- Practical estimates for the approximation formulas are derived and demonstrated through simulations and application to a microarray study (Sections 3 and 4).
- Delta-method arguments are used to extend the cdf results to more general summary statistics (Sections 3 and 4).
- Under reasonable assumptions, it is shown that z scores tend to have nearly normal distributions, even in nonnull situations (Section 5), justifying application of the theory to studies in which the individual variates are z-values.

Our main conclusion is that by dealing with normal variates, a practical assessment of large-scale correlation effects on statistical estimates is possible.

[Received March 2009. Revised June 2009.]

REFERENCES

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003), "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias," *Bioinformatics*, 19, 185–193. [1054]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1043]

Csörgő, S., and Mielniczuk, J. (1996), "The Empirical Process of a Short-Range Dependent Stationary Sequence Under Gaussian Subordination," *Probability Theory and Related Fields*, 104, 15–25. [1046]

Desai, K., Deller, J., and McCormick, J. (2009), "The Distribution of Number of False Discoveries for Highly Correlated Null Hypotheses," *The Annals of Applied Statistics*, to appear. [1043,1046]

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004), "Multiple Testing. I. Single-Step Procedures for Control of General Type I Error Rates," *Statistical Applications in Genetics Molecular Biology*, 3, Article 13 (electronic). [1043]

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103. [1042]

Efron, B. (1985), "Bootstrap Confidence Intervals for a Class of Parametric Problems," *Biometrika*, 72, 45–58. [1054]

——— (1987), "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association*, 82, 171–200. [1052,1053]

——— (2007a), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [1042-1046,1049,1051,1054]

——— (2007b), "Size, Power and False Discovery Rates," *The Annals of Statistics*, 35, 1351–1377. [1047-1051,1053]

——— (2008), "Microarrays, Empirical Bayes and the Two-Groups Model" (with discussion), *Statistical Science*, 23, 1–22. [1047-1050,1053]

——— (2009), "Are a Set of Microarrays Independent of Each Other?" *The Annals of Applied Statistics*, 3, 922–942. [1045,1047]

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. DOI: *10.1126/science.286.5439.531*. [1042,1043]

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*, New York: Springer-Verlag. [1052]

Johnson, N. L., and Kotz, S. (1970), *Distributions in Statistics. Continuous Univariate Distributions*, Vol. 1, Boston, MA: Houghton Mifflin. [1047,1051,1052]

Lancaster, H. O. (1958), "The Structure of Bivariate Distributions," *The Annals of Mathematical Statistics*, 29, 719–736. [1044]

Owen, A. B. (2005), "Variance of the Number of False Discoveries," *Journal of the Royal Statistical Society, Ser. B*, 67, 411–426. [1042,1043]

Qiu, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005), "The Effects of Normalization on the Correlation Structure of Microarray Data," *BMC Bioinformatics*, 6, 120. DOI: *10.1186/1471-2105-6-120*. [1043]

Qiu, X., Klebanov, L., and Yakovlev, A. (2005), "Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology for Finding Differentially Expressed Genes," *Statistical Applications in Genetics and Molecular Biology*, 4, Article 34 (electronic). [1043]

Schwartzman, A., and Lin, X. (2009), "The Effect of Correlation in False Discovery Rate Estimation," Biostatistics Working Paper 106, Harvard University. [1043]

Westfall, P., and Young, S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, New York: Wiley-Interscience. [1042]

Comment

T. Tony CAI

Professor Efron has given us an interesting article on the effects of correlation which is an important issue in multiple testing. He is to be congratulated for his significant contributions to large-scale multiple testing.

Much of the research on multiple testing has been focused on the independent case and many practical testing procedures have been developed under the independence assumption. However, in many interesting applications observations are correlated. It is known that correlation has significant effects on a multiple testing procedure. For example, both the expectation and variance of the number of Type I errors can be seriously affected by the correlation among the test statistics (see Finner and Roters 2002 and Owen 2005). Correlation can also substantially deteriorate the performance of many FDR procedures. Previous research on the effects of correlation in large-scale multiple testing has been mostly focused on the validity of various testing procedures under dependency. For example, Benjamini and Yekutieli (2001), Farcomeni (2007), and Wu (2009) show that the FDR is controlled at the nominal level by the BH step-up and adaptive p-value procedures under different dependence assumptions.

Among various aspects of the correlation effects on an FDR procedure, the validity issue is often overemphasized. FDR procedures developed under the independence assumption, even valid, may suffer from substantial efficiency loss when the dependence structure is ignored. These situations include the geographical disease mapping studies, multiple-stage clinical trials, functional Magnetic Resonance Imaging analyses and comparative microarray experiments, where the nonnull cases are often structured in some way, for example, correlated temporally, spatially, or functionally.

A critical step in the implementation of many FDR procedures is the estimation of several important quantities such as the proportion of the nonnull hypotheses, the empirical null distribution, and the true FDR level. Developing good estimators

T. Tony Cai is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *tcai@wharton.upenn.edu*). Research supported in part by NSF FRG grant DMS-0854973.

for these quantities is a challenging task, even in the independent case.

The present paper continues a prolific line of research of Professor Efron on multiple testing, but with a different focus from his previous papers. Efron (2007) considered the correlation effects on the null distribution of the z-values and suggested that an adjusted FDR estimate should be combined with the use of an Lfdr procedure to remove the bias caused by the correlation. The present paper focuses on the effects of correlation on certain summary statistics used in large-scale multiple testing procedures. It is demonstrated that when the test statistics are correlated, the distribution of the z-values and its functionals can be conveniently estimated with adjustments for correlation without the need of knowing the whole covariance matrix, which itself is very difficult, if not impossible, to estimate in the current setting. A key step of the proposed estimation procedure is the use of a clever approximation based on Mehler's identity and the root mean square correlation. The approximation can be efficiently carried out, which makes the method practical for applications.

The convenient correlation adjustments proposed in the paper are obtained through a sequence of approximations. It is illustrated in the paper that the method works well in several examples. It is of great interest to further study the precision of these estimators and to understand how well the targeted quantities can be optimally estimated under reasonable assumptions. Perhaps even more important questions are how to best use these estimators to construct more efficient multiple testing procedures under dependency and how the performance of these estimators affects the performance of the subsequent testing procedures.

As mentioned earlier, the estimation step plays an important role in a multiple testing procedure. Several alternative estimation methods have been developed in the literature. For example, in the independent case, Cai and Jin (2010) developed a frequency domain approach based on the empirical characteristic function and Fourier analysis for the estimation of both the parameters associated with the null distribution f_0 and the proportion of the nonnull effects p_0. These estimators were shown to attain the optimal rates of convergence under regularity conditions. In the correlated case, the estimators were shown to be uniformly consistent over a wide class of parameters when the dependency is short ranged or strongly mixing (see Jin and Cai 2007). Numerical results also showed that the estimators perform favorably in comparison to other existing methods.

It is true that, as pointed out in the paper, "correlation usually degrades statistical accuracy, an important question for the data analyst being the severity of its effects on the estimates and tests at hand." However, in many settings correlation effects can also be positive on the outcomes of a testing procedure. Intuitively it is clear that the dependency structure among hypotheses is highly informative in simultaneous inference and can be exploited to construct more efficient tests. For example, in comparative microarray experiments, it is found that changes in expression for genes can be the consequence of regional duplications or deletions, and significant genes tend to appear in clusters. Therefore, when deciding the significance level of a particular gene, the observations from its neighborhood should also be taken into account.

It is possible to construct significantly better multiple testing procedures in the correlated settings by modeling the dependency structure. Sun and Cai (2009) considered multiple testing under a particular dependency structure, the hidden Markov model (HMM). The HMM is an effective tool for modeling the dependency structure and has been widely used in areas such as speech recognition, signal processing as well as analysis biological sequences and processes. Using a compound decision theoretical framework, an oracle testing procedure is developed in an ideal setting where the HMM parameters are assumed to be known. Under mild conditions, the oracle procedure is shown to be optimal in the sense that it minimizes the false nondiscovery rate (FNR) subject to a constraint on the FDR. This approach is distinguished from the conventional methods in that the proposed procedure is built on a new test statistic (local index of significance, LIS) instead of the p-values. Unlike p-values, the LIS takes into account the observations in adjacent locations by exploiting the local dependency structure in the HMM. The precision of individual tests is hence improved by utilizing the dependency information.

A data-driven procedure is then constructed to mimics the oracle procedure by plugging in consistent estimates of the unknown HMM parameters. The data-driven procedure is shown to be asymptotically optimal in the sense that it attains both the FDR and FNR levels of the oracle procedure asymptotically. Numerical studies indicate the favorable performance of the LIS procedure. These findings show that the correlation among hypotheses can be highly informative in large-scale simultaneous inference and can be exploited to construct more efficient testing procedures.

Much research is still needed in order to fully understand the correlation effects on the accuracy of estimators used in multiple testing procedures as well as the testing procedures themselves under general dependency structures. The present paper raises important questions and will definitely stimulate new research in the future.

ADDITIONAL REFERENCES

Benjamini, Y., and Yekutieli, D. (2001), "The Control of False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1055]

Cai, T., and Jin, J. (2010), "Optimal Rates of Convergence for Estimating the Null and Proportion of Non-Null Effects in Large-Scale Multiple Testing," *The Annals of Statistics*, 38, 100–145. [1056]

Farcomeni, A. (2007), "Some Results on the Control of the False Discovery Rate Under Dependence," *Scandinavian Journal of Statistics*, 34, 275–297. [1055]

Finner, H., and Roters, M. (2002), "Multiple Hypotheses Testing and Expected Number of Type I Errors," *The Annals of Statistics*, 30, 220–238. [1055]

Jin, J., and Cai, T. (2007), "Estimating the Null and the Proportion of Non-Null Effects in Large-Scale Multiple Comparisons," *Journal of the Amererican Statistical Association*, 102, 495–506. [1056]

Sun, W., and Cai, T. (2009), "Large-Scale Multiple Testing Under Dependence," *Journal of the Royal Statistical Society, Ser. B*, 71, 393–424. [1056]

Wu, W. (2009), "On False Discovery Control Under Dependence," *The Annals of Statistics*, 36, 364–380. [1055]