Comment: Correlated z-values and the accuracy of large-scale statistical estimates

Ruth Heller[1]

Professor Efron has given us an interesting article on how to quantify the uncertainty in summary statistics of interest in large scale problems, when the summary statistics are based on correlated normal variates. It is shown that the inflation in the accuracy estimate due to correlation among the normal variates cannot be ignored (except possibly at the very far tails of distributions).

Using a series of simplifications of the covariance formula, a simple formula is derived and it is shown in a numerical example that the approximation is indeed very close to the truth. In particular it is shown that the entire correlation structure is captured by one parameter α, the rms correlation. Several methods of estimating α, as well as the other unknown parameters, are suggested.

In what follows I will discuss several topics in large scale significance testing that are related to the results of this paper.

Summary values of interest when z-values are correlated In large scale significance testing, methods that control or estimate the false discovery rate are often applied to identify the set of interesting discoveries. Professor Efron suggested the following two summary statistics and estimated their accuracy: the estimated tail-area false discovery rate $\hat{Fdr}(x) = p_0 F_0(x)/\hat{F}(x)$ and the estimated local false discovery rate $\hat{fdr}(x) = p_0 f_0(x)/\hat{f}(x)$. Keeping the same notations as in the manuscript, $F_0(\cdot)$ and $\hat{F}(\cdot)$ are the null and the empirical survival curve of the z-values.

[1]*Address for correspondence:* Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel. E-mail: ruheller@techunix.technion.ac.il.

Another summary statistics of interest is the following quantity that "monotonizes" the $\hat{Fdr}(z)$ curve: $\overline{Fdr}(z) = \inf\{z' \leq z : \hat{Fdr}(z')\}$. In practice, the cut-off value x is often chosen to be the smallest z-value so that $\hat{Fdr}(z) \leq q$, where q is a desired fraction chosen to be typically small (e.g. q=0.05,0.1,0.25), and the hypotheses with z-values above this cut-off x are reported as interesting discoveries. The same cut-off value x is selected when choosing the smallest z-value so that $\overline{Fdr}(z) \leq q$. This practice coincides with the celebrated Benjamini and Hochberg's false discovery rate controlling procedure ([Benjamini and Hochberg, 1995]), henceforth referred to as the BH procedure, up to the factor p_0, conservatively taken as 1.

The BH procedure appears to control the false discovery rate in most circumstances that are not highly artificial ([Yekutieli, 2008], [Romano et al., 2008]). When the test statistics are correlated, the false discovery proportion (FDP, the fraction of discoveries from null hypotheses out of all discoveries) of the specific data set at hand may be very high even though the false discovery rate, $FDR = E(FDP)$, is controlled at the nominal level q on average over (hypothetical) replications of the study ([Pawitan et al., 2006], [Efron, 2007]). For a given data set, the interest of the investigator is in the FDP, not the FDR. When the rms correlation is non-negligible, the FDP may be much higher than q.

Similarly, for a given dataset and a given cut-off value x, the interest is in the false discovery proportion $FDP(x)$, the fraction of z-scores above x from null hypotheses out of all z-scores above x. The variance of $FDP(x)$ depends critically on the correlation among the z-values. When the correlation is weak and the effect sizes are large, the variability of $FDP(x)$ is tight around its expectation, $FDR(x) = E(FDP(x))$. In this favorable setting there may be interest in the quantities $FDR(x)$, $Fdr(x) = p_0F_0(x)/F(x)$ or $fdr(x) = p_0f_0(x)/f(x)$. However, when the correlation is high or the effect sizes are small, interest may no longer be in $FDR(x)$, $Fdr(x)$ or $fdr(x)$ but in

the unknown random quantity $FDP(x)$. The estimated $Fdr(x)$ (and its variability) may therefore not be of interest when the estimated rms correlation in non-negligible. However, from [Efron, 2007] and [Pawitan et al., 2006] it appears that using the histogram of z-scores, it may be possible to identify whether the $FDP(x)$ is indeed much higher than expected for a given dataset.

A simulation study to illustrate the effect of correlation on the FDP Each of 1000 data sets was generated as follows: $N = 1000$ genes with expression values from two classes with parameters $(p_0, \mu_0, \sigma_0) = (0.95, 0, 1)$ and $(p_1, \mu_1, \sigma_1) = (0.05, 1, 1)$; 40 cases were generated each from $X_i \sim MVN(\vec{0}, \Sigma)$ and 40 cases were generated each from $X_i \sim MVN(\vec{\mu}, \Sigma)$; the first 50 entries in $\vec{\mu}$ are one and the remaining 950 entries are zero; Σ is a block diagonal correlation matrix, each block of size 200 with symmetric correlation of 0.5. The data matrix $\mathbf{X}$ was either standardized by subtracting off its column-wise means (as done in Professor Efron's paper) or left unstandardized. The correlation in each data set was substantial: the rms correlation was $\alpha = 0.2$. For comparison, 1000 datasets were also generated under independence (i.e. Σ was the identity matrix).

The z-score for row i was $z_i = \Phi^{-1}(F_{78}(t_i))$, where $t_i = (\sum_{j=41}^{80} x_{ij}/40 - \sum_{j=1}^{40} x_{ij}/40)/\hat{SE}$ is the t-statistic for comparing the mean of the last 40 cases with that of the first 40 cases, and F_{78} and Φ are the cumulative distribution functions for a Student-t distribution with 78 degrees of freedom and a standard normal respectively. The BH procedure was applied at nominal level $q = 0.01, 0.02, \ldots, 0.25$ to the z-scores in each data set. Figure 1 top shows the 50th, 75th, and 95th quantiles of the FDP for each q after applying the BH procedure to each of 1000 simulated data sets when the data matrix was not standardized (left) and when it was standardized (right). For comparison, the same analysis was repeated when the data was independent and the results are displayed in Figure 1 bottom. The blue line in Figure 1 is the average

FDP, the red line is the nominal FDR level $0.95 * q$. The variability of the FDP was much higher when the data was correlated than when it was independent. The standardization of X reduced the variability of the FDP. The variability in the correlated case was reduced as the nominal value q decreased. The average FDP was below the nominal level for all q as expected, and moreover it was almost the same as the nominal level when X was not standardized and the data was independent. When the data was standardized yet independent, the average FDP was below the nominal level since after standardization the p-values were no longer independent, nor were they uniformly distributed but had instead a distribution that was stochastically larger than the uniform.

For each dataset, $FDP(x)$ and $\hat{Fdr}(x)$ were computed for the following cut-off values: $x = 2.00, 2.25, 2.50, 2.75, 3.00, 3.50, 4.00, 5.00, 6.00$. Table 1 shows summary statistics of $FDP(x)$ and $\hat{Fdr}(x)$ for the correlated case as well as for the independent case (the columns of the data matrix X were not standardized). The average and standard deviation are summarized. For $FDP(x)$, which may be highly skewed in the correlated case, the 50th, 85th and 95th quantiles are also summarized. While the average $FDP(x)$ was below the average $\hat{Fdr}(x)$ for the correlated case, the variability of $FDP(x)$ was very large for the smaller x's and diminished as x increased. For the independence case, the average $FDP(x)$ was very close to the average $\hat{Fdr}(x)$, and the variability of $FDP(x)$ was much smaller compared with the correlated case. For example, the effect of correlation cannot be ignored at $x = 2.25$: for correlated data, the average $FDP(2.25)$ was 0.15, but the 85th and 95th quantiles of $FDP(2.25)$ were respectively 0.31 and 0.48; for independent data, the average $FDP(2.25)$ was 0.19, and the 85th and 95th quantiles were respectively 0.23 and 0.26.

Correlation in relation to Signal In Professor Efron's paper it is assumed that the correlation is not informative of where the signal lies. However, this assump-

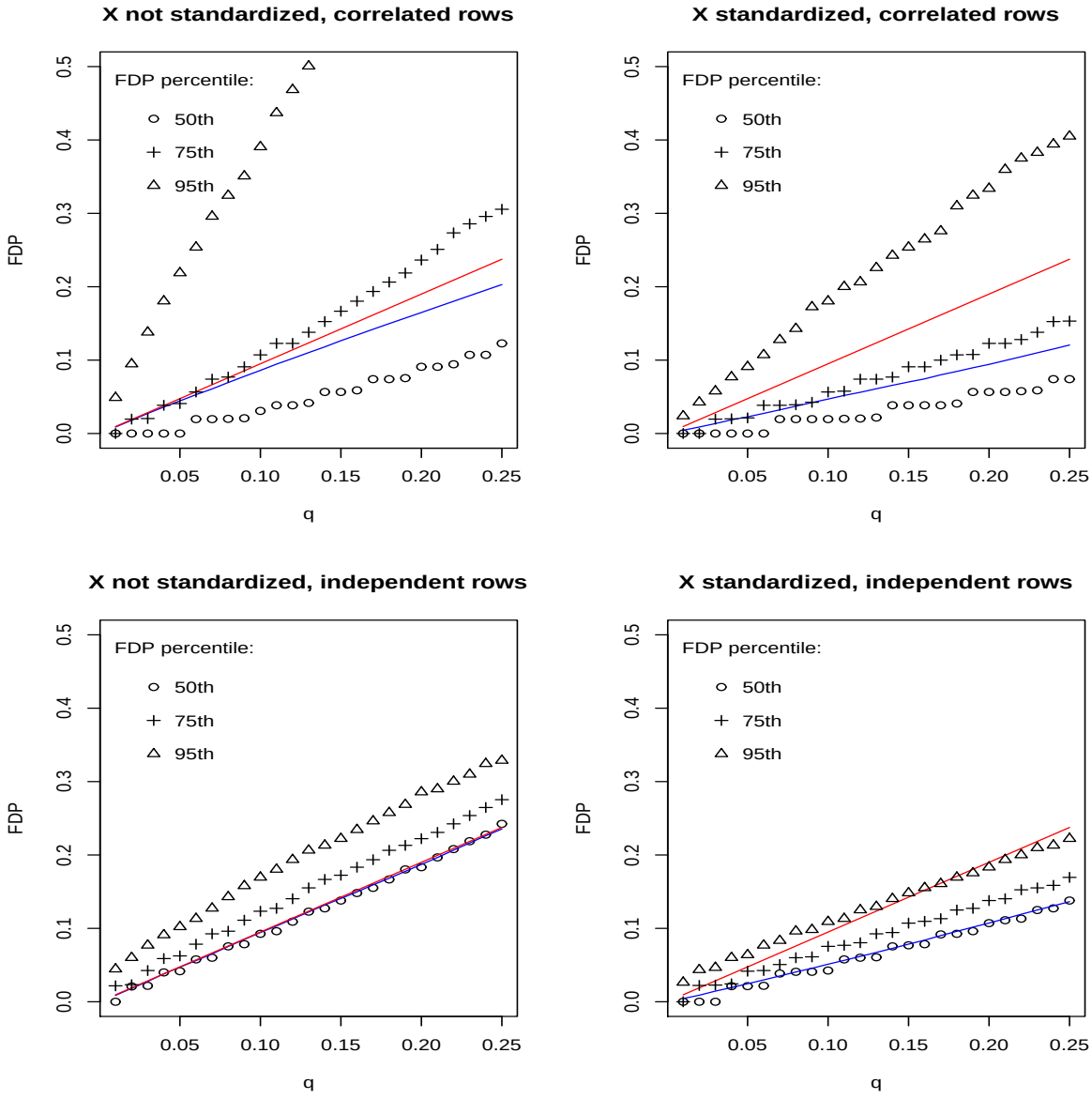

Figure 1: A plot of the 50th, 75th, and 95th FDP percentiles for each level q after applying the BH procedure to each of 1000 simulated data sets. Solid red line is the nominal level $0.95 * q$, blue line is the average FDP.

	Correlated rows		Independent rows	
	$FDP(x)$	$\hat{Fdr}(x)$	$FDP(x)$	$\hat{Fdr}(x)$
x	$mean(SD), Q_{0.5}, Q_{0.85}, Q_{0.95}$	$mean(SD)$	$mean(SD), Q_{0.5}, Q_{0.85}, Q_{0.95}$	$mean(SD)$
2.00	0.24 (0.18), 0.21, 0.45, 0.61	0.33 (0.08)	0.30 (0.04), 0.30, 0.34, 0.37	0.31 (0.02)
2.25	0.15 (0.15), 0.10, 0.31, 0.48	0.20 (0.04)	0.19 (0.05), 0.19, 0.23, 0.26	0.19 (0.01)
2.50	0.09 (0.12), 0.04, 0.18, 0.33	0.11 (0.02)	0.11 (0.04), 0.11, 0.14, 0.17	0.11 (0.01)
2.75	0.05 (0.09),0.02, 0.09, 0.21	0.06 (0.01)	0.05 (0.03),0.06, 0.09, 0.11	0.06 (0.00)
3.00	0.02 (0.06),0.00, 0.04, 0.12	0.03 (0.01)	0.03 (0.02), 0.02, 0.04, 0.06	0.03 (0.00)
3.50	0.01 (0.03),0.00, 0.00, 0.04	0.01 (0.01)	0.00 (0.01), 0.00, 0.02, 0.03	0.01 (0.00)
4.00	0.00 (0.01),0.00, 0.00, 0.00	0.00 (0.00)	0.00 (0.01),0.00, 0.00, 0.00	0.00 (0.00)

Table 1: Summary statistics of $FDP(x)$ and $\hat{Fdr}(x)$ for the correlated case as well as for the independent case (the columns of the data matrix X were not standardized). The average and standard deviation are summarized. For $FDP(x)$, which may be highly skewed in the correlated case, the 50th, 85th and 95th quantiles are also summarized.

tion may not always apply. For example, different genes may cluster into groups that participate in the same molecular functions or biological process and exhibit high correlation. If these groups are known a-priori, this knowledge can be incorporated into the multiple comparisons procedure to gain power (e.g. [Benjamini and Heller, 2007], [Heller et al., 2009]). Aggregates of statistical estimates within each group (and their accuracy) can be useful in this setting. Incorporating the correlation structure without a-priori knowledge of the grouping is a greater challenge ([Sun and Cai, 2009] model the unknown correlation structure assuming a hidden Markov model for the hypotheses).

In closing, I congratulate Professor Efron for the interesting article, and I thank the editor for giving me an opportunity to contribute to the discussion.

References

[Benjamini and Heller, 2007] Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *JASA - Theory and Methods*, 102(480):1272–1281.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

[Efron, 2007] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102 (477):93–103.

[Heller et al., 2009] Heller, R., Manduchi, E., Grant, G., and Ewens, W. (2009). A flexible two stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics*, 25 (8):1019–1025.

[Pawitan et al., 2006] Pawitan, Y., Calza, S., and Ploner, A. (2006). Estimation of false discovery proportion under general dependence. *Bioinformatics*, 22 (24):3025–3031.

[Romano et al., 2008] Romano, J., Shaikh, A., and Wolf, M. (2008). Rejoinder on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17 (3):461–471.

[Sun and Cai, 2009] Sun, W. and Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society, Series B*, 71 (2):393–424.

[Yekutieli, 2008] Yekutieli, D. (2008). Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17 (3):458–460.