

# A Quantile-based Test for Detecting Differential Expression in Microarray Data

## *Un Test Basato sui Quantili per l'Individuazione di Espressione Differenziale in Dati di Microarray*

Lorenzo Maragoni and Monica Chiogna

**Abstract** In the last decade the development of statistical methods for microarray data analysis has been object of active research. A crucial problem in this area is the search for genes that show a differential expression between two or more groups like, for example, healthy and sick tissues. A huge number of statistical tools has been proposed for this task, most of which focus on testing equality of means. The aim of the present work is to introduce a simple quantile-based statistic that can test differences at different levels of the distribution, and has the desirable property of invariance with respect to monotone data transformations.

**Abstract** *Nello scorso decennio lo sviluppo di metodi statistici per l'analisi di dati di microarray è stato oggetto di crescente interesse. Un problema cruciale in questo campo è l'identificazione di geni che mostrano una differente espressione tra due o più gruppi, come ad esempio tra tessuti sani e malati. Un grande numero di strumenti statistici è stato proposto per questo compito, la maggior parte dei quali ha l'obiettivo di testare un'ipotesi di uguaglianza tra medie. Lo scopo del presente lavoro è di introdurre una statistica semplice basata sui quantili, in grado di testare differenze a diversi livelli della distribuzione, e che gode della proprietà desiderabile di invariata rispetto a trasformazioni monotone dei dati.*

**Key words:** microarray, differential expression, bootstrap

---

Lorenzo Maragoni  
Department of Statistical Sciences University of Padua, 241 Via Battisti, Padova 35121 (Italy)  
e-mail: maragoni@stat.unipd.it

Monica Chiogna  
Department of Statistical Sciences University of Padua, 241 Via Battisti, Padova 35121 (Italy)  
e-mail: monica@stat.unipd.it

## 1 Differential Expression in Microarray Studies

In microarray experiments, the expression of large number of genes is measured over a few replicates. Situations where tens of thousands of genes are sampled over a few tens of biological or technical replicates are not infrequent. In this setting, the experimenter's goal is usually to identify a small subset of genes that show differential expression between two groups of replicates, such as between healthy and sick tissues, or different arms of a clinical trial. The selected genes will then be the object of further biological analysis, impractical or impossible to be done over the original whole set of genes. The choice of appropriate statistical tools in the field of gene expression analysis is of crucial interest, and attempts in this direction have been numerous and diverse. Since comparison of two groups is at the core of the problem, classical  $t$ -statistics and many subsequent modifications have been object of particular interest. Modifications have often aimed at artificially increasing some of the smallest variances, which might have happened by chance and lead to false positives in the search for differentially expressed genes. Among the most notable examples, the *SAM* procedure introduced a stabilization factor for experimental variability [1], while a few hierarchical models provide a broad Bayesian framework for inference [2] [3]. The research on the topic is still very active, and both modifications of the above methods and original solutions have been proposed. For a comparison of different alternatives we refer, for example, to [4] and [5]. In the following, we will build a simple approach based on quantile comparison to identify differentially expressed genes. To the best of our knowledge, not many attempts at using the tools of quantile regression, as introduced by [6], have been made in the area of microarray data analysis, being one of the few examples [7]. One of the key aspects of the approach is related to the invariance property of quantiles, which guarantees that testing hypotheses on different scale will yield the same results. This property is not shared by the mean estimator, but is instead a key aspect in a field where data transformation is often a routine operation. Moreover, quantile comparison allows to test differences on different aspects of the distribution of the data, not only at its center. Since the concept of "differential expression" has not a rigid definition, quantile inference might then be useful in opening new possibilities for microarray data analysis, a field where the relationship between statistical and biological aspects is complex and constantly in evolution. In Section 2 we will propose a novel statistic for testing differential expression and study its most prominent features. In Section 3 we will provide some simulation results, and we will provide a brief conclusion in Section 4.

## 2 A Simple Proposal

Let  $X_j^A \sim f_{X_j^A}$  and  $X_j^B \sim f_{X_j^B}$  be two absolutely continuous random variables, modeling expression data for gene  $j = 1, \dots, p$  in samples  $A$  and  $B$ , respectively. Assume

$Y_j^A = g(X_j^A)$  and  $Y_j^B = g(X_j^B)$ , where  $g(\cdot)$  is a strictly monotone function and let  $Y_j^A \sim f_{Y_j^A}$  and  $Y_j^B \sim f_{Y_j^B}$ . Let also  $\xi_U(\tau)$  denote the quantile  $\tau \in (0, 1)$  of the random variable  $U$ , and  $n = n_A + n_B$  the total sample size. We wish to test the hypothesis:

$$H_j^0 : \xi_{X_j^A}(0.5) = \xi_{X_j^B}(0.5), \quad j = 1, \dots, p,$$

or more in general the hypothesis:

$$H_j^0(\tau) : \xi_{X_j^A}(\tau) = \xi_{X_j^B}(\tau), \quad j = 1, \dots, p.$$

Before proceeding, we highlight the consistency with respect to transformations, a fundamental property of quantiles:

$$\xi_{Y_j^A}(\tau) = \xi_{Y_j^B}(\tau) \iff \xi_{g(Y_j^A)}(\tau) = \xi_{g(Y_j^B)}(\tau),$$

for any strictly monotone function  $g(\cdot)$ . This implies that testing on the original scale is equivalent to testing on the transformed scale, while this would in general not be true while testing equality of means, due to Jensen's inequality. Now let  $X_{j1}^A, \dots, X_{jn_A}^A$  and  $X_{j1}^B, \dots, X_{jn_B}^B$  be simple random samples of size  $n_A$  and  $n_B$  respectively, independent from each other, generated from  $f_{X_j^A}$  and  $f_{X_j^B}$ . The proposed statistic  $QT(\tau)$  is then based on a raw difference between quantiles, namely

$$\hat{\xi}_{X_j^A}(\tau) - \hat{\xi}_{X_j^B}(\tau) \tag{1}$$

where  $\hat{\xi}_{X_j^K}(\tau)$  is the sample quantile of level  $\tau$  for sample  $K \in \{A, B\}$ . Since the asymptotic distribution of the estimator  $\hat{\xi}_{X_j^K}(\tau)$  is [6]:

$$\sqrt{n_K}(\hat{\xi}_{X_j^K}(\tau) - \xi_{X_j^K}(\tau)) \sim N(0, \omega_{X_j^K}^2), \quad \omega_{X_j^K}^2 = \tau(1 - \tau) / f_{X_j^K}^2(\xi_{X_j^K}(\tau)),$$

for  $K \in \{A, B\}$ , then an estimate of the density function would be necessary to properly Studentize the quantity (1) as in [8]. Several choices would be available for this task, such as the usual kernel estimator [9] and the bootstrap estimator [10], but they all can be computationally intensive. Therefore, in analogy with mean-based statistics, we will compare the quantity in (1) to a suitable measure of variability, following a structure similar to the one of a Student's  $t$ . In order to compute this measure, we make the assumption of equal populations implicit in the classic Student's  $t$  statistic. Under this hypothesis, and assuming  $n_A = n_B$ , we choose as a measure of variability the quantity

$$\hat{\xi}_{X_j^P}(1 - \varepsilon) - \hat{\xi}_{X_j^P}(\varepsilon) \tag{2}$$

where  $\varepsilon \in (0, 1)$  and  $P$  denotes the pooled sample. The choice of pooling the samples seems reasonable (under  $n_A = n_B$ ) since the two populations are assumed

to have the same distribution and therefore the same variability. Therefore we define our pseudo-Studentized statistic as the ratio of quantities (1) and (2):

$$QT_j(\tau) = \frac{\hat{\xi}_{X_j^A}(\tau) - \hat{\xi}_{X_j^B}(\tau)}{\hat{\xi}_{X_j^P}(1 - \varepsilon) - \hat{\xi}_{X_j^P}(\varepsilon)}. \quad (3)$$

Based on this statistic, we can provide a  $p$ -value via resampling (for example bootstrap) or via its asymptotic distribution, which we will derive in the next subsection. After possible corrections, for example due to multiple testing if  $p$  is large, genes whose  $p$ -value exceeds a prefixed threshold can then be candidates for differential expression.

## 2.1 Asymptotic distribution of $QT(\tau)$

Under mild assumptions on  $f_{X_j^A}$  and  $f_{X_j^B}$ , it is possible to prove asymptotic normality of  $QT(\tau)$ . First, we will show that, because of independence of the two samples, the quantity (1) is asymptotically normal. In fact, if the sample sizes converge at a comparable rate, i.e. if  $n_A/n_B \rightarrow c \neq 0$ , then:

$$(\hat{\xi}_{X_j^A}(\tau) - \hat{\xi}_{X_j^B}(\tau)) - (\xi_{X_j^A}(\tau) - \xi_{X_j^B}(\tau)) \sim N(0, \omega_{X_j^A}^2/n_A + \omega_{X_j^B}^2/n_B).$$

Second, we will show that the quantity (2) converges in probability, applying the Cramér and Wold device, which reduces the convergence of multivariate distribution functions to the convergence of univariate distribution functions, to the asymptotic joint distribution of quantile estimators. The resulting distribution is:

$$\sqrt{n} \left( (\hat{\xi}_{X_j^P}(1 - \varepsilon) - \hat{\xi}_{X_j^P}(\varepsilon)) - (\xi_{X_j^P}(1 - \varepsilon) - \xi_{X_j^P}(\varepsilon)) \right) \xrightarrow{d} N(0, \sigma_F^2),$$

where

$$\sigma_F^2 = \frac{\varepsilon(1 - \varepsilon)}{f_{X_j^P}(\xi_{X_j^P}(1 - \varepsilon))^2} + \frac{\varepsilon(1 - \varepsilon)}{f_{X_j^P}(\xi_{X_j^P}(\varepsilon))^2} - \frac{(1 - \varepsilon)^2}{f_{X_j^P}(\xi_{X_j^P}(\varepsilon))f_{X_j^P}(\xi_{X_j^P}(1 - \varepsilon))}$$

and  $f_{X_j^P}$  is the density function for the pooled sample. The result implies convergence in probability:

$$\hat{\xi}_{X_j^P}(1 - \varepsilon) - \hat{\xi}_{X_j^P}(\varepsilon) = \xi_{X_j^P}(1 - \varepsilon) - \xi_{X_j^P}(\varepsilon) + o_p(1).$$

Asymptotic normality of  $QT_j(\tau)$  now can be proven by Slutsky's Theorem. Provided that  $\xi_{X_j^P}(1 - \varepsilon) - \xi_{X_j^P}(\varepsilon) \neq 0$ , one has that, under  $H_j^0(\tau)$ :

$$QT_j(\tau) \sim N(0, \omega_j^2)$$

where

$$\omega_j^2 = (\omega_{X_j^A}^2/n_A + \omega_{X_j^B}^2/n_B)/(\xi_{X_j^P}(1-\varepsilon) - \xi_{X_j^P}(\varepsilon))^2.$$

Therefore, if  $n_A$  and  $n_B$  are sufficiently large, one can use the asymptotic distribution to compute  $p$ -values for testing  $H_j^0(\tau)$ . However, the above mentioned problem of density estimation would persist. In this case, since a proper Studentization is not available and therefore results of [8] do not apply, we can compute  $p$ -values via bootstrap resampling, although no guarantee exists that the unconditional type I error levels are matched. In the next section we will test the performance of  $QT(\tau)$  with a bootstrap  $p$ -value in two small simulation studies.

### 3 Simulation Studies

We have conducted several simulation studies to test the performances of  $QT(\tau)$ . We report results obtained in two simple settings: as a first study we simulate  $p = 1000$  genes from  $X_j^A \sim N(0, 1)$  and  $X_j^B \sim N(0, 1)$ , and as a second study we simulate  $p = 1000$  genes from  $X_j^A \sim \exp(1)$  and  $X_j^B \sim \exp(1)$ . We perform both studies for  $\bar{n} = n_A = n_B \in \{5, 10, 100\}$  and compare the estimated type I error obtained using the true density function and the one obtained via bootstrap with the nominal one, fixed to  $\alpha = 0.05$ . In the first example we use  $\tau = 0.9$ , and in the second one we use  $\tau = 0.5$ , while we let  $\varepsilon \in \{0.10, 0.25, 0.45\}$ . Results for the two studies are reported in Table 1 and Table 2, respectively.

**Table 1** Estimated type I error for standard Normal distributions ( $\tau = .9, \alpha = 0.05$ ).

| $\varepsilon$ |                                    | $\bar{n} = 5$ | $\bar{n} = 10$ | $\bar{n} = 100$ | $\bar{n} = 1000$ |
|---------------|------------------------------------|---------------|----------------|-----------------|------------------|
| 0.45          | $\hat{\alpha}(\tau)$               | 0.270         | 0.156          | 0.066           | 0.059            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.100         | 0.021          | 0.037           | 0.058            |
| 0.25          | $\hat{\alpha}(\tau)$               | 0.018         | 0.069          | 0.050           | 0.060            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.097         | 0.022          | 0.033           | 0.058            |
| 0.10          | $\hat{\alpha}(\tau)$               | 0.070         | 0.028          | 0.044           | 0.058            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.102         | 0.017          | 0.038           | 0.055            |

### 4 Conclusions

From these preliminary studies, the use of  $QT(\tau)$  seems promising. The obtained bootstrap type I errors do not look far from the expected ones for a sufficiently large  $\bar{n}$ , and further modifications could be implemented for refining the method and pos-

**Table 2** Estimated type I error for standard Exponential distributions ( $\tau = .5, \alpha = 0.05$ ).

| $\varepsilon$ |                                    | $\bar{n} = 5$ | $\bar{n} = 10$ | $\bar{n} = 100$ | $\bar{n} = 1000$ |
|---------------|------------------------------------|---------------|----------------|-----------------|------------------|
| 0.45          | $\hat{\alpha}(\tau)$               | 0.220         | 0.090          | 0.035           | 0.049            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.013         | 0.010          | 0.026           | 0.048            |
| 0.25          | $\hat{\alpha}(\tau)$               | 0.026         | 0.031          | 0.044           | 0.053            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.013         | 0.011          | 0.028           | 0.047            |
| 0.10          | $\hat{\alpha}(\tau)$               | 0.110         | 0.068          | 0.042           | 0.054            |
|               | $\hat{\alpha}_{\text{boot}}(\tau)$ | 0.015         | 0.013          | 0.031           | 0.050            |

sibly achieving a faster convergence. The adequate performances of the statistic, together with the ability to study differences far from the mean and with the invariance property of quantiles might propose  $QT(\tau)$  as a valuable integration or alternative to current popular methods for detecting differential expression in microarray data.

## References

- [1] V.G. Tusher, R. Tibshirani, and G. Chu. “Significance analysis of microarrays applied to the ionizing radiation response”. In: *Proceedings of the National Academy of Sciences* 98.9 (2001), p. 5116.
- [2] B. Efron et al. “Empirical Bayes analysis of a microarray experiment”. In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1151–1160.
- [3] G.K. Smyth et al. “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.” In: *Statistical applications in genetics and molecular biology* 3.1 (2004), p. 3.
- [4] C. Kooperberg et al. “Significance testing for small microarray experiments”. In: *Statistics in medicine* 24.15 (2005), pp. 2281–2298.
- [5] I.B. Jeffery, D.G. Higgins, and A.C. Culhane. “Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data”. In: *Bmc Bioinformatics* 7.1 (2006), p. 359.
- [6] Roger Koenker and Gilbert Bassett Jr. “Regression quantiles”. In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.
- [7] Liping Huang et al. “A novel application of quantile regression for identification of biomarkers exemplified by equine cartilage microarray data”. In: *BMC bioinformatics* 9.1 (2008), p. 300.
- [8] EunYi Chung, Joseph P Romano, et al. “Exact and asymptotically robust permutation tests”. In: *The Annals of Statistics* 41.2 (2013), pp. 484–507.
- [9] LP Devroye and TJ Wagner. “The strong uniform consistency of kernel density estimates”. In: *Multivariate Analysis* 5 (1980), pp. 59–77.
- [10] Bradley Efron. “Bootstrap methods: another look at the jackknife”. In: *The annals of Statistics* 7.1 (1979), pp. 1–26.