Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Rejoinder

Jianqing Fan , Xu Han & Weijie Gu

To cite this article: Jianqing Fan , Xu Han & Weijie Gu (2012) Rejoinder, Journal of the American Statistical Association, 107:499, 1046-1048, DOI: 10.1080/01621459.2012.720911

To link to this article: https://doi.org/10.1080/01621459.2012.720911

Rejoinder

Jianqing FAN, Xu HAN, and Weijie GU

Controlling false discovery rate (FDR) in large-scale multiple testing problems has received significant attention in statistics and bioinformatics, thanks to the arrival of high-throughput genomics and genetic data. The article by Benjamini and Hochberg (1995) has been cited 15,000 times, according to Google Scholar. The impact of the field on scientific discovery can easily be seen. Since the publication of Benjamini and Hochberg, hundreds if not thousands of articles on the subject have been published. Many early articles focus on the situation where the test statistics are independent or weakly dependent. Then, the work was extended to the dependent situations but focused predominantly on the robustness of the FDR control to the dependence. Recently, efforts have been made to use the dependence of the test statistics to account for FDR and to improve the efficiency of the test. However, no satisfactory answers have been found. It is against this background that three laymen of the field work together to find a generally applicable method for consistently estimating the false discovery proportion (FDP) and controlling FDR. We begin by solving the problem for the case when the covariance matrix is known. In our follow-up manuscript (Fan, Han, and Gu 2012), we address the problem of how accurately the covariance matrix of dependence needs to be estimated in order for the substitution method to work and what kind of covariance structure is necessary to obtain this accuracy. These two sister articles give a comprehensive view of how to consistently estimate FDP and how to control FDR, when the covariance dependence is unknown. They also provide formal statistical frameworks to the FDR control for correlated tests, pioneered by Efron (2007, 2010).

1. FDR AND FDP

FDP refers to the false discovery proportion in a particular experiment, whereas the FDR is the expected FDP across many repeated experiments. It is clear that scientists are more interested in estimating the FDP for a particular experiment and in controlling the FDP by adjusting the critical value or thresholding level t.

Using the notation of Section 3.1 of the main article, in the definition of $\mathrm{FDP}(t) = V(t)/R(t)$, only $V(t)$, the number of falsely discovered null hypotheses, is unknown. When the test statistics are independent, it is the number of successes in p_0 independent Bernoulli trials and should always be close to $p_0 t$, in relative terms. Therefore, $\mathrm{FDP}(t) = V(t)/R(t) \approx p_0 t/R(t)$ for large-scale hypothesis testing problems. On the other hand, when the test statistics are dependent, $V(t)$ is the number of successes in a sequence of dependent Bernoulli trials, and its value can vary significantly from one experiment to another, depending on the degree of the dependence. In our decomposition, the dependence depends on the principal factor $\mathbf{W}$ in (10). Thanks to the sparsity of $\{\mu_i\}$, the realized principal factor $\mathbf{W}$ can be estimated consistently for each experiment, and hence $V(t)$ can be estimated with precision from each experiment, if the remaining variables K_i are weakly dependent.

A very simple example to illustrate the main idea behind our method is Example 1 in the main article, in which

$$Z_i = \sqrt{\rho} W + \sqrt{1-\rho} K_i$$

for the true null hypotheses. For simplicity of notation, consider the one-sided tests that reject the ith null hypothesis when $Z_i > u$. In this case,

$$\begin{aligned} V(u) &= \sum_{\text{true null}} I(\sqrt{\rho} W + \sqrt{1-\rho} K_i > u) \\ &= p_0 \Phi((-u + \sqrt{\rho} W)/\sqrt{1-\rho}) + O_P\big(p_0^{1/2}\big), \end{aligned}$$

where the central limit theorem is applied to the sequence of independent Bernoulli trials for each given W. When $\rho = 0$, $V(u) \approx p_0 \Phi(-u)$. When $\rho \neq 0$, the result depends on the realization of W. For example, when $\rho = 0.64$, $p_0 = 1000$, and $u = 2.5$, we have

$$V(u) \approx p_0 \Phi((-2.5 + 0.8W)/0.6),$$

which is approximately 0, 2.3, 66.8, and 433.8 for $W = 0, 1, 2,$ and 3, respectively. This is in contrast to $V(u) \approx 13.5$ for the case $\rho = 0$. Thus, the number of false discoveries depends critically on the realization of W. Yet, this realized value of W can be well estimated from each experiment: $W \approx \bar{Z}/\sqrt{\rho}$ and hence the FDP can be better estimated for each experiment.

We truly appreciate Professor Armin Schwartzman's conscientious effort in explaining the main idea behind our technique and its relation to Efron (2007) and Schwartzman (2010). The comparisons and comments are fair and insightful. They help readers understand better the intuition behind our technique. We are also grateful to his efforts to demonstrate that the FDP is

Jianqing Fan is Frederick L. Moore'18 professor, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, and honorary professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China (E-mail: *jqfan@princeton.edu*). Xu Han is assistant professor, Department of Statistics, Fox Business School, Temple University, Philadelphia, PA 19122 (E-mail: *hanxu3@temple.edu*). Weijie Gu is graduate student, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544 (E-mail: *wgu@princeton.edu*). The article was completed while Xu Han was a postdoctoral fellow at Princeton University. This research was partly supported by National Science Foundation NSF grants DMS-0704337 and DMS-1206464 and National Institutes of Health (NIH) grant R01-GM072611. We would like to thank the editors Leonard A. Stefanski and Xuming He for organizing this stimulating discussion, with a conscientious effort to invite outstanding researchers from diverse backgrounds that make the discussion more thought-provoking. We are also very grateful to all discussants (Aurore Delaigle, Peter Hall, Jiashun Jin, Armin Schwartzman, and Larry Wasserman) for their insightful and stimulating comments, touching practical, methodological, and philosophical aspects of large-scale hypothesis testing problems, and offering deep insights to the problem. Their contributions are very timely and helpful. We treasure the opportunity to respond the questions raised by the discussants.

Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.720911

not only more relevant, but also easier to estimate. For example, for the equiv-correlation test statistics, Schwartzman and Lin (2011) showed that FDR cannot be consistently estimated, whereas we show in Example 1 that the FDP can be consistently estimated. In general, the mathematics behind this is that

$$\operatorname{var}(V(t)) \geq \operatorname{var}(V(t)|\mathbf{W}).$$

Therefore, the variability of $V(t)$ from one realization to another is in general larger than the conditional one. The latter is similar to the independent Bernoulli trials (with varying probability of success, assuming $\{K_i\}$ are weakly dependent). The unconditional variance of $V(t)$ is shown in Table 3 of the main article, which can be much larger than the conditional variance.

2. ESTIMATING REALIZED PRINCIPAL FACTORS

The success of the estimation of FDP depends very much on the precision of estimating the realized principal factors $\mathbf{W}$. Three techniques are proposed in Section 3.2. The most natural one is the penalized least-squares, which minimizes

$$\sum_{i=1}^{p} \left(Z_i - \mu_i - \mathbf{b}_i^T \mathbf{W}\right)^2 + \sum_{i=1}^{p} p_\lambda(|\mu_i|).$$

Given $\mathbf{W}$, the above minimization can easily be found. For example, when the L_1-penalty is used, $\hat{\mu}_i$ is just the soft-thresholding of $Z_i - \mathbf{b}_i^T \mathbf{W}$. Substituting this into the above equation, we obtain the profile least-squares as

$$\sum_{i=1}^{p} \psi\left(Z_i - \mathbf{b}_i^T \mathbf{W}\right),$$

where ψ is Huber's robust ψ-loss function. See the recent manuscript by Fan, Tang, and Shi (2012) where the asymptotic behavior of $\mathbf{W}$ is thoroughly studied. In other words, the penalized least-squares in this setting is simply a form of the robust regression. This establishes the link between the second approach and the first approach (robust-based approach) in the main article. From this connection, the L_1-regression can even be applied to the whole data, namely, in (22) of the main article, m can even be taken as p.

The simplest approach to understand the estimability of $\mathbf{W}$ is the least-squares approach. As shown in Theorem 4, the realized factor can be estimated more precisely when the eigenvalues are spiked. However, deviating somewhat from Professor Jin's comment, our applications do not limit us to the covariance $\mathbf{\Sigma}$ with spiked eigenvalues. For example, if all eigenvalues are an order of magnitude smaller than $\sqrt{p}$, Theorem 1 allows us to take $k = 0$ and we do not need to estimate any realized principal factors. In other words, the solution is the same as the weak dependence case. In general, as we remarked after Theorem 4, we need only to estimate the realized principal factors with $\lambda_k > c\sqrt{p}$. As long as $k/\sqrt{p} \to 0$, namely, the large eigenvalues with magnitude of order at least $O(\sqrt{p})$ concentrate on the first $o(\sqrt{p})$ locations, our method continues to work.

3. POWER CONSIDERATIONS

Professor Jiashun Jin contributes some nice insights on power improvements. He is right that when the test statistics are correlated, the power of the test can be improved. He shows that the innovated transform and prewhitening can improve the power of the tests (without transform). However, the p-values computed under three different transforms have very different meaning, as their null hypotheses are very different. His covariance matrix has a very special block-diagonal structure. It is not clear whether the improved power is due to the structure or due to the transform. We hasten to add that the dependence does help in improving the power. A simple approach was outlined in Section 3.4 by extracting the common dependence structure out, which results in improving the signal-to-noise ratios. Conditioning is another possible approach to improve the power of the test.

We also appreciate the connections made by Professor Jin, between the multiple testing with dependent statistics and the sparse linear model. His optimality criterion seems reasonable. We welcome further contributions along this line.

4. FAMILYWISE ERROR RATE (FWER)

Professor Larry Wasserman raises a number of excellent questions. Among them, how does our PFA compare with the method of van der Laan, Dudoit, and Pollard (2004) and Genovese and Wasserman (2006). First of all, their method is designed to control the k-FWER, that is,

$$P(\text{reject at least } k \text{ true hypotheses } H_i) \leq \alpha$$

or $P(\text{FDP} > c) \leq \alpha$ for a given c. This controls the upper bound of FDP with high confidence rather than the point estimation and is a nice framework. But the suggested procedure does not use at all the covariance information. Indeed, the suggested procedure is conservative and crude: after an initial procedure that controls the 1-FWER, it rejects in addition the next k most significant hypotheses not rejected thus far. It makes the worst case calculation in FDP: the k next most significant hypotheses are all true null hypotheses. Moreover, the k additional hypotheses are always rejected, even if the corresponding test statistics are not significant even without the Bonforroni adjustment. In contrast, our method uses the covariance information and consistently estimates the FDP.

Professor Wasserman also asks about the connection between our method and that of Romano and Wolf (2007). Romano and Wolf's method was also constructed to control the k-FWER, but it incorporates the dependence information from the test statistics via bootstrap and other resampling methods. There are no obviously direct connections between these two approaches, since they aim at somewhat different goals. Our method is to consistently estimate the realized FDP in a given experiment and their method can only control FDP via a probabilistic statement $P(\text{FDP} > c) \leq \alpha$. It is not clear whether their upper limit is tight.

5. ROBUSTNESS TO ASSUMPTIONS

Professors Aurore Delaigle and Peter Hall raise an excellent question about the robustness of the results to the normality assumption. This question is also echoed by Professor Jiashun Jin. The normality assumption is an idealization, like many other assumptions such as independent realizations of the data. It is equivalent to the assumption made in the literature on the large-scale hypothesis testing that the p-value is computed precisely

under the null hypothesis. The senior author of the article is aware of this restrictive assumption, as evidenced in the work by Fan, Hall, and Yao (2007). On the other hand, these kind of assumptions are frequently made in the literature, including by readers and discussants. Even under this idealized assumption, statisticians do not have good tools to assess FDP consistently. That is the background of the article.

In general, for large-scale statistical learning, including multiple hypothesis testing, results are sensitive to the assumptions on the underlying distributions. However, for our problem, $V(t)$ and $R(t)$ are the sum of many weakly dependent random variables and the impact of individual random variable is limited. Therefore, we would expect that there is a certain degree of robustness in the applications of our method. The robustness of the FDP to the normality assumption deserves a thorough investigation.

We are grateful to Professors Delaigle and Hall for providing extra insights when the normality assumption fails. The examples they gave are the situations under which the test statistics S_j in their Equation (5) are concentrated on a particular component. Since our inferences are conditional, the weights $\{X_j^i - \bar{X}_j\}$ are known in advance, and the data analysts should know in advance whether the applications of the central limit theorem are reasonable for their problems. Nevertheless, we appreciate the caveat given by Professors Delaigle and Hall.

We would like to thank to Professors Delaigle and Hall for the comments on the elementary results in Section 2. Their results are correct and so are ours. The difference lies in notation. For example, in Equation (1), the expectation is meant to be taken conditioning on $\{X_j^i\}_{i=1}^n$ and our s_{ii} denotes the standard deviation whereas theirs are sample variance. Their comment led us to change the text accordingly to make things clearer.

6. MISCELLANEOUS

Professor Larry Wasserman raises the question on the interpretability of the marginal regression coefficients. The marginal regressions are used in the main article only to show that the dependence of test statistics can be known. We did not invent the method. The marginal regressions are frequently used in the study of Quantitative Trait Loci (QTL) or expression QTL (eQTL). There are several reasons for the use of such a marginal regression approach. First of all, the signal-to-noise ratio is too low for using a high-dimensional multiple regression with genotypes as regressors. Second, geneticists are more interested in understanding which SNPs have correlation with the outcome (phenotype or gene expressions). This is a more realistic question than a multiple regression that attempts to quantify the contributions of each SNP. A nonvanishing marginal regression coefficient is equivalent to a nonvanishing correlation coefficient. Third, the nonvanishing marginal regression coefficients provide useful probes to their contributions in the multivariate regression. This has been proved recently by Fan and Lv (2008) and extended by Fan and Song (2010).

Professor Larry Wasserman raises a very provocative question whether "we are all focusing too much on interpreting and testing parameters." That is indeed a very good question. Even for multiple regression, we have problems of nonlinearity, measurement errors, endogeneity, missed variables in the regression equations, among others. These make the interpretation of coefficients difficult. On the other hand, disciplinary scientists, from biologists and epidemiology to economists and social scientists, do pay a lot of attention to the interpretation and testing parameters. As we mentioned earlier, according to Google Scholar, the multiple testing article by Benjamini and Hochberg (1995) has been cited 15,000 times. This shows the need for hypothesis testing in various scientific disciplines. On the other hand, for many machine learning problems, such as text classification and pattern recognition, it is not important at all to interpret and test parameters. It is not even important which variables are chosen as long as the resulting rules have good prediction power. The question can then also be asked if our profession focuses too much on risk properties, without interpreting and testing parameters, even when we solve a disciplinary science problem.

REFERENCES

Fan, J., Hall, P., and Yao, Q. (2007), "To How Many Simultaneous Hypothesis Tests Can Normal, Student's t or Bootstrap Calibration be Applied?" *Journal of the American Statistical Association*, 102, 1282–1288. [1048]

Fan, J., Han, X., and Gu, W. (2012), "Estimation of FDP With Unknown Dependence," unpublished manuscript. [1046]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [1048]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1048]

Fan, J., Tang, R., and Shi, X. (2012), "Partial Consistency in Linear Model With a Sparse Incidental Parameters," unpublished manuscript. [1047]

Schwartzman, A. (2010), Comment on "Correlated z-Values and the Accuracy of Large-Scale Statistical Estimates" by Bradley Efron, *Journal of the American Statistical Association*, 105, 1059–1063. [1046]

Schwartzman, A., and Lin, X. (2011), "The Effect of Correlation in False Discovery Rate Estimation," *Biometrika*, 98, 199–214. [1047]