of scalability, complexity, noise, and dependence. However, general methodological development for pharmacogenomic data analysis, mirroring numerous other scientific settings, has lagged behind the rapid development of new technologies and new datasets. For example, most existing methods assume the data are independently sampled from a parametric distribution (e.g., Gaussian), and most of them use Pearson's sample covariance matrix as the sufficient statistic and thus are not robust to outliers and possible data contamination. To bridge these gaps, new statistical methodology is urgently needed. In the following sections, we discuss several recent efforts that aim at addressing these problems.

8.1. Handling Complex Data with Semiparametric Modeling

To deal with complex nonGaussian data, [97] has proposed a semiparametric modeling framework based on the *transelliptical distribution family*.

The transelliptical family is a semiparametric extension of the elliptical family. Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with mean $\boldsymbol{\mu}$ and correlation matrix $\boldsymbol{\Sigma}$. We say that $\boldsymbol{X}$ follows an elliptical distribution if its density can be written in the form of $p(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$. Elliptical family contains many multivariate distributions, including multivariate Gaussian, multivariate t-distribution, Cauchy, logistic and Kotz distributions. We define $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ follows a transelliptical distribution, denoted by $\boldsymbol{X} \sim TE(\boldsymbol{\Sigma}, g; f)$, if there exists a set of increasing functions $\{f_j\}_{j=1}^d$ such that $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))^T$ follows an elliptical distribution.

Figure 6(a) illustrates the relationships of the transelliptical, elliptical [98], and nonparanormal families [99, 100, 101]. Both the nonparanormal and elliptical distributions are proper subsets of the transelliptical family. They share the Gaussian family as a common subset. Figure 6(b) visualizes a 2-dimensional transelliptical density. Clearly the transelliptical family is much richer than the Gaussian family.

Similar to the Gaussian graphical model, we could also construct the transelliptical graphical model based on the transelliptical family (More details can be found in [97]). To understand the semantics of a transelliptical graph, we have proved that a transelliptical distribution must admit a three-layer hierarchical latent variable representation as illustrated in Figure 6(c): The observed vector, denoted by $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ and presented in the first layer, has a transelliptical distribution, and a

latent random vector, $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T$ in the second-layer, is elliptically distributed. Variables in the first and second layers are related through the transformation $Z_j = f_j(X_j)$ with f_j being an unknown monotone function. The latent vector $\boldsymbol{Z}$ can be further represented by a third-layer latent random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^T$, which is multivariate Gaussian with a correlation matrix $\boldsymbol{\Sigma}$ (called latent correlation matrix) and an inverse correlation matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ (called latent inverse correlation matrix). We define the transelliptical graph $G = (V, E)$ with the node set $V = \{1, \ldots, d\}$ and the edge set E encoding the nonzero entries of $\boldsymbol{\Theta}$. We provide interpretations of the graph G for the variables in different layers: (i) For the observed variables in the first layer, the absence of an edge between two variables means the absence of a certain rank-based association of the pair given other variables; (ii) For the latent variables in the second layer, the absence of an edge means the absence of the conditional Pearson's correlation of the pair; (iii) For the third layer variables, the absence of an edge means the conditional independence of the pair. Compared with the Gaussian graphical model, the transelliptical graphical model has richer structure with more relaxed modeling assumptions.

8.2. Handling Noisy and Dependent Data with Robust Methods

To handle noisy data, we need to develop robust statistical methods. We introduce a rank-based method using the aforementioned transelliptical graphical model as an illustrative example. Recall that estimating the transelliptical graph is equivalent to estimating the latent inverse correlation matrix, [97] develops a rank-based estimator using the Kendall's tau statistics. This estimator has been proved to have good theoretical property and is simultaneously robust to outliers, missing values, and data dependence.

More specifically, let $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of a random vector $\boldsymbol{X} \in \mathbb{R}^d$. The population Kendall's tau statistic is

$$\tau_{jk} = \mathrm{Corr}(\mathrm{sgn}(X_j - \widetilde{X}_j), \mathrm{sgn}(X_k - \widetilde{X}_k)).$$

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be n observed data points with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$. The sample version Kendall's tau statistic is

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le s < t \le n} \mathrm{sgn}(x_{sj} - x_{tj})(x_{sk} - x_{tk}).$$

This is a monotone-transformation-invariant measure of association between the empirical realizations of two random variables X_j and X_k. Let $\widehat{\mathbf{S}} = (\widehat{s}_{jk}) \in \mathbb{R}^{d \times d}$ with $\widehat{s}_{jk} = \sin(\frac{\pi}{2}\widehat{\tau}_{jk}) \cdot I(j \neq k) + I(j = k)$, where $I(\cdot)$ is the indicator function. A rank-based estimator $\widehat{\boldsymbol{\Theta}}$ can be obtained by plugging $\widehat{\mathbf{S}}$ into a sparse inverse covariance matrix estimation algorithm like CLIME [16] or TIGER [17]. Such a rank-based estimator is robust and easy to compute.

Besides the above rank-based method, there are many other ways we can develop robust estimators. In summary, statistical analysis of Big pharmacogenomics data is both promising and challenging. On one hand, significant progress has been made towards developing new statistical method and theory to explore and predict massive amounts of high dimensional data. On the other hand, there are many new challenges and open problems remain to be solved.

Acknowledgements

We thank Ronglin Wu for his helpful comments and discussions. Jianqing Fan is supported by NSF Grant DMS-1206464 and NIH Grants R01GM100474 and R01-GM072611. Han Liu is supported by NSF Grant III-1116730 and a NIH sub-award from Johns Hopkins University.

References

[1] W. E. Evans, M. V. Relling, Pharmacogenomics: translating functional genomics into rational therapeutics, Science 286 (5439) (1999) 487–491.

[2] A. J. Wood, W. E. Evans, H. L. McLeod, Pharmacogenomics-drug disposition, drug targets, and side effects, New England Journal of Medicine 348 (6) (2003) 538–549.

[3] K. Jain, Applications of biochip and microarray systems in pharmacogenomics, Pharmacogenomics 1 (3) (2000) 289–307.

[4] P. J. Mishra, J. R. Bertino, Microrna polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine, Pharmacogenomics 10 (3) (2009) 399–416.

[5] B. R. Winkelmann, W. März, B. O. Boehm, R. Zotz, J. Hager, P. Hellstern, J. Senges, Rationale and design of the luric study-a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease, Pharmacogenomics 2 (1) (2001) 1–73.

[6] H. E. Wheeler, M. L. Maitland, M. E. Dolan, N. J. Cox, M. J. Ratain, Cancer pharmacogenomics: strategies and challenges, Nature Reviews Genetics (2013) 23–34.

[7] S. Ross, S. S. Anand, P. Joseph, G. Paré, Promises and challenges of pharmacogenetics: an overview of study design, methodological and statistical issues, JRSM Cardiovascular Disease 1 (1) (2012) 1–7.

[8] R. Wu, M. Lin, Statistical and computational pharmacogenomics, Chapman & Hall/CRC, 2008.

[9] B. J. Grady, M. D. Ritchie, Statistical optimization of pharmacogenomics association studies: key considerations from study design to analysis, Current pharmacogenomics and personalized medicine 9 (1) (2011) 41–66.

[10] S.-J. Wang, R. T O'Neill, H. J. Hung, Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials, Clinical Trials 7 (5) (2010) 525–536.

[11] S. D. Turner, D. C. Crawford, M. D. Ritchie, Methods for optimizing statistical analyses in pharmacogenomics research, Expert review of clinical pharmacology 2 (5) (2009) 559–570.

[12] E. Topić, Pharmacogenomics and personalized medicine, The 7th EFCC Continuous Postgraduate Course in Clinical Chemistry: New trends in diagnosis, monitoring and management using molecular diagnosis methods.

[13] P. Bickel, E. Levina, Covariance regularization by thresholding, Annals of Statistics 36 (6) (2008) 2577–2604.

[14] J. Fan, Y. Liao, M. Mincheva, High dimensional covariance matrix estimation in approximate factor models, Annals of statistics 39 (6) (2011) 3320–3356.

[15] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements (with discussion), Journal of the Royal Statistical Society Series B, to appear.

[16] T. Cai, W. Liu, X. Luo, A constrained ℓ_1 minimization approach to sparse precision matrix estimation, Journal of the American Statistical Association 106 (494) (2011) 594–607.

[17] H. Liu, L. Wang, Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models, Tech. rep., Massachusett Institute of Technology (2012).

[18] J. Fan, X. Han, W. Gu, Estimating false discovery proportion under arbitrary covariance dependence (with discussion), Journal of American Statistical Association 107 (3) (2012) 1019–1048.

[19] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (456) (2001) 1348–1360.

[20] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (5) (2008) 849–911.

[21] J. S. Yap, J. Fan, R. Wu, Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci, Biometrics 65 (4) (2009) 1068–1077.

[22] S. Geman, A limit theorem for the norm of random matrices, Annals of Probability 8 (2) (1980) 252–261.

[23] Y. Q. Yin, Z. D. Bai, P. R. Krishnaiah, On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix, Probability Theory Related Fields 78 (1988) 509–521.

[24] W. Wu, M. Pourahmadi, Nonparametric estimation of large covariance matrices of longitudinal data, Biometrika 90 (1) (2003) 831–844.

[25] P. Bickel, E. Levina, Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations, Bernoulli 10 (6) (2004) 989–1010.

[26] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, Journal of Econometrics 147 (1) (2008) 186–197.

[27] P. Bickel, E. Levina, Regularized estimation of large covariance matrices, Annals of Statistics 36 (1) (2008) 199–227.

[28] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Annals of Statistics 37 (2009) 42–54.

[29] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, Journal of the American Statistical Association 106 (494) (2011) 672–684.

[30] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants, Journal of Multivariate Analysis 98 (2) (2007) 227–255.

[31] J. Huang, N. Liu, M. Pourahmadi, L. Liu, Covariance matrix selection and estimation via penalised normal likelihood, Biometrika 93 (1) (2006) 85–98.

[32] E. Levina, A. Rothman, J. Zhu, Sparse estimation of large covariance matrices via a nested lasso penalty, The Annals of Applied Statistics (2008) 245–263.

[33] A. Rothman, E. Levina, J. Zhu, A new approach to cholesky-based covariance regularization in high dimensions, Biometrika 97 (3) (2010) 539–550.

[34] T. Cai, C. Zhang, H. Zhou, Optimal rates of convergence for covariance matrix estimation, Annals of Statistics 38 (4) (2010) 2118–2144.

[35] D. L. Donoho, J. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455.

[36] A. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, Journal of the American Statistical Association 104 (485) (2009) 177–186.

[37] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58 (1) (1996) 267–288.

[38] S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1) (1998) 33–61.

[39] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Annals of Statistics 38 (2) (2010) 894–942.

[40] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction (2nd ed.), Springer New York, 2009.

[41] A. Antoniadis, J. Fan, Regularization of wavelet approximations, Journal of the American Statistical Association 96 (455) (2001) 939–967.

[42] H. Liu, L. Wang, T. Zhao, Sparse covariance estimation with eigenvalue constraints, Journal of Computational and Graphical Statistics, to appear.

[43] L. Xue, S. Ma, H. Zou, Positive definite ℓ_1 penalized estimation of large covariance matrices, Journal of the American Statistical Association 107 (500) (2012) 1480–1491.

[44] A. Rothman, Positive definite estimators of large covariance matrices, Biometrika, to appear.

[45] A. Dempster, Covariance selection, Biometrics 28 (1972) 157–175.

[46] N. Meinshausen, P. Bühlmann, High dimensional graphs and variable selection with the lasso, Annals of Statistics 34 (3) (2006) 1436–1462.

[47] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.

[48] A. d'Aspremont, O. Banerjee, L. El Ghaoui, First-order methods for sparse covariance selection, SIAM Journal on Matrix Analysis and Applications 30 (1) (2008) 56–66.

[49] E. Candes, T. Tao, The dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics 35 (6) (2007) 2313–2351.

[50] M. Yuan, High dimensional inverse covariance matrix estimation via linear programming, Journal of Machine Learning Research 11 (8) (2010) 2261–2286.

[51] T. Sun, C.-H. Zhang, Sparse matrix inversion with scaled lasso, Tech. rep., Department of Statistics, Rutgers University (2012).

[52] W. Liu, X. Luo, High-dimensional sparse precision matrix estimation via sparse column inverse operator, arXiv/1203.3896.

[53] A. Belloni, V. Chernozhukov, L. Wang, Square-root lasso: Pivotal recovery of sparse signals via conic programming, Biometrika 98 (2012) 791–806.

[54] B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical bayes analysis of a microarray experiment, Journal of the American Statistical Association 96 (456) (2001) 1151–1160.

[55] S. Dudoit, J. Fridlyand, T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American statistical association 97 (457) (2002) 77–87.

[56] J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proceedings of the National Academy of Sciences 100 (16) (2003) 9440–9445.

[57] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 289–300.

[58] C. Genovese, L. Wasserman, A stochastic process approach to false discovery control, The Annals of Statistics 32 (3) (2004) 1035–1061.

[59] E. Lehmann, J. P. Romano, J. P. Shaffer, On optimality of stepdown and stepup multiple test procedures, Annals of statistics (2005) 1084–1108.

[60] B. Efron, Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, Vol. 1, Cambridge University Press, 2010.

[61] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Annals of statistics (2001) 1165–1188.

[62] J. D. Storey, J. E. Taylor, D. Siegmund, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66 (1) (2003) 187–205.

[63] S. Clarke, P. Hall, Robustness of multiple testing procedures against dependence, The Annals of Statistics 37 (1) (2009) 332–358.

[64] B. Efron, Correlation and large-scale simultaneous significance testing, Journal of the American Statistical Association 102 (477) (2007) 93–103.

[65] J. T. Leek, J. D. Storey, A general framework for multiple testing dependence, Proceedings of the National Academy of Sciences 105 (48) (2008) 18718–18723.

[66] B. Efron, Correlated z-values and the accuracy of large-scale statistical estimates, Journal of the American Statistical Association 105 (491) (2010) 1042–1055.

[67] A. Schwartzman, X. Lin, The effect of correlation in false discovery rate estimation, Biometrika 98 (1) (2011) 199–214.

[68] J. Fan, R. Tang, X. Shi, Partial consistency with sparse incidental parameters.

[69] J. Fan, X. Han, Estimation of false discovery proportion with unknown dependence, Manuscript.

[70] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statistica Sinica 20 (1) (2010) 101–148.

[71] P. Bühlmann, S. Van De Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer, 2011.

[72] S. A. Van de Geer, High-dimensional generalized linear models and the lasso, The Annals of Statistics 36 (2) (2008) 614–645.

[73] J. Fan, J. Lv, Nonconcave penalized likelihood with NP-dimensionality, Information Theory, IEEE Transactions on 57 (8) (2011) 5467–5484.

[74] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, Annals of statistics 36 (4) (2008) 1509–1533.

[75] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, The Annals of Applied Statistics 1 (2) (2007) 302–332.

[76] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communi-

cations on pure and applied mathematics 57 (11) (2004) 1413–1457.
[77] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences 2 (1) (2009) 183–202.
[78] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of statistics 32 (2) (2004) 407–499.
[79] J. Fan, R. Song, Sure independence screening in generalized linear models with np-dimensionality, The Annals of Statistics 38 (6) (2010) 3567–3604.
[80] P. Hall, H. Miller, Using generalized correlation to effect variable selection in very high dimensional problems, Journal of Computational and Graphical Statistics 18 (3) (2009) 533–550.
[81] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, Journal of the American Statistical Association 106 (494) (2011) 544–557.
[82] R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, Journal of the American Statistical Association 107 (499) (2012) 1129–1139.
[83] G. Li, H. Peng, J. Zhang, L. Zhu, Robust rank correlation based screening, The Annals of Statistics 40 (3) (2012) 1846–1877.
[84] J. Fan, R. Samworth, Y. Wu, Ultrahigh dimensional feature selection: beyond the linear model, The Journal of Machine Learning Research 10 (2009) 2013–2038.
[85] V. Vapnik, S. E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, Advances in neural information processing systems (1997) 281–287.
[86] Y. Freund, R. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, Springer, 1995, pp. 23–37.
[87] L. Breiman, Arcing classifier (with discussion and a rejoinder by the author), The annals of statistics 26 (3) (1998) 801–849.
[88] D. M. Witten, R. Tibshirani, Covariance-regularized regression and classification for high dimensional problems, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (3) (2009) 615–636.
[89] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis by thresholding for high dimensional data, The Annals of Statistics 39 (2) (2011) 1241–1265.
[90] T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, Journal of the American Statistical Association 106 (496) (2011) 1566–1577.
[91] J. Fan, Y. Feng, X. Tong, A road to classification in high dimensional space: the regularized optimal affine discriminant, Journal of the Royal Statistical Society: Series B (2012) 745–771.
[92] A. Wille, P. Zimmermann, E. Vranova, A. Frholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, P. Bühlmann, Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana, Genome Biology 5 (11) (2004) R92.
[93] T. Zhao, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, The huge package for high-dimensional undirected graph estimation in r, Journal of Machine Learning Research 13 (2012) 1059–1062.
[94] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, et al., Gene-expression profiles in hereditary breast cancer, New England Journal of Medicine 344 (8) (2001) 539–548.
[95] B. Efron, Correlation and large-scale simultaneous significance testing, Journal of the American Statistical Association 102 (477) (2007) 93–103.
[96] K. H. Desai, J. D. Storey, Cross-dimensional inference of dependent high-dimensional data, Journal of the American Statistical Association 107 (497) (2012) 135–151.
[97] H. Liu, F. Han, C. Zhang, Transelliptical graphical models, in: Advances in Neural Information Processing Systems 25, 2012, pp. 809–817.
[98] K. Fang, S. Kotz, K. Ng, Symmetric multivariate and related distributions, Chapman & Hall, 1990.
[99] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: Semiparametric estimation of high dimensional undirected graphs, Journal of Machine Learning Research 10 (2009) 2295–2328.
[100] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, High dimensional semiparametric gaussian copula graphical models, Annals of Statistics 40 (40) (2012) 2293–2326.
[101] L. Xue, H. Zou, Regularized rank-based estimation of high-dimensional nonparanormal graphical models, Annals of Statistics 40 (5) (2012) 2541–2571.