A Simulation Based Evaluation of the Bootstrap Bias Corrected Percentile Interval Estimators of the Local False Discovery Rates

Tasneem Zaihra*

Department of Mathematics, SUNY-Brockport

tzaihra@brockport.edu

Abstract

Large scale data, such as the one collected in microarray, proteomics, MRI imaging, and massive social science surveys etc., often requires simultaneous consideration of hundreds or thousands of hypothesis tests, which leads to inflated type I error rate. A popular way to account for it is to use local false discovery rates (LFDR), which is the probability that a gene is truly not differentially expressed given the observed test statistic. The purpose of this report is to evaluate the Bootstrap Bias Corrected Percentile (BBCP) method proposed by Shao and Tu (1995) for estimating the lower bound for the LFDR. The method didn't perform as expected. The overall coverage probability for null genes as well as non null genes was far from nominal coverage level of 50%.

KEY WORDS: Bootstrap Bias Corrected Percentile, Interval Estimator, False Discovery Rates, Local False Discovery Rates, Type I Error, Microarray Data

1 Introduction

Nowadays, we often come across large scale data, which are produced by experiments in microarray analysis, spectroscopy, proteomics, etc. Thus, large scale simultaneous hypothesis testing problems with hundreds of thousands of cases considered together have become very common. This leads to the problem of multiple testing, which leads to inflation in the overall experimentwise Type I error rate. One simple method for adjustment of Type I error rate in

multiple testing is Bonferroni's correction; many other methods are described by Westfall and Young (1993). This control of the experimentwise error rate is conservative and a popular alternative to it is the false discovery rate (FDR), first proposed by Benjamini and Hochberg (1995), which relies on null hypothesis tail areas. FDR is equal to the expected proportion of rejected null hypothesis that are true if the probability of rejection is greater than zero, that is, if null hypothesis is almost never rejected. In genomic, FDR is expected proportion of genes falsely called differentially expressed.

The literature on false discovery rate procedures can be divided into two areas Ghosh (2009). The ones that deal with methods which control the FDR, for details of such methods refer Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001) and Sarkar (2002) and the others are dedicated to point and interval estimation of the of the false discovery rate, for further details refer Efron et al. (2001), Storey (2002) and Genovese and Wasserman (2002).

Although, in microarray data analysis problems, the focus is on determining which null hypotheses are false, but in many cases, interest might focus on constructing estimators and associated confidence intervals, Ghosh (2009) for FDR.

A variant of false discovery rates, based on tail areas, is local false discovery rates (LFDR) introduced by Efron et al. (2001). Analogous to FDR, LFDR is also a measure of uncertainty refereing to single genes. It is defined as the probability that gene is truly not differentially expressed given the observed test statistic or p-value. In this report we evaluate one such, interval estimator for LFDR proposed by Shao and Tu Shao and Tu (1995).

Interval estimators are usually preferred over the point estimators because we can associate some guarantee of capturing parameter of interest with our estimator in the form of confidence coefficient. Good confidence interval has two desirable properties. Firstly it is narrow and secondly it has large confidence co-efficient. The main focus of this study is to evalaute the performance of the interval estimates of LFDR in terms of their coverage probability.

In Section 2 we provide a brief overview of the Bootstrap Bias Corrected Percentile(BBCP) estimators for the LFDR proposed by Shao and Tu (1995) based on Efron and Stein (1981) and Efron (1982). In Section 3 we discuss the simulation studies to evaluate them. In Section 4 we report the results of the pilot study. A Discussion along with some future directions follows in Sections 5 and 6 respectively.

2 Methods

2.1 Large Scale Simultaneous Hypothesis testing

Traditional approach to multiple inference control familywise error rate, resulting in excessive false negative rates. One such approach for example, is Bonferoni bound, it changes the rejection level for each test from α to α/k, where k is the number of hypotheses being tested, in order to prevent false positive rates. Several other multiple comparison procedures are used to control false positive rates. A more balanced approach involves controlling false discovery rates. It is used to control the expected proportion of incorrectly rejected null hypotheses. It has lower false negative rates than Bonferoni correction and other traditional methods of controlling family wise error rate. A popular way to account for it is to use local false discovery rates, which represents the posterior probability that the null hypothesis is true, in genomic, it represents the probability that the gene is truly not differentially expressed given the observed test statistic. For further details on FDR and LFDR refer to Benjamini and Hochberg (1995) and Efron and Stein (1981)

2.2 The Bootstrap Bias Corrected Percentile Interval Estimator

Suppose that the data $X_1, ..., X_n$ are i.i.d. from a distribution, with cumulative distribution function (cdf) given by $F(x)$. Suppose $x = (x_1, x_2, \ldots, x_n)$ is an observed random sample

from a distribution with cdf $F(x)$. If X^* is selected at random from x, then

$$P(X^* = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

If we resample from the observed random sample, x, with replacement, it generates a random sample $X_1^*, ..., X_n^*$. The random variable X_i^* are iid uniformly distributed on the set $x = (x_1, x_2, \ldots, x_n)$, Rizzo (2008). The estimator of the cdf $F(x)$, namely, the empirical distribution function (ecdf) is denoted by $F_n(x)$, is the cdf of X^*, while the ecdf of the bootstrap replicates is denoted by $F_n^*(x)$. Therefore, in bootstrap there are two approximations,as represented by Rizzo (2008), in the below diagram:

$$F(x) \to X \to F_n(x)$$

$$F_n(x) \to X^* \to F_n^*(x)$$

Thus in non parametric bootstrap the empirical distribution is given by

$$F_n(x) \;\; = \;\; \frac{1}{n}\Sigma_{i=1}^{n} I\{X_i \leq x\}$$

Now, suppose we denote local false discovery rate by θ. Let $\hat{\theta}_n$ be an estimator of θ and $\hat{\theta}_n^*$ be the bootstrap analog based on the bootstrap sample $(X_1^*, ..., X_n^*)$. A bootstrap percentile interval uses the empirical distribution of bootstrap replicates $(F_n^*(x))$ as the reference distribution, that is, the quantiles of the empirical distribution are estimators of the quantiles of the sampling distribution of $\hat{\theta}_n$. Let the bootstrap distribution of $\hat{\theta}_n^*$ be defined by

$$F_n^*(x) \;\; = \;\; P_{F_n}\{\hat{\theta}_n^* \leq x\}$$

, which is approximated by

$$\hat{F}_n^*(x) = Number \;\; of \;\; \{\hat{\theta}^* \leq x\}/B,$$

where B is the number of bootstrap replicates. In general, the percentile method uses a $100(1-\alpha)\%$ confidence interval with equal lower and upper tail errors $\alpha/2$, which is given by $[\hat{F}_n^{*-1}(\alpha/2),\ \ \widehat{F}_n^{*-1}((1-\alpha)/2)]$. Thus, if $\theta^{\hat{(1)}},\ldots,\theta^{\hat{(B)}}$ are the bootstrap replicates of the statistic $\hat{\theta}$, from the ecdf of the replicates compute the $\alpha/2$ and $1-\alpha/2$ quantiles. Similarly, we can calculate lower or upper bounds. Several adjustments for percentile methods have been proposed to improve their theoretical properties and enhance their performance. Usually, for a $100(1-\alpha)\%$ confidence interval the quantiles are adjusted by two factors: a correction for bias and a correction for skewness. The bias correction factor measures the median bias of the bootstrap replicates $\hat{\theta}^*$ for $\hat{\theta}$, Shao and Tu (1995). The bias corrected intervals are transformation respecting but only first order accurate. Shao and Tu (1995) proposed bias corrected percentile confidence interval for the LFDR, which is based on the bootstrap percentile method by Efron (1981). It gives the lower confidence bound for θ as $\theta_{BP} = \hat{F}_n^{*-1}(\alpha)$.

Furthermore, they proposed an exact lower confidence bound for θ, given by

$$\theta_{exact} \quad = \quad \phi_n^{-1}(\hat{\phi_n} + z_\alpha),$$

where $Z_\alpha = \Psi^{-1}(\alpha)$ under the assumption that there exists an increasing transformation $\phi_n(x)$ such that

$$P\{\hat{\phi_n} - \phi_n(\theta) \leq x\} \quad = \quad \Psi(x) \tag{1}$$

holds for all possible F, where $\hat{\phi_n} = \phi_n(\hat{\theta_n})$ and Ψ is a continuous increasing and symmetric distribution, i.e., $\Psi(x) = 1 - \Psi(-x)$. If we assume $\Psi = \Phi$, the standard normal distribution, the function ϕ_n acts as the normalizing and variance stabilizing transformation, Shao and Tu (1995).

However, the above proposed lower confidence bound is exact if the assumption in equation (1) hold exactly. If the assumption holds approximately for large n then, the bound is

asymptotically valid and its performance depends on how good the approximation is. Usually the transformation ϕ_n is non linear and the bias $\hat{\phi_n} - \phi_n(\theta)$ does not vanish quickly as $n \to \infty$, since the assumption in (1) holds only when n is large. Therefore, the proposed bootstrap percentile confidence intervals are simple but need very large n for accuracy. This led to the proposal of bias corrected percentile bootstrap confidence intervals based on the following, more general assumption, originally proposed by Efron (1982)

$$P\{\hat{\phi_n} - \phi_n(\theta) + z_0 \leq x\} \quad = \quad \Psi(x) \tag{2}$$

where z_0 is a constant that may depend on F and n, ϕ_n is still an increasing transformation and Ψ is still assumed continuous, strictly increasing and symmetric. If ϕ_n, z_0 and Ψ are known an exact lower confidence bound as obtained by Shao and Tu (1995) is given below:

$$\theta_{exact} \quad = \quad \phi_n^{-1}(\hat{\phi_n} + z_\alpha + z_0)$$

Applying assumption (2) to $F = \hat{F}$ we obtain that

$$F_n^*(x) \quad = \quad P_{F_n}\{\hat{\phi_n^*} - \hat{\phi_n} + z_0 \leq z_0\} = \Psi(z_0)$$

where, $F_n^*(x) = P_{F_n}\{\hat{\theta_n^*} \leq x\}$,which implies

$$z_0 \quad = \quad \Psi^{-1}(F_n^*(\hat{\theta_n})) \tag{3}$$

Assuming, Ψ is known and using equation (3), the BBCP lower confidence bound for θ as obtained by Shao and Tu (1995) is:

$$\theta_{BC} \quad = \quad F_n^{*-1}\left(\Psi\big(z_\alpha + 2\Psi^{-1}(F_n^*(\hat{\theta_n}))\big)\right)$$

By taking bias into account the bootstrap does improve however, there are still many cases where assumptions can not be fulfilled and requires further assumes. In this study, the main focus is on to asses the performance of a confidence sets for the LFDR built using the above method.

2.3 Evaluation Criteria

A good method for calculating confidence interval will produce an appropriate coverage probability and avoid all abberations. Let θ denote the local false discovery rate, $\hat{\theta}$ it's corresponding point estimate and $[L, U]$ the corresponding lower and upper confidence limits. Then, the coverage probability (CP) is defined as $Pr[L \leq \theta \leq U]$. The exact criteria requires $CP \geq 1 - \alpha$ but by the smallest attainable margin that is $L \leq \theta \leq U$ should occur with probability $1 - \alpha$ and $L > \theta$ and $U < \theta$ each with probability $\alpha/2$. The main focus of this study is to check whether the coverage probability of the confidence sets converges to nominal level or not in various situations described in the next section.

3 Simulation Studies

We begin with the two-groups model, in which each of the N cases is either null or non-null, with prior probability p_0 or p_1.

$$p_0 = Pr\{Null\}$$

with $f_0(z)$ as the null density and

$$p_1 = Pr\{NonNull\}$$

and $f_1(z)$ is the non-null density. We generate 200 random gene expression dataset with $N = 10,000$ genes, each with $n = 5$ replications. For our simulation study, we assume that the N cases are divided into two classes, null and non null, occurring with prior probabilities $p_0 = Pr\{Null\}$ with standard normal density $N(0, 1)$ or $p_1 = Pr\{NonNull\}$ with normal density $N(\mu_{alt}, 1)$,where $1 \leq \mu_{alt} \leq 5$. Thus, we generate the data using mixture of two normal distributions with mixing proportions p_0 and p_1. Since, practical applications of large scale testing usually assume p_0 large as the goal of the studies is to identify a relatively small

set of interesting non-null cases, therefore we consider values of $p_0 \geq 0.9$ for data generation purpose. For each of these 200 random gene expression data we calculate non parametric bias corrected percentile bootstrap confidence intervals. The number of bootstrap ($n.boot$) samples that we use is 500. Also, to avoid the possibility of all elements of the sample to be same while bootstrapping, as it involves sampling with replacement we use R-function jitter to break ties for the situation. Since, the estimated time with 8 multi cores for the following configuration: Number of genes ($n.genes$)= 10000, Sample size ($n.samp$)= 5, Number of bootstrap Samples ($n.boot$) =500, alternate mean (μ_{alt})=2 , Number of simulations (n) = 32 was 37145.45 seconds. Therefore, we estimated that 1000 simulations would take around 13 days. So, instead we did a pilot study with 200 simulation using FUNDY cluster of the Atlantic Computational Excellence Network (ACEnet), which is a consortium of Atlantic Canadian Universities providing researchers with high performance computing (HPC) resources.

4 Results

In the table below we provide estimated coverage rates for one sided non-parametric percentile bootstrap confidence intervals based on the method proposed by Shao and Tu (1995) at 50% nominal coverage level.

Table 1: Estimated coverage rates for one sided non-parametric percentile bootstrap confidence intervals based on Shao and Tu's method

$n.genes$	$n.samp$	n	$n.boot$	μ_{alt}	$CP_{NullGenes}$	$CP_{NonNullGenes}$	$Avg.Covg$
10000	5	200	500	2	0.1126661	0.095900	0.1109895
1000	5	200	500	3	0.04312167	0.027095	0.041519
1000	5	200	500	4	0.02787611	0.007995	0.025888
1000	5	200	500	5	0.04312167	0.027095	0.041519

We also present some plots of the bootstrap distributions indicating their deviation from

the true values e.g., a plot of each 50% upper limit, and original bootstrap estimates versus the corresponding true values. Basically, we plotted two plots for the simulation, one plot for the null cases and another plot for the alternative cases for each iteration. The x-axis gives the true value of the LFDR. The y-axis gives the following estimated values of the original estimate obtained before bootstrapping (circles) and the upper limit of the confidence interval (triangles). For brevity, we report a selection of four plots each for null and non null gene case in the report, most of which show non-coverage (with the upper limit lower than the true value).

5 Discussion

The main objective of the pilot study was to evaluate bias corrected percentile bootstrap methods used for confidence interval estimation of local false discovery rate. Provided the coverage probability of the confidence interval converged to the nominal level, we planned to explore the viability of the method for interval estimation of LFDR under different configurations, for instance, alternate mean and proportion of interesting non-null cases. Also, we planned to apply the methods to some real life data sets. However, as it can be noted from Table 1, the method didn't perform as expected. The overall coverage probability for null genes as well as non null genes is far from nominal coverage level of 50%. Furthermore, most of the plots (only few of which are reported in results section for brevity, others can be obtained from the authors) show non-coverage with the upper limit lower than the true value.

5.1 Future Directions

In order to assess future directions, we further tried a simpler method due to Storey (2002). It works better for interval estimates of false discovery rate and we plan to further study it for interval estimation of local false discovery rate. To calculate interval estimate of

Figure 1: Plots for Non-Null Genes Cases

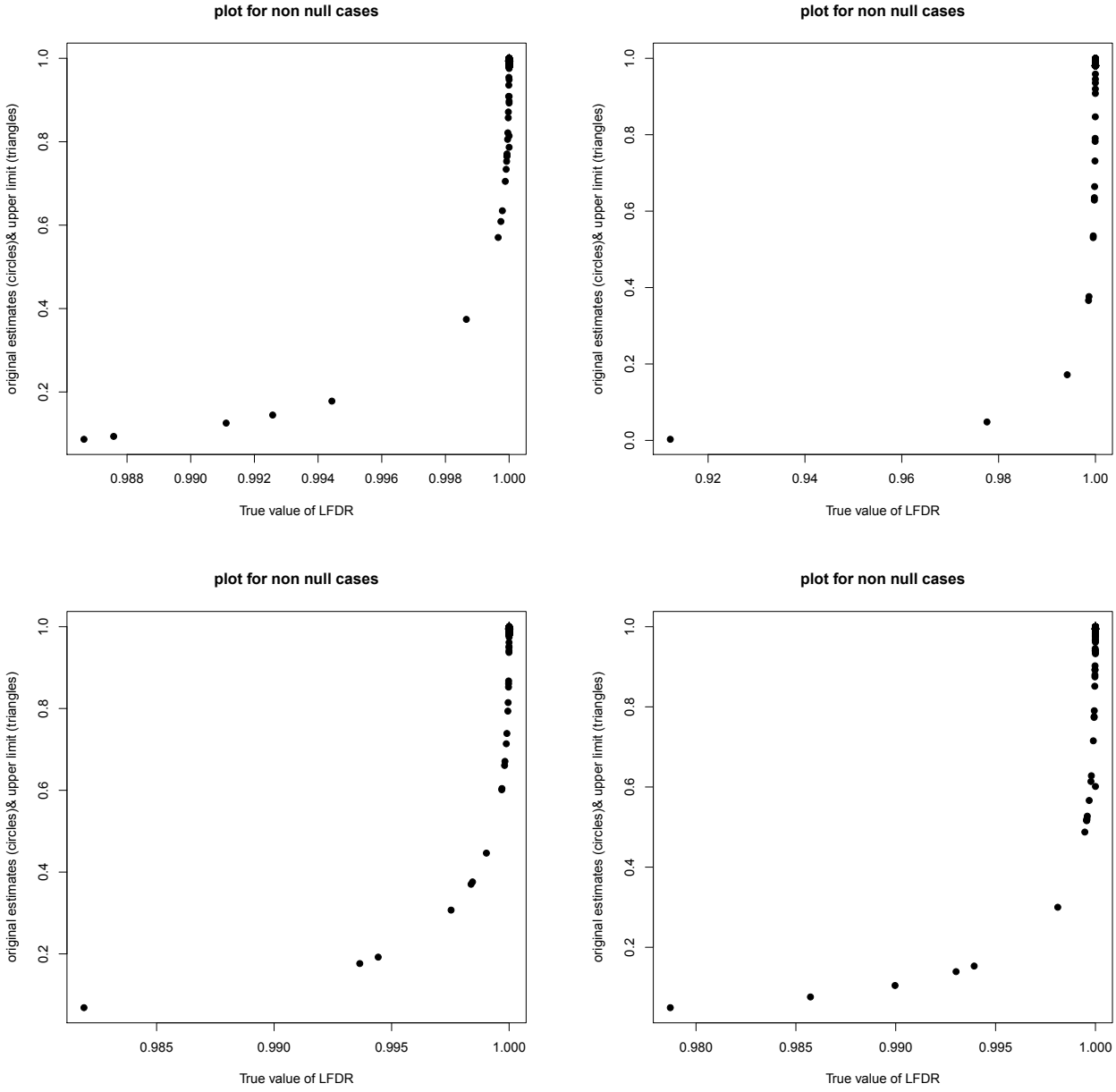
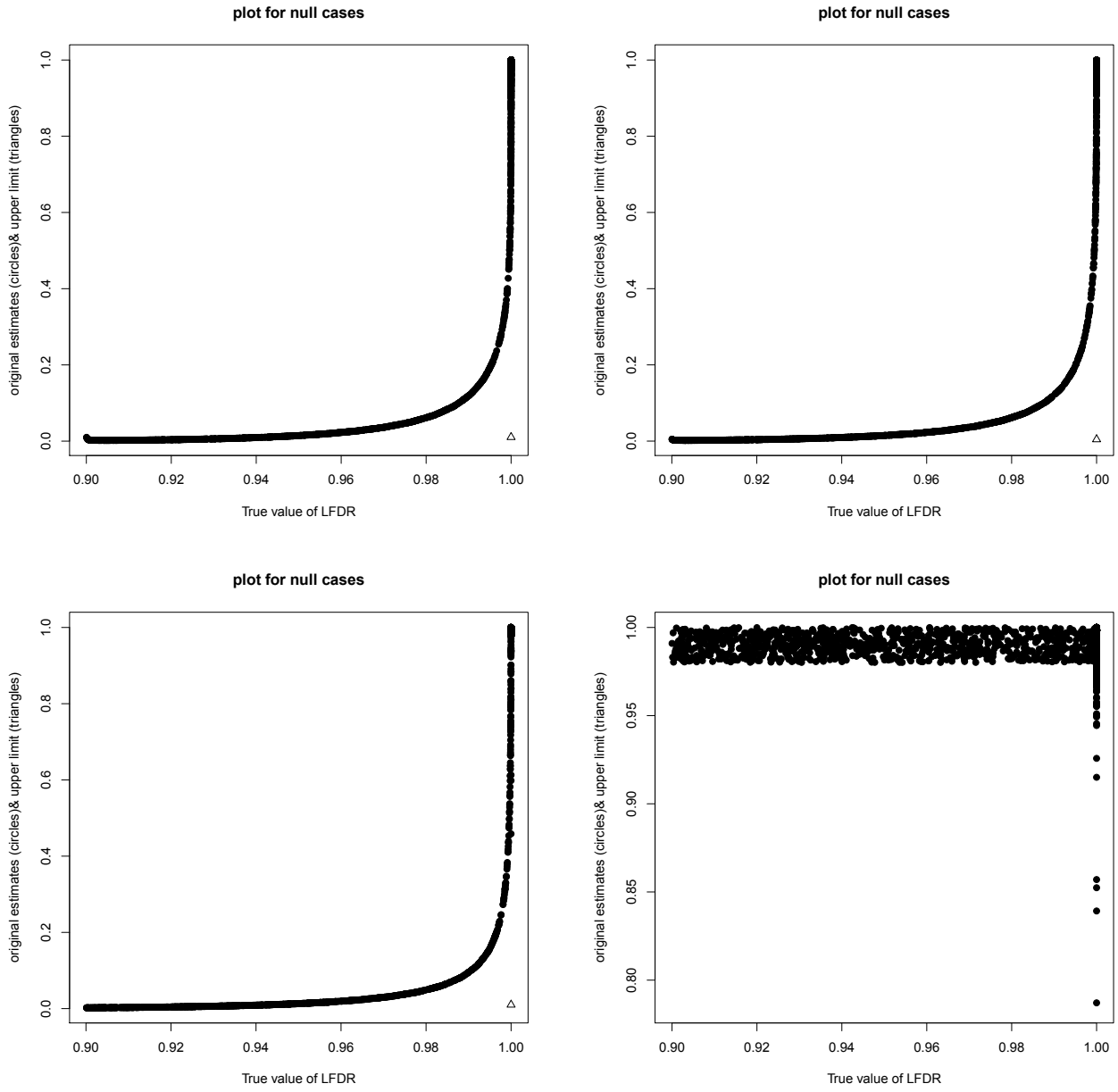


Figure 2: Plots for Null Genes Cases



false discovery rate using the method based on Storey (2002) we first took a nonparametric bootstrap sample of the p-values, i.e., resample them with replacement. Then we computed the estimated false discovery rate as $0.01 * m$ divided by the number of p-values less than 0.01, where m is the number of p-values. The above two steps were repeated 1000 times. We then took the $5th$ and $95th$ quantiles of the estimated false discovery rates to get a 95% confidence interval. These steps were repeated for each simulated vector of p-values, thus we simulated 1000 vectors of p-values and then we calculated the coverage probability i.e., out of the 1000 confidence intervals, the proportion that contains the true value of the false discovery rate. Following is the table for the coverage probabilities:

Table 2: Estimated Coverage Rates for non-parametric percentile bootstrap confidence interval for false discovery rate based on John Storey's Method.

$n.genes$	$n.samp$	$n.boot$	$n[n_{good}]$	π_0	μ_{alt}	$Covg.Prob$
10000	5	1000	1000[893]	0.9	5	0.822
10000	5	1000	1000[878]	0.9	4	0.828
10000	5	1000	1000[302]	0.9	2	0.811

As we can note from Table 2, above, this methods perform much better in terms of coverage probability for false discovery rates. For future study we plan to extend the above method for interval estimation of local false discovery rates and evaluate it in terms of coverage probability. Note in some iterations upper confidence limit becomes infinite when all the p-values sampled while obtaining bootstrap sample are > 0.01, which leads to infinite FDR, n_{good} in Table 2 stands for the number of good iterations, which are the iterations in which upper confidence limit doesn't become Infinite.

6 Acknowledgements

This work was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University

of Ottawa. I would like to thank Prof David Bickel, at the Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa for his guidance throughout the pilot study and for the permission to use the R functions developed by him. I would also like to thank Corey Yanofsky, for the useful discussions related to the simulation study, while I was getting trained at Dr.Bickel's Lab. Furthermore, the computational resources were provided by ACENET, the regional advanced research computing consortium for post-secondary institutions in Atlantic Canada. ACENET is funded by the Canada Foundation for Innovation (CFI), the Atlantic Canada Opportunities Agency (ACOA), and the provinces of Newfoundland and Labrador, Nova Scotia, and New Brunswick.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139–158.

Efron, B. (1982). *The Jackknife, The Bootstrap and Other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

Efron, B. and Stein, C. (1981). The jackknife estimator of variance. *Annals of Statistics*, 9:586–596.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of American Statictical Association*, 96:1151–1160.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of Royal Statistical Society Series B*, 64:499–517.

Ghosh, D. (2009). Empirical bayes methods for estimations and confidence intervals in high dimnesional problems. *Statistica Sinica*, 19:125–143.

Rizzo, M. (2008). *Statistical Computing with R.* Chapman & Hall, CRC, Boca Raton, FL.

Sarkar, S. (2002). Some results on false discovery rates in multiple testing procedures. *Annals of Statistics*, 30:239–257.

Shao, J. and Tu, D. (1995). *The Jacknife and Bootstrap.* Springer.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of Royal Statistical Society Series B*, 64:479–498.

Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment.* John Wiley, New York.