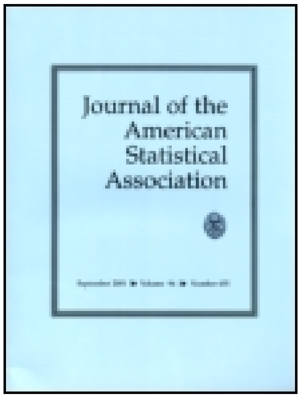

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Comment

Jiashun Jin [a]
[a] Department of Statistics, Baker Hall , Carnegie Mellon University , Pittsburgh , PA , 15213

Published online: 08 Oct 2012.

To cite this article: Jiashun Jin (2012) Comment, Journal of the American Statistical Association, 107:499, 1042-1045, DOI: 10.1080/01621459.2012.713878

To link to this article: http://dx.doi.org/10.1080/01621459.2012.713878

Comment

Jiashun JIN

I would like to congratulate Fan, Han, and Gu for a very interesting and thought-provoking article. The idea in this article sounds very promising to me, and I have no doubt that it could be useful in the areas of large-scale multiple testing and genomics. I also believe that the idea can be useful in solving many other problems in the area of high-dimensional data analysis, such as variable selection, low-rank matrix recovery, and sparse principal component analysis (PCA).

The article considers a p-dimensional Gaussian vector

$$Z \sim N(\mu, \Sigma), \tag{1}$$

where $\Sigma = \Sigma_{p\times p}$ is a correlation matrix that is assumed as known, and μ is a $p \times 1$ vector that is unknown but is presumably sparse in the sense that only a small fraction of its coordinates is nonzero. The primary interest is to test simultaneously that, for each $1 \le j \le p$, which of the following two hypotheses is true:

$$H_{0j}: \quad \mu_j = 0 \quad \text{versus} \quad H_{1j}: \quad \mu_j \neq 0. \tag{2}$$

Driven by the recent development of "big data" in many scientific areas, this problem has received a lot of attention in the past decade.

In practice, it is frequently desirable to control the so-called quantity of false discovery rate (FDR; see, e.g., Benjamini and Hochberg 1995). In the simplest case where Σ is the $p \times p$ identity matrix I_p, the problem is well understood, and two popular approaches are the FDR-controlling method by Benjamini and Hochberg (1995), and the local FDR-controlling approach by Efron et al. (2001). However, the case for general $\Sigma \neq I_p$ is much more challenging, and it remains an open problem to date.

Fan, Han, and Gu propose a very interesting approach that paves the way for solving the above problem in the case where Σ has spiky eigenvalues. In detail, suppose the eigenvalues of Σ are sparse in the sense that all eigenvalues are small except for a few very large spikes (such a situation can be found in many applications including, but not limited to, factor analysis). In this case, Σ can be well approximated by a low rank matrix, so the multiple testing problem is much easier to tackle. Fan, Han, and Gu (2012) further develop this idea and derive an elegant formula for the FDR that holds for general correlation matrix Σ. The approach is interesting both from theoretic and scientific point of views.

The article also raises some very interesting and closely related questions.

The first question is: What kind of role does the correlation matrix Σ play in the multiple testing? The study by Fan, Han, and Gu (2012) focuses on the false discovery proportion (FDP; as well as the FDR) associated with "marginal" methods. Using these methods, we calculate a (two-sided) p-value for each coordinate of Z as follows:

$$P_i = 2(1 - \Phi(|Z_i|), \qquad 1 \le i \le p,$$

where Φ is the cumulative distribution function of $N(0, 1)$. Now, for a threshold $t \in (0, 1)$, we call the ith hypothesis a discovery if and only if $P_i \le t$. The goal of Fan, Han, and Gu (2012) is to study the false discovery proportion $\mathrm{FDP}(t) = V(t)/R(t)$, where $V(t)$ is the number of false discoveries and $R(t)$ is the number of total discoveries. Seemingly, their study starts from p-values obtained "marginally," where the correlation structure of Σ is neglected. A natural question then is: Could we better the results of multiple testing by using the correlation structure in Σ?

The second question concerns the optimality of large-scale multiple testing. In the work by Fan, Han, and Gu (2012) and many recent works in large-scale multiple testing, the emphasis has been placed on how to control the FDR. While it is desirable to develop FDR-controlling procedures, we must note that merely controlling the FDR only tells one side of the story. What is more satisfying is to develop procedures that control the FDR at a prescribed level, say, $0 < q < 1$ (in the literature, q is referred to as the FDR control parameter), but at the same time have powers that is as large as possible (In the context of large-scale multiple testing, the power is referred to as the number of the true discoveries; see e.g., Genovese, Roeder, and Wasserman 2006). Procedures satisfying both properties can be viewed as optimal in multiple testing.

In this note, I would like to contribute some preliminary thoughts on the two intertwined questions above. The discussion covers a different angle from that by Fan, Han, and Gu (2012) where the eigenvalues of Σ are not spiky.

1. THE ROLE OF Σ IN LARGE-SCALE MULTIPLE TESTING

In this section, we use a simple example to illustrate that we can improve the results of multiple testing by using the matrix Σ properly.

In this example, we assume $p = 3m$ for some integer m, and let Σ be the following diagonal block-wise matrix:

$$\Sigma = \begin{pmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D \end{pmatrix}. \tag{3}$$

Jiashun Jin is Associate Professor, Department of Statistics, Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *jiashun@stat.cmu.edu*).

© 2012 American Statistical Association
Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.713878

In (3), D is a 3×3 equicorrelation matrix having the form of

$$D = (1-a)I_3 + a\mathbf{1}\mathbf{1}',$$

where I_3 is the 3×3 identity matrix and $\mathbf{1}$ is the 3×1 vector of ones. The parameter a satisfies $-0.5 < a < 1$ so D is positive definite. To be consistent with Fan, Han, and Gu (2012), we assume Σ as known, and so a is known.

At the same time, we model the vector μ as follows. Fixing $\epsilon \in (0,1)$ and $\tau > 0$, let k be the integer that is closest to $m\epsilon$. According to the partition of Σ, we partition the set $\{1, 2, \ldots, p\}$ into m different blocks: $\{1,2,3\}$, $\{4,5,6\}$, $\ldots$, $\{p-2, p-1, p\}$. We randomly generate k indices $1 \le j_1 < j_2 < \cdots < j_k \le m$, and let

$$S = \{3j_1 + 1, 3j_2 + 1, \ldots, 3j_k + 1\}.$$

We then model μ by

$$\mu_i = \begin{cases} \tau, & i \in S, \\ 0, & i \notin S. \end{cases}$$

Note that in this model, S is actually the support of μ. When ϵ is small and τ is moderately large, the model is an example of the so-called *rare and weak* signal model (see, e.g., Jin, Zhang, and Zhang 2012).

Suppose we are interested in using the FDR-controlling methods by Benjamini and Hochberg (1995). How to use this method depends on how we calculate the p-values. Below are three reasonable approaches to calculate the p-values.

In the first approach, we neglect the correlation structure in Σ, and compute the (two-sided) p-values by

$$P_i = 2[1 - \Phi(|Z_i|)], \qquad 1 \le i \le p. \tag{4}$$

This is the approach by Fan, Han, and Gu (2012) and many recent works in this area. To differentiate it from the approaches below, we call this the *naive approach*.

In the second approach, we first take the transformation

$$Z \mapsto \tilde{Z} \equiv \Sigma^{-1} Z.$$

We call this the innovated transform, as it is related to the notion of innovation in the context of time series (see, e.g., Hall and Jin 2009). It is seen that

$$\tilde{Z} \sim N(\Sigma^{-1}\mu, \Sigma^{-1}),$$

where

$$\Sigma^{-1} = \begin{pmatrix} D^{-1} & 0 & \ldots & 0 \\ 0 & D^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & D^{-1} \end{pmatrix}, \qquad \text{with}$$

$$D^{-1} = \frac{1}{1-a} I_3 - \frac{a}{(1-a)(1+2a)} \mathbf{1}\mathbf{1}'.$$

Especially, all diagonals of Σ^{-1} equal to $(1+a)/[(1-a)(1+2a)]$. As a result, we can compute the (two-side) p-values by

$$\tilde{P}_i = 2\left(1 - \Phi\left(|\tilde{Z}_i| / \sqrt{(1+a)/[(1-a)(1+2a)]}\right)\right), \qquad 1 \le i \le p. \tag{5}$$

Still another approach is based on whitening. In this approach, we first take the transform of

$$Z \mapsto \tilde{\tilde{Z}}_i \equiv \Sigma^{-1/2} Z.$$

Note that $\tilde{\tilde{Z}} \sim N(\Sigma^{-1/2}\mu, I_p)$ and the noise is whitened. A natural way to calculate the (two-sided) p-values is then

$$\tilde{\tilde{P}}_i = 2(1 - \Phi(|\tilde{\tilde{Z}}_i|)), \qquad 1 \le i \le p.$$

It turns out that the innovated transform could largely improve the results of multiple testing. The whitening approach could also improve the results, but the improvement is not as large as that of the innovated transform. For this reason, we compare only the naive approach and the innovated approach below.

Toward this end, we note that by our constructions, in the naive approach,

$$Z_i \sim \begin{cases} N(\tau, 1), & i \in S, \\ N(0,1), & i \notin S, \end{cases}$$

and in the innovated approach,

$$\frac{\tilde{Z}_i}{\sqrt{(1+a)/[(1-a)(1+2a)]}} \sim \begin{cases} N(\sqrt{(1+a)/[(1-a)(1+2a)]}\tau, 1), & i \in S, \\ N(-a\tau/\sqrt{(1-a^2)(1+2a)}, 1), & i-1 \in S, \\ N(-a\tau/\sqrt{(1-a^2)(1+2a)}, 1), & i-2 \in S, \\ N(0,1), & \text{otherwise}. \end{cases}$$

In other words, we have the following observations regarding Z and $\tilde{Z}/\sqrt{(1+a)/[(1-a)(1+2a)]}$.

- At a signal location (i.e., an index $i \in S$), the innovated transform increases the (marginal) signal-to-noise ratio (SNR) by a factor of $\sqrt{(1+a)/[(1-a)(1+2a)]}$; note for all $a \neq 0$ and $-0.5 < a < 1$, the constant $\sqrt{(1+a)/[(1-a)(1+2a)]} > 1$.
- At a location corresponding to a noise (i.e., an index $i \notin S$) but in the same block of a signal location, the vector $\tilde{Z}/\sqrt{(1+a)/[(1-a)(1+2a)]}$ may contain a "signal" (i.e., $\Sigma^{-1}\mu$ is nonzero at this location). We call this signal a *fake signal* as $\mu_i = 0$.

The increased SNR at all signal locations imply that we can have larger testing powers if we apply the Benjamini and Hochberg procedure to the p-values computed from Equation (4) instead of from Equation (5).

For illustration, we have conducted a small-scale simulation experiment, in which we compare the receiver operating characteristic (ROC) corresponding to the vector Z and the transformed vector $\tilde{Z}/\sqrt{(1+a)/[(1-a)(1+2a)]}$. In this experiment, we take $p = 3000$ and four different combinations of (ϵ, τ, a), where $(\epsilon, \tau, a) = (0.1, 1, 0.75)$, $(0.025, 15, 0.75)$, $(0.2, 1.5, 0.5)$, and $(0.15, 1.25, -0.45)$. The ROC curves are displayed in Figure 1, where the results in each panel are based on one replication, but similar results hold for multiple repetitions.

The ROC curves suggest that the innovated transform can substantially improve the power of multiple testing, especially when local correlations are strong. We mention that a more sophisticated method can be developed to filter out the "fake

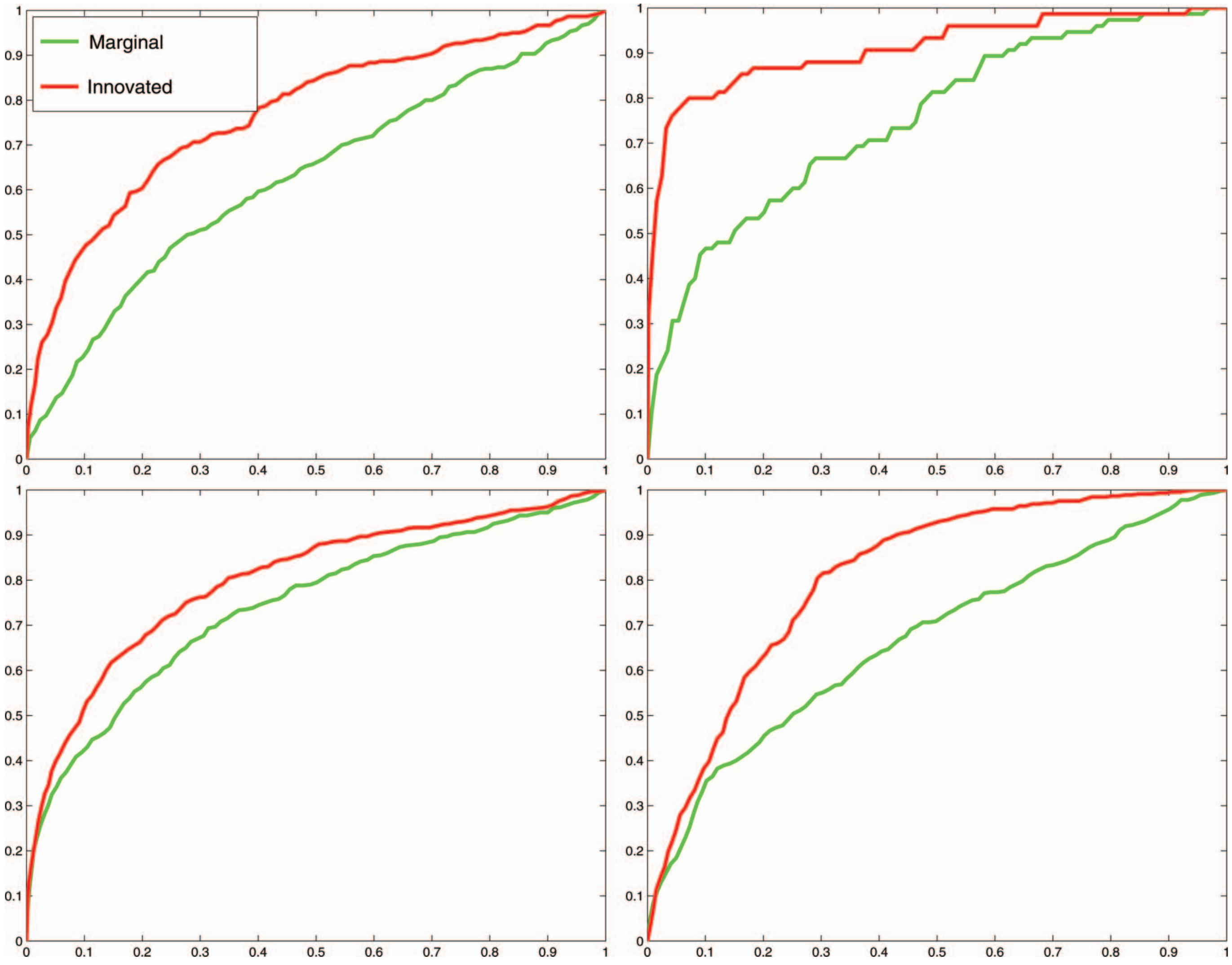

Figure 1. ROC curves associated with Z (green) and $\tilde{Z}/\sqrt{(1+a)/[(1-a)(1+2a)]}$ (red). x-axis: false positive rate; y-axis: true positive rate. From top to bottom, left to right: $(\epsilon, \tau, a) = (0.1, 1, 0.75), (0.025, 15, 0.75), (0.2, 1.5, 0.5), (0.15, 1.25, -0.45)$. The online version of this figure is in color.

signals" created artificially by the transform. In that case, the difference between two ROC curves in each panel can be even larger, in the favor of the innovated transform.

How general are the above findings? In the literature by Hall and Jin (2009), we have carefully investigated this, where the following lemma plays a key role.

Lemma 1.1 For any positive definite matrix Σ with unit diagonals, all diagonals of Σ^{-1} are at least as large as 1.

The lemma follows directly from Cholesky factorization so we omit the proof. In the literature by Hall and Jin (2009), we showed that the aforementioned increases in (marginal) SNR can be found in a broad context. We termed this phenomenon as: *the correlation is a blessing rather than a curse* (see details therein).

2. OPTIMAL PROCEDURES FOR MULTIPLE TESTING

The above discussion suggests that by using the correlation matrix Σ properly, we could improve the power of the Benjamini and Hochberg procedure (and also many other methods driven by the p-values). This raises the question: What could be the optimal procedure for large-scale multiple testing? Clearly, this is a challenging problem. In this section, we suggest some ideas that could be useful.

Toward this end, note that if we let $X = \Sigma^{-1/2}$ and $Y = \Sigma^{-1/2}Z$, then Equation (1) can be equivalently rewritten as a linear regression model

$$Y = X\mu + z, \qquad z \sim N(0, I_p). \tag{6}$$

In this model, the problem of large-scale multiple testing (Equation (2)) is the same as the problem of variable selection.

For any variable selection procedure $\hat{\mu}$, it is natural to measure the risk with the *weighted-L^0*-loss. In detail, fix a weight parameter

$$\lambda > 0.$$

The weighted L^0-risk is defined as the (expected) weighted sum of Type I and Type II errors:

$$wH_p(\hat{\mu}, \lambda) = \left[\sum_{j=1}^{p} P(\hat{\mu}_j \neq 0, \mu_j = 0)\right] + \lambda\left[\sum_{j=1}^{p} P(\hat{\mu}_j = 0, \mu_j \neq 0)\right].$$

The parameter λ is the ratio between the cost of making a Type II error over that of making a Type I error.

Fix $\lambda > 0$. It is desirable to obtain variable selection procedures that minimize the weighted L^0-risk. This is a well-known challenging problem. However, recently, progresses have been made in the case where the Gram matrix

$$G = X'X$$

is sparse, or is sparsifiable (see e.g., Jin, Zhang, and Zhang 2012; Ke, Jin, and Fan 2012). Note that in Equation (6), the Gram matrix coincides with the concentration matrix Σ^{-1}.

Let $\hat{\mu}_\lambda$ be an optimal variable selection procedure. It can be argued (e.g., Sun and Cai 2007) that in a general setting,

- $\hat{\mu}_\lambda$ controls FDR at a level $q = q(\lambda) > 0$, asymptotically, and
- among all procedures that control FDR at the level $q = q(\lambda)$, $\hat{\mu}_\lambda$ is asymptotically most powerful.

Therefore, variable selection procedures that optimize the weighted L^0-risk simultaneously control the FDR at some levels and maximize the testing power. This suggests an intimate relationship between optimal procedures for multiple testing and optimal procedures for variable selection, and the solution of one is also the solution of the other. Note that, however, the mapping $\lambda \mapsto q(\lambda)$ may have a complicated form.

3. OTHER COMMENTS

Below are two additional comments to the authors.

- It seems that the work focuses on the case where the eigenvalues of Σ is spiky. What happens if Σ is structured in a different way, say, Σ is a Toeplitz matrix where the eigenvalues are not spiky. Is the approach extendable to such situations?
- How robust is the proposed approach to the Gaussianity of the noise?

4. CONCLUSION

This is a very nice article that provides a fresh perspective in solving a long-standing hard problem in large-scale multiple testing. I particularly like the idea of reducing Σ to a low rank matrix by exploiting the sparsity of the eigenvalues.

On the other hand, the article focuses on estimating the FDR associated with some types of marginal methods. It seems that exploiting the graphical structure of Σ^{-1} could be very helpful, especially when the eigenvalues of Σ is not spiky. I wonder what is the authors' opinion on the role of Σ and on what could be done in the line of pursuing "optimal" procedures for large-scale multiple testing.

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society,* Series B, 57, 289–300. [1042,1043]

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160. [1042]

Genovese, C., Roeder, K., and Wasserman, L. (2006), "False Discovery Control With *p*-value Weighting," *Biometrika*, 93, 509–524. [1042]

Hall, P., and Jin, J. (2009), "Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise," *The Annals of Statistics*, 38, 1686–1732. [1043,1044]

Jin, J., Zhang, C.-H., and Zhang, Q. (2012), "Optimality of Graphlet Screening in High Dimensional Variable Selection," arXiv:0465.053. [1043,1045]

Ke, T., Jin, J., and Fan, J. (2012), "Covariance Assisted Screening and Estimation," arXiv:1205.4645. [1045]

Sun, W., and Cai, T. (2007), "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of the American Statistical Association*, 102, 901–912. [1045]