ON ESTIMATING THE PROPORTION OF FALSE HYPOTHESES IN MULTIPLE TESTING PROCEDURE

Bobosharif Shokirov

bobosari@karlin.mff.cuni.cz
Department of Probability and Statistics, Faculty of Mathematics and Physics, Charles University in Prague

SUMMARY

The poster presents an approach for estimating the proportion of false positives - rejection of the null hypotheses in multiple testing procedure when they are true. Under independence assumptions, the explicit form the estimator of ratio of the false hypotheses to number of all hypotheses is given. Some of properties of the estimator is discussed and the given results are illustrated by simulation.

INTRODUCTION

The usual approach in multiple testing problems is to control the family-wise error rate (FWER), the probability of committing one or more false rejection [5]. There are many alternatives to FWER (for example, Bonferroni method). In [6] is proposed a step-down multiple testing procedure (MTP) that has more power than Bonferroni method, but still controls FWER at the same level. In [16] is designed a step-down resampling algorithm to control FWER, where to gain more power the dependence of test statistics is utilized. Once the number of hypotheses is huge, rejection of the individual hypothesis have little chance to be detected, therefore, in [9] it is suggested to replace FWER by so-called k-FWER.

In this sequel a new concept, proposed in [2], was false discovery rate (FDR), defined as the expected value of false discovery proportion (FDP), which is the number of false rejections, divided by the total number of rejections. A modification of FDR, so-called positive false discovery rate (pFDR) was introduced in [13] as conditional expected value of FDP on the event that positive findings have occurred. Introduced in [1] so-called sequential goodness of fit (SGoF) metatest, where given α, assuming all n nulls are true, it compares the observed number k to the expected number of rejections $n \times \alpha$. As SGoF test is used exact binomial and for $n \geq 100$ its approximation by chi-squared distribution with one degree of freedom. Rejection occurs if $k \geq k_\alpha$: given α, k_α, SGoF concludes that $k - k_\alpha + 1$ hypotheses are false. FDR rejects all hypotheses with $p_i \leq \alpha i/n$ (cf. [3]) while the SGoF test does not decide which hypotheses are erroneously rejected. Although under SGoF method the test power does not decrease once the number of hypotheses increases, it can not be precise on the number of rejections as it only informs about false rejections relative to the expected number of rejections. Based on the empirical distribution function of the p-values the lower bound λ of the proportion of false hypotheses with the property $\mathbb{P}(\hat{\lambda} \leq \lambda) \geq 1 - \alpha$ was constructed ([10]- [11]), asserting that the proportion of false null hypotheses is at least $\hat{\lambda}$. Closely related to the last by using different approach we focus on estimating the number of false hypotheses.

Let for $i = 1, \ldots, n$ we test n null hypotheses $H_0^i : F_i(x) = G_i(x)$ against $H_A^i : F_i(x) \neq G_i(x)$ with independent test statistics T_i and let as a result of n independent comparisons we obtain a sample $Z_1, Z_2, \ldots, Z_n$ of size n, where $Z_i = p_i \in [0,1], \quad i = 1, 2, \ldots, n$ and p_i be corresponding to i-th hypotheses p-value. Now taking n p-values as an independent random variables, we test the hypothesis H_0: sample $Z_1, Z_2, \ldots, Z_n$ has a df $F(x)$ against H_A: it has a df $G(x)$. Let k denotes the number of p-values from H_A. We are interested to know k in testing n hypotheses. Clearly, some $Z_{k_j}, 1 \leq k_j \leq n$ correspond to df $G(x)$ and others to $F(x)$. If H_0 is true then $k = 0$ and if H_A is true then $k = n$. Therefore, we focus on cases when $0 < k < n$.

Although there are approaches ([3], [14]-[15]) where control of multivariate analog of the univariate type I error are proposed under certain dependence assumptions and applications with strongly dependent data (for instance, gene expression data [7]-[8]), we assume independence. Dependent data could be reduced to weakly dependent or almost independent ([12], [7]).

MOTIVATION AND DESCRIPTION OF THE METHOD

Let us have n points from the interval $[0,1]$. Then the ratio of the average number of points less than x to the number of all points will approximately be equal to x, that is, one can suggest that this number is distributed as $\mathbb{U}(0,1)$. Take n points (random numbers) from $[0,1]$. Assume that $n-k$ of them $\sim \mathbb{U}[0,1]$ and remaining $k \sim \mathbb{U}[0, 1-\delta]$ ($\delta > 0$). Now we take $x \in [0,1]$ and would like to test a hypothesis: $x \sim \mathbb{U}[0,1]$ against $x \sim \mathbb{U}[0, 1-\delta]$. Then the share of points from the null hypothesis which are greater than x would approximately be equal to $1-x$, this is, (number of points $> x$)$/(n-k) \approx 1-x$ and the share of points from the null hypothesis which are less than x would approximately be equal to x, this is, (number of points $< x$)$/(n-k) \approx x$. Then the total number of points which are less then x approximately equals to $x(n-k)+k$. Thus, we have the distribution of the rv x on the whole interval $[0,1]$.

Now we take a random number x from the interval $[0,1]$ and would like to test a hypothesis $H_0 : x \sim F(x)$ against $H_A : x \sim G(x)$ (not only uniform as before). Let $Z_1, Z_2, \ldots, Z_n$ be n random numbers on $[0,1]$; $n-k$ of them produce sample $\mathbb{X} = (X_1, X_2, \ldots, X_{n-k})$ and the remaining k sample $\mathbb{Y} = (Y_1, Y_2, \ldots, Y_k)$. Let $T_{nZ}(x)$ denotes the "average number" of points greater than $x \in [0,1]$, $T_{nX}(x)$ be the "average number" of points greater than x corresponding to the null hypothesis and $T_{nY}(x)$ the "average number" of points greater than x corresponding to the alternative. Then, clearly,

$$T_{nZ}(x) = T_{nX}(x) + T_{nY}(x), \tag{1}$$

where $T_{nZ}(x) = \sum_{j=1}^{n} I_{\{Z_j > x\}}$; $I_{\{Z_j > x\}}$ is the indicator function of the event $\{Z_j > x\}$.

Denote $T_Z(x) = \mathbb{E}[T_{nZ}(x)]$. Since $T_{nX}(x)$ and $T_{nY}(x)$ are the "average number of points"then their expected value give the "exact" number of points: $\mathbb{E}[T_{nX}(x)] = (n-k)(1-F(x))$ and $\mathbb{E}[T_{nY}(x)] = k(1-G(x))$. From (1) we get

$$T_Z(x) = (n-k)\,(1-F(x)) + k\,(1-G(x))\,. \tag{2}$$

We cannot estimate k in such settings, therefore we assume that

(A1) $G(x) > F(x), \quad \forall x \in [0,1]$

(A2) $\mathrm{supp} G(x) \subset [0, 1-\delta], \quad$ for some $\quad \delta > 0.$

By virtue of $(A2)$ from equation (2) for the estimator of the ratio k/n we obtain

$$p^*(x) = 1 - \frac{T_Z(x)}{n(1-F(x))}. \tag{3}$$

So, we take $p^*(x)$ as an estimator of the ratio $k/n = p$. Now if we look back into the equation (2), we get

$$H_p(x) = (1-p)\,(1-F(x)) + p\,(1-G(x))\,, \tag{4}$$

where $H_p(x) = T_Z(x)/n$. Thus, the problem of estimating the number k of false hypothesis leads to the problem of estimating the parameter p from the distribution of $H_p(x)$, as a contamination of $F(x)$ and $G(x)$.

$P^*(X)$ AND ITS PROPERTIES

Lemma 1 . *If* $(A1)$ *holds, then*

$$\mathbb{E}[p^*(x)] = p\left[1 - \frac{1-G(x)}{1-F(x)}\right], \tag{5}$$

where $p = k/n$.

Corollary 1 *For* $x \in (1-\delta, 1]$ $p^*(x)$ *is an unbiased estimator of* p.

Corollary 2 *If in addition to* $(A1)$

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}, \tag{6}$$

then the expected value of $p^*(x)$ *is a monotonic nondecreasing on the interval* $[0,1]$ *function. Moreover, since*

$$0 \leq 1 - \frac{1-G(x)}{1-F(x)} \leq 1,$$

then $0 \leq \mathbb{E}[p^*(x)] \leq p \ \forall x \in [0,1]$.

Theorem 1 *If random vectors* $\mathbb{X}$ *and* $\mathbb{Y}$ *are independent, then standard deviation of the estimator* $p^*(x)$ *has the form*

$$\sigma^2_{p^*(x)} = \frac{(1-p)F(x)}{n(1-F(x))} + \frac{pG(x)(1-G(x))}{n(1-F(x))^2}, \tag{7}$$

Corollary 3 *If condition (A2)holds then*

$$\sigma^2_{p^*(x)} = \frac{(1-p)F(x)}{n(1-F(x))}, \quad \textit{if} \quad 1-\delta < x \leq 1 \tag{8}$$

and

$$\sigma^2_{p^*(x)} = \frac{(1-p)F(x)}{n(1-F(x))} + \frac{pG(x)(1-G(x))}{n(1-F(x))^2}, \quad \textit{if} \quad 0 \leq x \leq 1-\delta. \tag{9}$$

Theorem 2 *Let conditions (A1) and (A2) satisfied. Then* $\sigma^2_{p^*(x)}$ *is a monotonic nondecreasing function of* x $\forall x \in [0,1]$.

SIMULATION STUDY

We simulated data from different distributions on the interval $[0,1]$. Almost in all cases both expected value and the standard deviation of $p^*(x)$ were increasing with x (Corollary 1 and Theorem 2). The case of uniform distribution is the most suitable to our theoretical explanation. With other distributions we observe some deviation from the uniform case but in general behavior of $p^*(x)$ is very close to the uniform case: it expected value increases with x in the interval $[0, 1-\delta]$ and "stabilizes" once x reaches the right border of the support. Standard deviation is always increasing and started from the right border of the support of the alternative distribution it increases more rapidly. If in the uniform case expected value of $p^*(x)$ remains constant above the support of distribution function $G(x)$, in other cases its behaves as a function of bounded variation. So, the right border of the support of $G(x)$ can serve as a lower bound for the estimator p. Since we want to choose our estimator as close as possible to 1, we can choose p greater than $1-\delta$ with a minimal standard deviation.

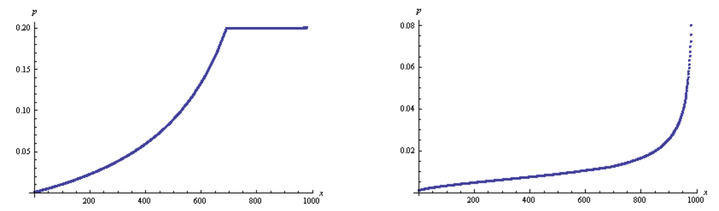

Figure 1: $\mathbb{E}[p^*(x)]$ and $\sigma^2_{p^*(x)}$ for $\mathbb{X} \sim \mathbb{U}(0,1)$ and $\mathbb{Y} \sim \mathbb{U}(0, 0.7)$.

In Figure 1 distribution under the null hypothesis is $\mathbb{U}(0,1)$ and under the alternative $\mathbb{U}(0, 0.7)$. Here we generated n-dimensional random vector $\mathbb{Z}$ with $n = 7500$, 10000 and 12500, and corresponding $k = 1500$, 2000 and 2500, such that in all cases $p = 0.2$. The number of simulations is $M = 10000$ ($n = 7500$ and $n = 12500$ is not shown here).

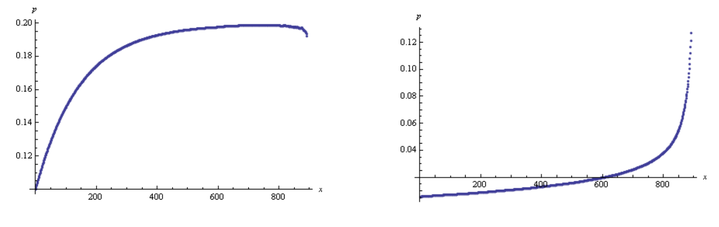

Figure 2: $\mathbb{E}[p^*(x)]$ and the $\sigma^2_{p^*(x)}$ for $\mathbb{X} \sim \mathbb{U}(0,1)$ and $\mathbb{Y} \sim 1 - e^{-\lambda x}$.

In Figure 2 the null distribution is again $\mathbb{U}(0,1)$ and under the alternative we took $G(x) = 1 - e^{-\lambda x}$ with $\lambda = 8$ and $\mathrm{supp} G(x) = [0, 0.7]$. As before we generated n-dimensional random vector $\mathbb{Z}$ with $n = 7500$, 10000 and 12500, and corresponding $k = 1500$, 2000 and 2500, such that in all cases $p = 0.2$. The number of simulations is $M = 15000$ ($n = 7500$ and $n = 12500$ is not shown here).

Acknowledgement. Author would like to express his thanks to Prof. Lev Klebanov for his generous support, valuable comments and help.

References

[1] Carvajal-Rodriguez, A., Una-Alvarez, J., Rolan-Alvarez, E., A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests, *BMC Bioinformatics*. Available from http://www.biomedcentral.com/content/pdf/1471

[2] Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. R. Statist. Soc.*, **Vol. 57**, 1995.

[3] Benjamini, Y., Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics*, **Vol. 25**,2000.

[4] Farcomeni, A. Multiple testing procedures under dependence, with applications. Ph.D. thesis, Univ Roma "La Sapienza", 2004.

[5] Hochberg, Y. and Tamhane, A. Multiple Comparison Procedures. New York, Wiley, 1987.

[6] Holm, S. A simple sequentially rejective multiple procedure, *Scand. J. Statist.*, **Vol, 6**, 1979.

[7] Klebanov, L. B. Yakovlev, A. Diverse correlation structures in gene expression data and their utility in improving statistical inference, *Statistics and Probability Letters*, **Vol. 31**, 2000.

[8] Klebanov, L. B. Yakovlev, A. A nitty-gritty Aspects of correlation and network inference from gene expression data, *Biology Direct*, available at: http://www.biology-direct.com/content/3/1/35.

[9] Lehmann, E. L., Romano, J. P. Generalization of he Familywise Error Rate, *Annals of Statistics*, **Vol. 34**, 2006.

[10] Meinhausen, N., Rice, J. P. Estimating the proportion of false null hypotheses among alarge number of independently tested hypotheses, *Annals of Statistics*, **Vol. 33**, 2006.

[11] Meinhausen, N., Bu¨hlmann, P. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures, *Biometrika*, **Vol. 92**, 2005.

[12] Qiu, X., Brooks, A. I., Klebanov, L. B. and Yakovlev, A., The effect of normalization on the correlation structure of microarray data *BMC Bioinformatics*, **Vol. 6**, 2005.

[13] Storey, J. D., A direct approach to false discovery Rate, *Journal of Royal Statistical Society,*, **Vol. 64**, 2002.

[14] Wu, W. B., Nonlinear system theory: Another look at dependence, *Proc. Natl. Acad. Sci. USA*, **Vol. 102**, 2005.

[15] Wu, W. B., On false discovery control under dependence, *The Annals of Statistics*, **Vol. 36**,2008.

[16] Westfall, P. H., Young, S. S. Resampling-based multiple testing: Examples and Methods for p-value Adjustment, Wiley, New York, 1993.