Estimating False Discovery Proportion Under Arbitrary Covariance Dependence

Jinxi Liu

October 26, 2017

Introduction

Jianqing Fan et al. proposed a method based on principal factor approximation, which successfully subtracts the common dependence and weakens significantly the correlation structure, to deal with an arbitrary dependence structure. They derived an approximate expression for false discovery proportion (FDP) in large scale multiple testing when a common threshold is used and provided a consistent estimate of realized FDP.

Basic idea

Consider the test statistics,

$$(Z_1, ..., Z_p)^T \sim N((\mu_1, ..., \mu_p)^T, \Sigma)$$

where σ is known and p is large. We would like to simultaneously test $H_{0i} : \mu_i = 0$ *vs* $H_{1i} : \mu_i \neq 0$ for $i = 1, ..., p$.

The basic idea is to first take out the principal factors that derive the strong dependence among observed data $Z_1, ..., Z_p$ and to account for such dependence in calculation of false discovery proportion (FDP).

Basic idea

This is accomplished by the spectral decomposition of Σ and taking out the largest common factors so that the remaining dependence is weak.

We then derive the asymptotic expression of the FDP, defined as V/R, that accounts for the strong dependence.

The realized but unobserved principal factors that derive the strong dependence are then consistently estimated. The estimate of the realized FDP is obtained by substituting the consistent estimate of the unobserved principal factors.

We are especially interested in estimating FDP under the high dimensional sparse problem, that is, p is very large, but the number of $\mu_i \neq 0$ is very small.

Estimating False Discovery Proportion

Assume that among all the p null hypotheses, p_0 of them are true and p_1 hypotheses are false, and p_1 is supposed to be very small compared to p, $p_0 \to \infty$ *when* $p \to \infty$.

For a fixed rejection threshold t, we will reject those P-values no greater than t and regard them as statistically significance.

$$FDP(t) = V(t)/R(t)$$

We need the following definition for weakly dependent normal random variables;

Suppose $(K_1, ..., K_p)^T \sim N((\theta_1, ..., \theta_P), A)$. Then $K_1, ..., K_p$ are called weakly dependent normal variables if

$$\lim_{p\to\infty} p^{-2} \sum_{i,j} |a_{ij}| = 0,$$

where a_{ij} denote the $(i,j)th$ element of the covariance matrix A.

Our procedure is called principal factor approximation (PFA). The basic idea is to decompose any dependent normal random vector as a factor model with weakly dependent normal random errors. The details are shown as follows.

Estimating False Discovery Proportion

Firstly apply the spectral decomposition to the covariance matrix Σ. Suppose the eigenvalues are $\lambda_1, ..., \lambda_p$ which have been arranged in decreasing order. If the corresponding orthonormal eigenvectors are denoted as $\gamma_1, ..., \gamma_p$ then

$$\Sigma = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T.$$

If we further denote $A = \sum_{i=k+1}^{p} \lambda_i \gamma_i \gamma_i^T$ for an integer k, then

$$||A||_F^2 = \lambda_{k+1}^2 + ... + \lambda_p^2$$

where $|| \cdot ||_F$ is the Frobenius norm.

Let $L = (\sqrt{\lambda_i}\gamma_1, ..., \sqrt{\lambda_i}\gamma_k)$,which is a $p \times k$ matrix. Then

$$\Sigma = LL^T + A,$$

and $Z_1, ..., Z_p$ can be written as

$$Z_i = \mu_i + b_i^T W + K_i, \ \ i = 1, ..., p, \tag{1}$$

where $b_i = (b_{i1}, ..., b_{ik})^T$, $(b_{1j}, ..., b_{pj})^T = \sqrt{\lambda_j}\gamma_j$, the factors are $W = (W_1, ..., W_k)^T \sim N_K(0, I_k)$ and the random errors are $(K_1, ..., K_p)^T \sim N(0, A)$. Furthermore,$W_1, ..., W_k$ are independent of each other and independent of $K_1, ..., K_p$.

Estimating False Discovery Proportion

FDP is only a function of $Z_1, ..., Z_p$. Although (1)is not exactly a classical multifactor model because of the existence of dependence among $K_1, ..., k_p$, we can show that $(K_1, ..., K_P)^T$ is a weakly dependent vector if the number of factors k is appropriately chosen.

Estimating False Discovery Proportion

Theorem 1

Suppose $(Z_1, .., Z_p)^T \sim N((\mu_1, .., \mu_p)^T, \Sigma)$. Choose an appropriate k such that

$$(C0) \qquad \frac{\sqrt{\lambda_{k+1}^2 + ... + \lambda_p^2}}{\lambda_{k+1} + ... + \lambda_p} = O(p^{-\delta}) \ \mathit{for}\ \delta > 0$$

Let $\sqrt{\lambda_j}\gamma_j = (b_{1j}, ..., b_{pj})^T$ *for* $j = 1, .., k$. Then,

$$\lim_{p\to\infty} \left\{ FDP(t) - \frac{\sum_{i \in \{true\ \ null\}} [\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))]}{\sum_{i=1}^p [\Phi(a_i(z_{t/2} + \eta_i + \mu_i)) + \Phi(a_i(z_{t/2} - \eta_i - \mu_i))]} \right\} = 0, a.s., \quad (2)$$

Estimating False Discovery Proportion

where $a_i = (1 - \sum_{h=1}^{k} b_{ih}^2)^{-1/2}$, $\eta_i = b_i^T W$ with $b_i = (b_{i1}, ..., b_{ik})$ and $W \sim N_k(0, I_k)$.

Note that condition (C0) implies that $K_1, , K_p$ are weakly dependent random variables, but (**??**eqq2)) converges to zero at some polynomial rate of p.

Estimating Realized FDP

In Theorem 1, the summation over the set of true null hypotheses is unknown. However, due to the high dimensionality and sparsity, both p and p_0 are large and p_1 is relatively small. Therefore, we can use

$$V(t) \approx \sum_{i=1}^{p} \left[\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))\right]$$

as a as a conservative surrogate for

$$\sum_{i \in \{true\ \ null\}} \left[\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))\right]$$

Estimating Realized FDP

Let

$$FDP_A(t) = \left(\sum_{i=1}^{p} \left[\Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i)) \right] \right) / R(t),$$

if $R(t) \neq 0$ and $FDP_A(t) = 0$ when $R(t) = 0$.

If the unobserved but realized factors $W_1, , W_k$ can be estimated by $\hat{W}_1, ..., \hat{W}_k$, then we can obtain an estimator of $FDP_A(t)$ by plugging in $\hat{\eta}_i = \sum_{h=1}^{k} b_{ih} \hat{W}_h$.

Estimating Realized FDP

The following procedure is one practical way to estimate W based on the data. For observed values $z_i, ..., z_p$, we choose the smallest 90% of $|z_i|'s$, say. For ease of notation, assume the first m $z_i's$ have the smallest absolute values. Then approximately

$$Z_i = b_i^T W + K_i \quad i = 1, ..., m \tag{3}$$

The approximation from (1)to (3) stems from the intuition that large $|\mu_i|'s$ tend to produce large $|z_i|'s$ and the sparsity makes approximation errors negligible.

Finally we apply the robust L1-regression to the equation set (3) and obtain the least absolute deviation estimates $\hat{W}_1, ..., \hat{W}_k$.