Effects of dependence in high-dimensional multiple testing problems

Kyung In Kim and Mark van de Wiel[†]

† Department of Mathematics, Vrije Universiteit Amsterdam.

Contents

High-dimensional multiple testing problem

Consider a multiple testing problem with m hypotheses and m_1 false null hypotheses.

n(the number of replicates) $\ll$ m(the number of hypotheses).

Controlling type I error rates adjusting for multiplicity is main concern.

Decision setting (Benjamini and Hochberg (1995)):

Decision	Declared non-significant	Declared significant	Total
true null	U	V	m_0
false null	T	S	m_1
	$m-R$	R	m

False Discovery Rate

- FDR (False Discovery Rate) is a popular type I error rate for multiple testing problems.
- FDR is defined as $\mathsf{E}[V/R]$, the the expected proportion of the number of falsely rejected hypotheses among total number of rejected hypotheses.
- Benjamini and Hochberg (1995) finds the maximal k such that $p_{(k)} \leq (k/m)\alpha$ where $p_{(1)}, \ldots, p_{(m)}$ are the ordered p-values.
- Benjamini and Hochberg (1995) is known to control

$$\text{FDR} \leq \frac{m_0}{m}\alpha = \pi_0\alpha \leq \alpha$$

.

- Effective estimations of π_0 can give more powerful results. (SAM, pFDR, Adaptive Benjamini-Hochberg etc)

Motivation

How do pairwise correlations affect the result of multiple testing problem?

1. Simulate 'general' dependence circumstances to see the correlation effects to FDR.

2. Examine the validity of various FDR implementations.

Modeling general dependence circumstances is difficult.

- Arbitrary pairwise correlations do not guarantee positive definiteness of correlation matrix.
- Equicorrelated model(single or block diagonal structure): simple, easy to understand but not realistic.
- Simple generation of random correlation matrices: too general and hard to compare.

Conditional independence structures in random correlation matrices are considered as a measure of dependence.(Whittaker (1990),Wille et al. (2004),Dobra et al. (2004))

Generating constrained random correlation matrices

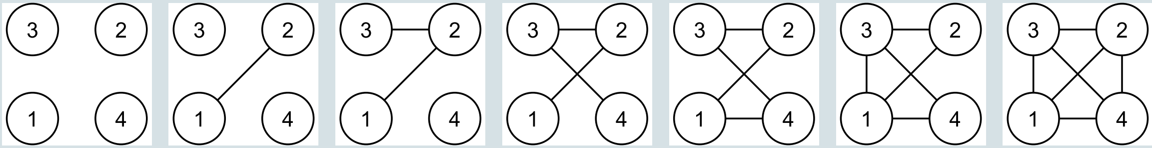
Goal: Generate a sequence of "nested" random correlation matrices with conditional independence structures.

Conditional independence: When $X = (X_1, \ldots, X_m)^T \sim N_m(\mu, \Sigma)$,

$$X_i \perp\!\!\!\perp X_j \mid \{\text{the rest variables}\} \quad \text{if and only if} \quad [\Sigma^{-1}]_{ij} = 0.$$

Example: When $m = 4$, maximally 7 "nested" random correlation matrices can be considered according to the proportions of non-zero partial correlations. Each random correlation matrices can be described by graphs.

Graphical representations of conditional independence structures

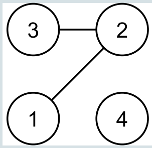


(a) $d = 0/6 = 0$ (b) $d = 1/6$ (c) $d = 2/6$ (d) $d = 3/6$ (e) $d = 4/6$ (f) $d = 5/6$ (g) $d = 6/6 = 1$

Inverse correlation matrices (* : non-zero elements)

$$\begin{bmatrix} * & 0 & 0 & 0 \\ & * & 0 & 0 \\ & & * & 0 \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & 0 & 0 \\ & * & 0 & 0 \\ & & * & 0 \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & 0 & 0 \\ & * & * & 0 \\ & & * & 0 \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & 0 & 0 \\ & * & * & 0 \\ & & * & * \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & 0 & * \\ & * & * & 0 \\ & & * & * \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ & * & * & 0 \\ & & * & * \\ & & & * \end{bmatrix} \begin{bmatrix} * & * & * & * \\ & * & * & * \\ & & * & * \\ & & & * \end{bmatrix}$$

Example: Construction of random correlation matrix with given structure



$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ & 1 & 1 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}$$

1. Generate $Z = [z_1\ z_2\ z_3\ z_4]$ where z_i is M-dimensional standard normal vector($M > m = 4$).
2. $\tilde{z}_1 = z_1$.
3. $\tilde{z}_2 = z_2$.
4. $\tilde{z}_3 = z_3 - P_3 z_3$ where $P_3 = \tilde{z}_1(\tilde{z}_1^T \tilde{z}_1)^{-1}\tilde{z}_1^T$.
5. $\tilde{z}_4 = z_4 - P_4 z_4$ where $P_4 = [\tilde{z}_1\ \tilde{z}_2\ \tilde{z}_3]([\tilde{z}_1\ \tilde{z}_2\ \tilde{z}_3]^T[\tilde{z}_1\ \tilde{z}_2\ \tilde{z}_3])^{-1}[\tilde{z}_1\ \tilde{z}_2\ \tilde{z}_3]^T$.
6. Let $\tilde{Z} = [\tilde{z}_1\ \tilde{z}_2\ \tilde{z}_3\ \tilde{z}_4]$. Then $\Sigma = (\tilde{Z}^T\tilde{Z})^{-1}$ is a random covariance matrix with constraint matrix J.

Controlling average correlation by M parameter

For unrestricted random covariance matrices, that is, $Z_{M\times m} = \tilde{Z}$, the expectation and the variance of pairwise correlation of random correlation matrices are

$$\begin{aligned} \mathsf{E}(\rho_{ij}) &= O((M-m+2)^{-2}), \\ \mathsf{var}(\rho_{ij}) &= \frac{1}{M-m+2} + O((M-m+2)^{-2}). \end{aligned}$$

If $\mathsf{var}(\rho_{ij})$ is small enough, we may expect the dependence structure of correlation matrix is almost same as the independence case.

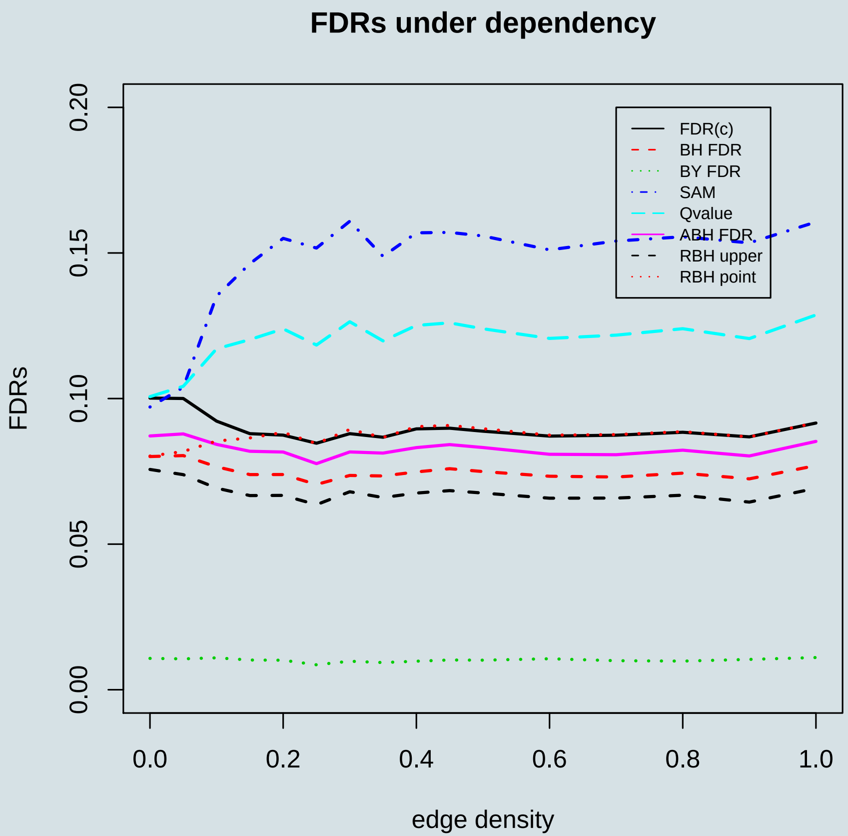
By controlling M, we can also control overall “correlation strength” of a random correlation matrix.

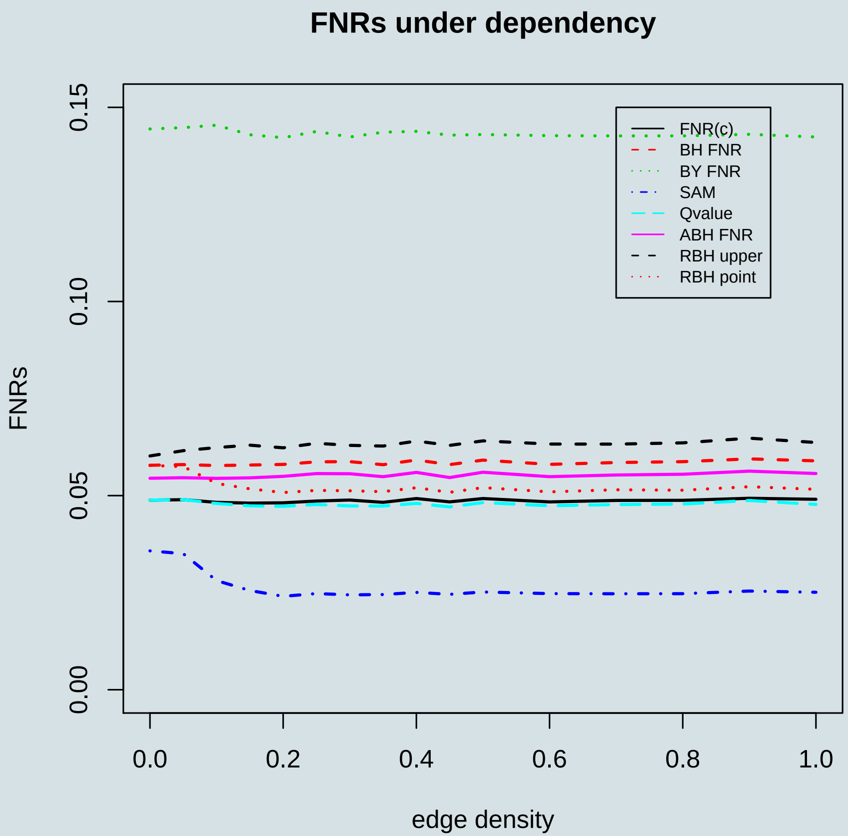
Simulation scheme

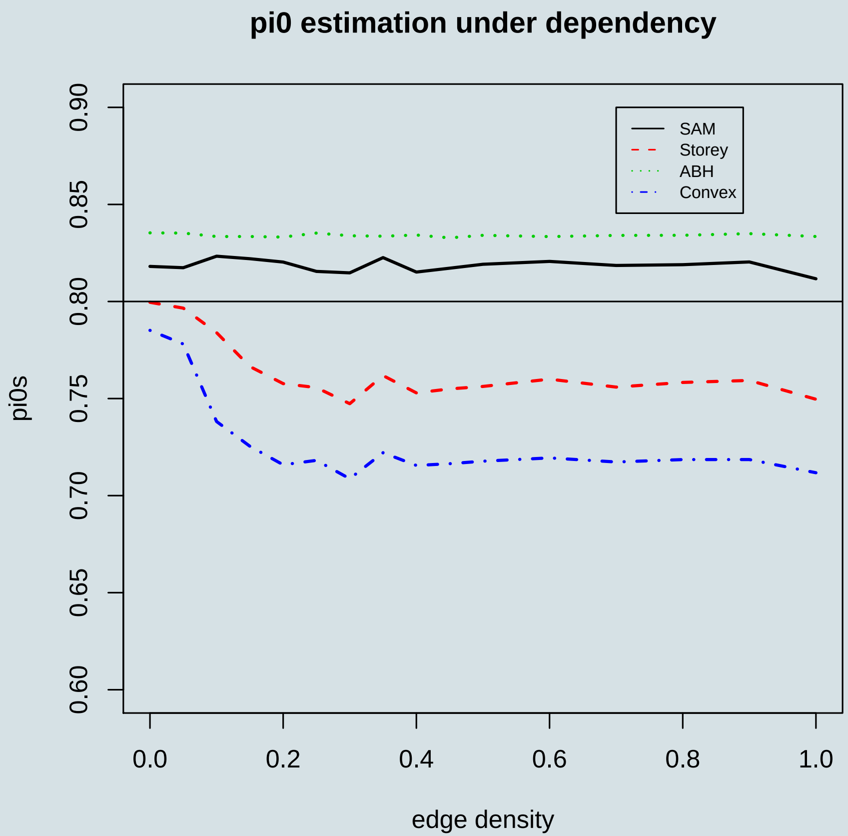
Purpose of this simulation is to investigate the effects of correlations for two-sample unpaired case.

1. Find c satisfying FDR(c) $= \alpha$ under independence assumption.
2. Generate random correlation matrices $\Sigma_1, \ldots, \Sigma_d$ from given structures.
3. For each Σ_j, $X_1, \ldots, X_{n_1} \sim N_m(\mu_X, \Sigma_j)$ and $Y_1, \ldots, Y_{n_2} \sim N_m(\mu_Y, \Sigma_j)$.
4. Apply various multiple testing methods to these data and compare their results of FDR, FNR and π_0 estimates.

Note $\mathrm{FNR} = \mathsf{E}[(m_1 - S)/(m - R)]$ (Genovese and Wasserman (2002)).


FDRs under dependency
FDR(c)
BH FDR
BY FDR
SAM
Qvalue
ABH FDR
RBH upper
RBH point
FDRs
0.00
0.05
0.10
0.15
0.20
0.0
0.2
0.4
0.6
0.8
1.0
edge density


FNRs under dependency
FNRs
edge density
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
FNR(c)
BH FNR
BY FNR
SAM
Qvalue
ABH FNR
RBH upper
RBH point


pi0 estimation under dependency
SAM
Storey
ABH
Convex
pi0s
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.0
0.2
0.4
0.6
0.8
1.0
edge density

Conclusions

1. Our simulation set-up allows for a structural study of the effect of dependencies on multiple testing criterions.
2. Most conventional implementations work well under independence assumption, but in the dependence conditions, they overestimate or underestimate FDR.
3. Benjamini-Hochberg type methods seem most robust in the dependence circumstances.
4. Adaptive methods are more powerful but estimates of π_0 depend on the dependence.

Bibliography

Benjamini, Yoav, and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**(1), 289–300.

Dobra, Adrian, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, **90**(1), 196–212.

Genovese, Christopher, and Larry Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**(3), 499–517.

Whittaker, Joe (1990). *Graphical models in applied multivariate statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Wille, Anja, Philip Zimmermann, Eva Vranova, Andreas Furholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelic, Peter von Rohr, Lothar Thiele, Eckart Zitzler, Wilhelm Gruissem, and Peter Buhlmann (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol*, **5**(11), R92.