Journal of Data Science **7**(2009), 219-234

Double Sampling Designs to Reduce the Non-discovery Rate: Application to Microarray Data

Maela Kloareg and David Causeur
Laboratoire de Mathématiques Appliquées Agrocampus Ouest

Abstract: Simultaneous tests of a huge number of hypotheses is a core issue in high flow experimental methods such as microarray for transcriptomic data. In the central debate about the type I error rate, Benjamini and Hochberg (1995) have proposed a procedure that is shown to control the now popular False Discovery Rate (FDR) under assumption of independence between the test statistics. These results have been extended to a larger class of dependency by Benjamini and Yekutieli (2001) and improvements have emerged in recent years, among which step-up procedures have shown desirable properties.

The present paper focuses on the type II error rate. The proposed method improves the power by means of double-sampling test statistics integrating external information available both on the sample for which the outcomes are measured and also on additional items. The small sample distribution of the test statistics is provided and simulation studies are used to show the beneficial impact of introducing relevant covariates in the testing strategy. Finally, the present method is implemented in a situation where microarray data are used to select the genes that affect the degree of muscle destructuration in pigs. A phenotypic covariate is introduced in the analysis to improve the search for differentially expressed genes.

Key words: Auxiliary covariate, double-sampling, false discovery rate, multiple tests, non-discovery rate.

1. Introduction

Although multiple testing issues have been widely discussed in the statistical literature for a long time, novel approaches have emerged in recent years to face situations where the number of tests is especially huge. Simultaneous tests have for instance become one of the core issues of the analysis of gene expression data measured in microarray experiments. In this situation, the main goal is to identify the genes that show good evidence of being differentially expressed under two or more conditions (eg. treatments, genotypes or times in kinetic studies).

The first approaches, among which Bonferroni's strategy is probably the most famous, aim at controlling the probability of a single false discovery, also called Family-Wise Error Rate and denoted FWER. As shown in Dudoit, Shaffer and Boldrick (2003), the power of such procedures can be improved by means of approximations of the distribution of the test statistics based on permutation or bootstrap methods. In the recent discussions about an alternative type I error rate that would yield to less conservative decision rules, a major innovation has come from Benjamini and Hochberg (1995), who define the false discovery rate (FDR) as the proportion of true H_0 among the tests for which H_0 is rejected. Benjamini and Hochberg (1995) also provide a decision rule that is shown by Benjamini and Yekutieli (2001) to control the FDR under a large class of positive dependency between the test statistics. Statistical properties of the former methodology have been explored by many authors (see for instance Storey, Taylor and Siegmund (2004)) under quite wide distributional assumptions.

As mentioned above, either for the FWER or the FDR, attempts to improve the existing methods involve a better knowledge of the responses' dependency structure. Unfortunately, the high dimensionality of the data usually prohibits the modelling of the whole set of variables' joint distribution. As mentioned by Kendziorski *et al.* (2003), treating variables as independent tend to be less efficient than some Bayesian approaches which take advantage of the shared information between variables. Similarly, some authors (see for instance Lönnstedt and Speed (2002); Smyth (2004)) proposed moderated versions of the t-statistic where the variable-specific variance estimator that appears in the denominator is augmented by a constant that is derived from the data of all variables.

In many situations, relating the responses to auxiliary variables can also give insight into the correlation structure of sets of variables. For instance, in the case of transcriptomic data, phenotypic variables, often much easier to measure than microarray data, can help interpreting the correlation between gene expressions. As mentioned by von Heydebreck, Huber and Gentleman (2004), integrating biological relevant knowledge and gene expressions in the differential analysis is not usual, though usually handled by canonical analysis in multivariate exploratory data analyses.

The aim of our paper is to propose a testing method, based on double-sampling t-statistics that integrates external information to improve the power of the existing testing strategies. This external information is supposed to be available in the sample for which the responses are measured but also on additional items for which the responses are not measured. Improving inference by use of auxiliary variables in such a double-sampling framework is not novel in some areas of statistics, although multiple sampling strategies are usually dedicated to improvements of estimation procedures and more rarely to testing issues.

Origins of such designs are probably to be found in Cochran (1963) that shows how to reduce the variance of the estimation of a mean by use of an auxiliary variable. The same idea can be found, for instance in Conniffe (1985), transposed to the estimation of the parameters of a multivariate normal regression model. In this situation, most of the papers have dealt with the optimal allocation of the measurements of the outcome and of the auxiliary variable (see Causeur and Dhorne (1998), Causeur (2005)). Analogous ideas can also be found in Breslow, McNeney and Wellner (2003), where the properties of estimation in non-parametric models are also investigated in a double-sampling framework. The starting point of the present paper comes from Causeur and Husson (2007) that adapted the methodology to testing issues.

In section 2, some basics about multiple testing are recalled and the impact of a high correlation on the distribution of the error rates is discussed. Section 3 is dedicated to the definition of double-sampling t-statistics and section 4 addresses the statistical properties of a double-sampling Benjamini-Hochberg procedure. In section 5, the method is illustrated by microarray data used to select the genes that affect the degree of muscle destructuration in pigs.

2. Simultaneous Test of a Large Number of Hypotheses

Let $Y_{ij}^{(k)}$ be the jth replicate, $j = 1, \ldots, n_i^{(k)}$ of the kth variable, $k = 1, \ldots, K$, for the ith level of a factor. Hereafter, the case of a factor with only two levels will be considered. Usually, for example in the case of microarray data, K is very large and $n_i^{(k)}$ can be quite small. For gene expression data, the sample sizes $n_i^{(k)}$ are most often the same for a given i since it corresponds to the number of slides under condition i. However, missing data can occur, resulting for instance from technical concerns that have led to flag some spots on the microarray. The usual framework, in most situations where such kind of problems arise, is assumed, namely $Y_{ij}^{(k)} \sim \mathcal{N}(\mu_i^{(k)}; \sigma_k^2)$. The main goal is to point out the variables $Y^{(k)}$ for which the null hypothesis $H_0^{(k)} : \mu_1^{(k)} = \mu_2^{(k)}$, $k = 1, \ldots, K$, has to be rejected in favor of the alternative hypothesis $H_1^{(k)} : \mu_1^{(k)} \neq \mu_2^{(k)}$.

2.1 The false discovery rate

Most of the multiple testing strategies are based on the ranked p-values $p_1 \leq p_2 \leq \ldots \leq p_K$ of the t-tests used to compare the mean levels of the K variables under both conditions. Basically, procedures rely on the choice of a cut-off t such that, if $p_k \leq t$, $H_0^{(k)}$ is rejected. For each cut-off t, call V_t the number of false discoveries (or false positives), namely the number of variables for which $H_0^{(k)}$ is rejected although it is true. Call also R_t the observable number of variables for which $H_0^{(k)}$ is rejected. The False Discovery Rate for the cut-off t (FDR_t) is defined as the expected rate of false discoveries among the variables for which the null hypothesis is rejected:

$$\mathrm{FDR}_t = \mathbb{E}\left[\frac{V_t}{R_t}|R_t > 0\right]\mathbb{P}(R_t > 0).$$

Benjamini and Hochberg (1995) suggest to choose t among the ordered p-values p_k. Suppose first that the number m_0 of true $H_0^{(k)}$ is known. If $t = p_k$, then $R_t = k$ and, assuming the p-values are independently and uniformly distributed, an intuitive estimator of FDR_t is given by $\widehat{\mathrm{FDR}}_{p_k} = m_0 p_k/k$. Now, if k^* denotes the largest k such that $\widehat{\mathrm{FDR}}_{p_k} \leq \alpha$, then the cut-off is p_{k^*}. Usually, $K - m_0$ is negligible with respect to K which allows the replacement of m_0 by K in the former procedure. Significant improvements have emerged from central discussions about a better estimation of m_0, for instance, by step-up strategies, as in a recent paper by Benjamini, Krieger and Yekutieli (2006). Under a quite general assumption of positive dependency between the test statistics, Benjamini and Yekutieli (2001) show that such a procedure controls the FDR at level α.

2.2 The non-discovery rate

The discussions about simultaneously testing many hypotheses have so far focused on the Type I error rate. Dudoit *et al.* (2003) have however explored different definitions of the power of a multiple testing strategy. Among these definitions, $1 - \mathbb{E}(T_t/m_1)$ has been widely used (see Storey *et al.* (2004), Li *et al.* (2005)), where m_1 is the number of true $H_1^{(k)}$ and T_t the number of non-rejected $H_0^{(k)}$ that should have been rejected (false negatives). From a mathematical point of view, measuring the type II error rate by the Non-Discovery Rate $\mathrm{NDR}_t = \mathbb{E}(T_t/m_1)$ is however not consistent with the choice of FDR_t as a type I error rate. This have led Genovese and Wasserman (2002) to propose an alternative type II error rate, the False Non-discovery Rate $\mathrm{FNR}_t = \mathbb{E}(T_t/(K - R_t)|R_t < K)\mathbb{P}(R_t < K)$. As it seems that NDR_t makes more sense for practitioners, it will hereafter be preferred to FNR_t.

2.3 Impact of a high correlation on the distribution of the error rates

Due to the variable-by-variable approach in the procedures cited above, the basic framework for studying their statistical properties has often been independence between the variables. However, in many situations such as microarray experiments, it is well-known that this assumption is far from true. This have led many authors to propose corrections of the initial procedures that better accounts for dependency to control the type I error rate. Moreover, Owen (2005) showed that dependencies between the hypothesis tests greatly affect the variance of the number of false discoveries and provided an estimator for this variance taking into account the correlations between test statistics.

The following simulation study is intended to evaluate the impact of a high correlation on the distribution of both error rates. First, in 1000 datasets with two groups of $n = 10$ rows, 100 independent variables are simulated according to a normal distribution with standard deviation 1 and expectation 0 for half of the variables. For the remaining 50 variables, $\mu_2^{(k)} - \mu_1^{(k)}$ is set to 1.25. In 1000 other datasets, the same feature is reproduced except that each pair of variables has the same intra-group correlation $\rho = 0.90$. For each dataset, a Benjamini-Hochberg procedure is performed with a control of the FDR at level $\alpha = 0.05$ and the proportions of false negatives and false positives are calculated.

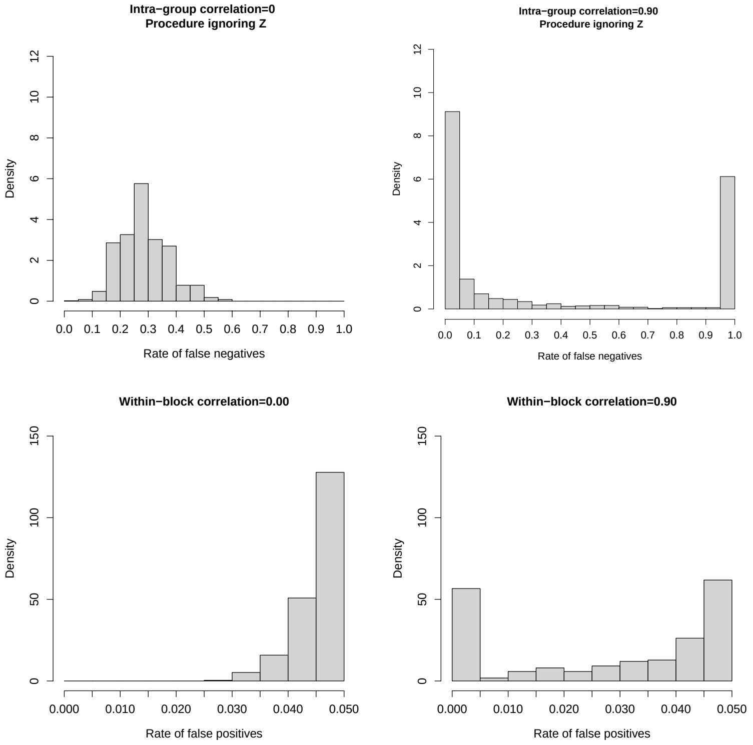

Figure 1: Distributions of the rates of false negatives and false positives

Figure 1, displaying the histograms of the error rates, shows a much larger dispersion of the error rates when the variables are highly correlated. In other words, in the case of a high correlation, the type II error rate is much more unstable than in the opposite case of independence. Note that the expected rate of false negatives, namely the NDR, is however, roughly speaking, the same.

3. Testing in the Presence of an Auxiliary Covariate

Although there is no methodological concern considering the case of many auxiliary covariates, the present paper focuses on the situation of only one covariate Z.

3.1 Double-sampling t-statistics

Suppose that measurements Z_{ij}, $j = 1, \dots, N_i$, of Z are available on a sample containing the $n^{(k)} = n_1^{(k)} + n_2^{(k)}$ items on which $Y^{(k)}$ is measured. In the following, Z_{ij} is assumed to be normally distributed with mean μ_i and standard deviation σ. Hereafter $N = N_1 + N_2$ denotes the size of the wider sample. Call $Y_{n^{(k)}}^{(k)}$ (resp. $Z_{n^{(k)}}$) the $n^{(k)}-$vectors of the observations of $Y^{(k)}$ (resp. Z) on the sample of size $n^{(k)}$. Call also Z_N the $N-$vector of the observations of the covariate Z on the whole sample of size N.

A useful way of defining the model in a double-sampling context consists in considering the $(n^{(k)}+N)-$vector obtained by concatenating $Y_n^{(k)}$ and Z_N. Under the assumptions mentioned above, $U^{(k)} = (Y_{n_1^{(k)}}^{(k)\prime}, Y_{n_2^{(k)}}^{(k)\prime}, Z'_{N_1}, Z'_{N_2})'$ is normally distributed with the following expectation and variance:

$$
\begin{array}{rcl}
\mathbb{E}(U^{(k)}) & = & \left[\begin{array}{cccc}
\mathbf{1}_{n_1^{(k)},1} & \mathbf{0}_{n_1^{(k)},1} & \mathbf{0}_{n_1^{(k)},1} & \mathbf{0}_{n_1^{(k)},1} \\
\mathbf{0}_{n_2^{(k)},1} & \mathbf{1}_{n_2^{(k)},1} & \mathbf{0}_{n_2^{(k)},1} & \mathbf{0}_{n_2^{(k)},1} \\
\mathbf{0}_{N_1,1} & \mathbf{0}_{N_1,1} & \mathbf{1}_{N_1,1} & \mathbf{0}_{N_1,1} \\
\mathbf{0}_{N_2,1} & \mathbf{0}_{N_2,1} & \mathbf{0}_{N_2,1} & \mathbf{1}_{N_2,1}
\end{array}\right]
\left(\begin{array}{c} \mu_1^{(k)} \\ \mu_2^{(k)} \\ \mu_1 \\ \mu_2 \end{array}\right), \\
\mathrm{Var}(U^{(k)}) & = & \left[\begin{array}{ccc}
\sigma_k^2 I_{n^{(k)}} & \rho_k \sigma \sigma_k I_{n^{(k)}} & \mathbf{0}_{n^{(k)},N-n^{(k)}} \\
\rho_k \sigma \sigma_k I_{n^{(k)}} & \multicolumn{2}{c}{\sigma^2 I_N} \\
\mathbf{0}_{N-n^{(k)},n^{(k)}} & &
\end{array}\right],
\end{array} \qquad (3.1)
$$

where ρ_k is the intra-condition correlation between $Y^{(k)}$ and Z and $\mathbf{0}_{m,p}$ (resp. $\mathbf{1}_{m,p}$) stands for the $m \times p$ matrix which all elements are 0 (resp. 1).

The above double-sampling context is a particular case of the general situation where the test of a General Linear Hypothesis, here $H_0^{(k)} : \mu_1^{(k)} = \mu_2^{(k)}$ against $H_1^{(k)} : \mu_1^{(k)} \neq \mu_2^{(k)}$, is considered under the assumption of a non-diagonal covariance structure. If the variance parameters are assumed to be known, the likelihood ratio-test statistic $T^{(k)}$ can be expressed as follows:

$$T^{(k)}(\rho_k, \sigma_k, \sigma) = \frac{\left[\bar{Y}_1^{(k)} - \bar{Y}_2^{(k)}\right] + \rho_k \frac{\sigma_k}{\sigma} \left\{\left[\bar{Z}_1 - \bar{Z}_2\right] - \left[\bar{z}_1 - \bar{z}_2\right]\right\}}{\sigma_k \sqrt{\frac{1}{n_1^{(k)}} + \frac{1}{n_2^{(k)}}} \sqrt{1 + \rho_k^2 \left[\frac{f_1^{(k)} f_2^{(k)}}{f^{(k)}} - 1\right]}}, \quad (3.2)$$

where $f_i^{(k)} = n_i^{(k)}/N_i$, $f^{(k)} = n^{(k)}/N$ are respectively the intra-condition and global sampling fractions, $\bar{Z}_i$ and $\bar{z}_i$ are the intra-condition means of Z on the samples of size N_i and $n_i^{(k)}$ respectively. Note that, if $\rho_k = 0$ or if the sampling fractions are 1, $T^{(k)}$ coincides with the usual t-statistic derived on the small sample, which means that no improvement is to be expected from the covariate.

According to Causeur (2005), the maximum-likelihood estimators of the variance parameters are given by the following expressions:

$$\begin{aligned}
\hat{\sigma}^2 &= \left\{\hat{\sigma}^2\right\}^{(N)}, \\
\hat{\sigma}_k^2 &= \left\{\hat{\sigma}_k^2\right\}^{(n)} + \left[\frac{\hat{\sigma}_{kz}^{(n)}}{\left\{\hat{\sigma}^2\right\}^{(n)}}\right]^2 \left[\left\{\hat{\sigma}^2\right\}^{(N)} - \left\{\hat{\sigma}^2\right\}^{(n)}\right], \\
\hat{\rho}_k &= \frac{\hat{\sigma}_{kz}^{(n)}}{\left\{\hat{\sigma}^2\right\}^{(n)}} \frac{\hat{\sigma}}{\hat{\sigma}_k}.
\end{aligned}$$

where $\hat{\sigma}_{kz}^{(n)}$, $\left\{\hat{\sigma}_k^2\right\}^{(n)}$, $\left\{\hat{\sigma}^2\right\}^{(n)}$ and $\left\{\hat{\sigma}^2\right\}^{(N)}$ are the usual maximum-likelihood estimators of the intra-group covariance σ_{kz} between $Y^{(k)}$ and Z and the intra-group variances σ_k^2 and σ^2, based on the residual sum-of-squares on the sample of size N or on the sub-sample of size $n^{(k)}$. The double-sampling t-statistics, $\hat{T}_k = T^{(k)}(\hat{\rho}_k, \hat{\sigma}_k, \hat{\sigma})$, integrating the measurements of the covariates, are obtained by plugging in the ML estimators of the variance parameters in expression (3.2).

3.2 Small-sample distribution

In some traditional fields of application of the multiple testing methods, a very small number of replications in each group is rather frequent. Therefore, there is an actual need in a non-asymptotic approximation of the double-sampling test statistics' distribution. It is straightforward deduced from Causeur and Husson (2007) that the distribution of $\hat{T}_k$ can be approximated by $T_{n^{(k)},N}(\rho_k^2)$, defined as follows:

$$\begin{aligned} T_{n^{(k)},N}(\rho_k^2) &= \sqrt{n_1^{(k)}+n_2^{(k)}-2}\sqrt{1+\frac{\rho_k^2}{1-\rho_k^2}\frac{f_1^{(k)}f_2^{(k)}}{f^{(k)}}} \\ &\quad \times \frac{T_1}{\sqrt{T_2+\frac{n_1+n_2-2}{N_1+N_2-2}\frac{f_1^{(k)}f_2^{(k)}}{f^{(k)}}T_3}}, \end{aligned} \tag{3.3}$$

where T_1, T_2 and T_3 are mutually independent. In addition, if $\delta_n^{(k)} = (\mu_2^{(k)} - \mu_1^{(k)})/\sigma_k$, T_1 is distributed according to a normal distribution with expectation

$$\delta_{n,N}(\rho_k) = \delta_n^{(k)}/\left\{\sqrt{\rho_k^2\left[(1/N_1)+(1/N_2)\right]+(1-\rho_k^2)\left[(1/n_1^{(k)})+(1/n_2^{(k)})\right]}\right\}$$

and standard deviation 1. T_2 is distributed according to a $\chi^2_{n_1+n_2-3}$ distribution. Suppose now that B and S are independent random variates following respectively a $\mathcal{B}([n_1^{(k)}+n_2^{(k)}-2]/2, [N_1+N_2-n_1^{(k)}-n_2^{(k)}]/2)$ and a $\chi^2_{N_1+N_2-2}$, then T_3 is conditionally distributed, given B and S, as the ratio between a non-central chi-square variable with 1 degree of freedom and non-centrality parameter $[\rho_k^2/(1-\rho_k^2)]BS$ and B.

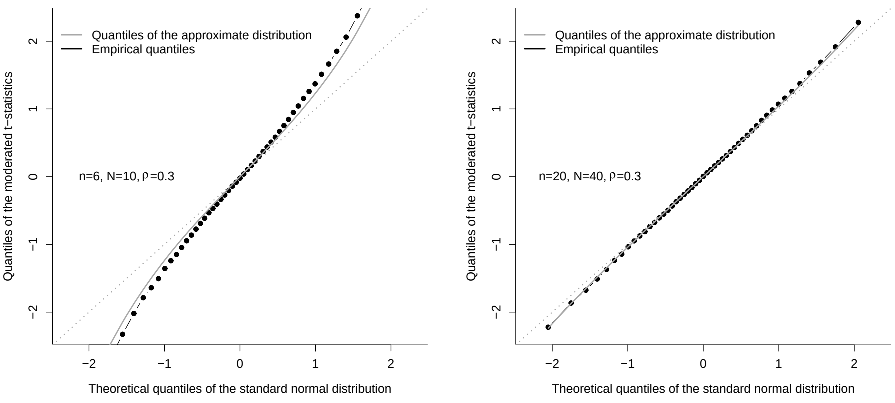

Figure 2: Quantile-quantile plots for the distribution of the moderated t-statistics. The empirical quantiles are derived from 5000 simulations whereas the quantiles of the approximate distribution are deduced from 3.3 by Monte-Carlo methods.

Even in small-sample conditions, $T_{n,N}(\rho_k^2)$ is a good approximation of the distribution of $\hat{T}_k$. In order to illustrate the former result, the empirical quantiles of the test statistic based on 5000 simulations are derived and compared to the theoretical quantiles of the approximate distribution under two schemes: $n = 6,\ N = 10,\ \rho_k = 0.3$ and $n = 20,\ N = 40,\ \rho_k = 0.3$. The quantile-quantile plot assessing the closeness of the two distributions is displayed in Figure 2.

3.3 Power of the double-sampling test

First, let us consider that the variance parameters are known. It is straightforward checked that the distribution of $T^{(k)}(\rho_k, \sigma_k, \sigma)$ is then normal with mean $\delta_{n,N}(\rho_k)$ where $\delta_{n,N}^{-2}(\rho_k)$ can be expressed as a convex linear combination of $\delta_n^{-2} = \delta_{n,N}^{-2}(0)$ and $\delta_N^{-2} = \delta_{n,N}^{-2}(1)$:

$$\delta_{n,N}^{-2}(\rho_k) = (1 - \rho_k^2)\delta_n^{-2} + \rho_k^2 \delta_N^{-2}. \tag{3.4}$$

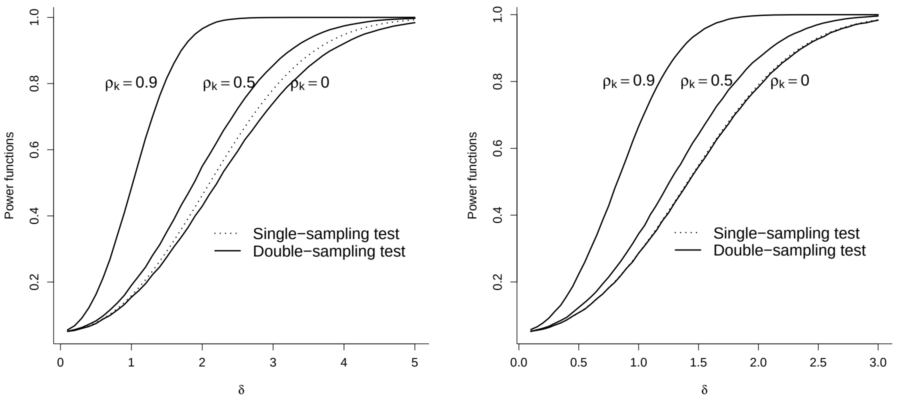

Figure 3: Power functions for the double-sampling test. On the left plot, $n_1^{(k)} = n_2^{(k)} = 3$ and $N_1 = N_2 = 20$. On the right plot, $n_1^{(k)} = n_2^{(k)} = 5$ and $N_1 = N_2 = 20$.

Note that δ_n and δ_N are the expectations of the test statistics calculated on the sample of size $n^{(k)}$ and N respectively. Therefore, expression (3.4) implies that the power of the double-sampling test is always larger than the power of the test based on the small sample only (equality holds if $\rho_k = 0$) and always smaller than the test that would be based on the sample of size N (equality holds if $\rho_k = 1$).

When the variance parameters are no longer assumed to be known, the preceding result remains true asymptotically. However, in the case of a small value of ρ_k^2 and a small sample size $n^{(k)}$, the usual single-sampling test on the small sample shall be preferred to the double-sampling strategy. This is particularly obvious in the left plot of Figure 3 displaying the power functions for various values of ρ_k together with the power function of the t-test on the small-sample in the case $n_1^{(k)} = n_2^{(k)} = 3$ and $N_1 = N_2 = 20$. On the right plot, showing the same functions with $n_1^{(k)} = n_2^{(k)} = 5$, the difference between the power function of the single-sampling t-test and the double-sampling t-test for values of ρ_k close to zero is much thinner.

4. Double-sampling Benjamini-Hochberg Procedure

Obviously, the testing method described above show desirable properties when the double-sampling scheme can take advantage of an auxiliary variable that is well correlated with the response. In the studies of high flow experimental data, it is of course most probably impossible to find relevant auxiliary covariates that can exhaustively be used to improve the power of each test. However, as illustrated in the next section, a pre-filtering of the variables with respect to their high correlation with some auxiliary covariates points out sets of variables for which the double-sampling tests can be beneficial.

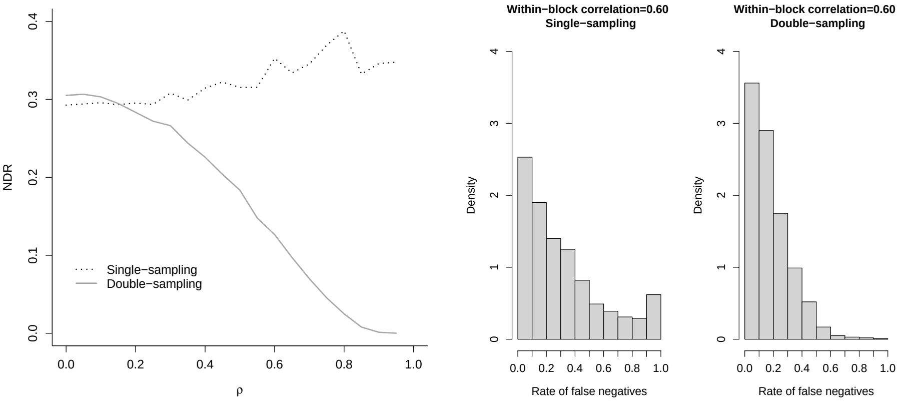

Figure 4: Left plot: NDR for various values of ρ. Right plot: Histograms of the rates of false negatives for the single-sampling and the double-sampling approaches ($\rho = 0.6$)

The multiple tests procedure focusing on these particular sets of variables only differs from the Benjamini-Hochberg approach described in section 2 by the variable-by-variable tests on which it is based. Let us denote $\tilde{p}_1 \leq \tilde{p}_2 \leq \ldots \leq \tilde{p}_K$ the p-values of the double-sampling t-tests used to compare the mean levels of the K variables under both conditions. For a given cut-off $t = \tilde{p}_k$, such that, if $\tilde{p}_k \leq t$, $H_0^{(k)}$ is rejected, the proposed estimator of FDR_t is now given by $\widetilde{\mathrm{FDR}}_{\tilde{p}_k} = m_0 \tilde{p}_k / k$. By analogy with the Benjamini-Hochberg procedure, if k^* denotes the largest k such that $\widetilde{\mathrm{FDR}}_{\tilde{p}_k} \leq \alpha$, then the chosen cut-off is $\tilde{p}_{k^*}$.

The following simulation study aims at showing the impact of the above modification on the power of the procedure. For various values ρ, 1000 datasets are simulated, with two groups of $n_1 = n_2 = 10$ rows and 100 variables, normally distributed with a null difference between the means in both groups for half of the variables and a difference of 1.25 for the second half of the variables, standard deviation 1 and an equal intra-group correlation ρ with an auxiliary covariate Z. Z is itself normally distributed with standard deviation 1 and $\mu_2 - \mu_1 = 2$. $N_1 = N_2 = 75$ observations of Z are available in each group, among which the $n_1 = n_2 = 10$ rows for which the responses are observed. For each dataset, both the usual Benjamini-Hochberg procedure based on the single-sampling t-tests and the modified Benjamini-Hochberg based on the double sampling-scheme are performed with a control of the FDR at level $\alpha = 0.05$.

Figure 4 shows the decrease of the mean NDR of the double-sampling procedure when ρ increases. It also shows that the mean NDR of the single-sampling strategy remains quite unchanged for all the values of ρ. In fact, the perturbed form of the plot is due to the high dispersion, already mentioned above, in the distribution of the rate of false negatives. Figure 4 also displays histograms of the rates of false negatives for the two approaches for $\rho = 0.6$ and shows that the double-sampling method reduces the dispersion of the error rates.

5. Application to Microarray Data

The above double-sampling method is implemented in a situation where microarray data ($n_1 = n_2 = 7$) containing information on $K = 3442$ genes are analyzed to select the genes that affect the degree of muscle destructuration in pigs. The two groups are coded as 1 for high quality and 2 for structureless meat. The original gene expressions were normalized in two steps: a logarithmic transformation and a global mean normalization (same mean for each microarray) were performed. The auxiliary variable Z introduced in the double-sampling scheme is the pH, measured for $N = 163$ pigs ($N_1 = 87,\ N_2 = 76$). This covariate is indeed known to play a role in the destructuration of meat, which explains the highly significant difference between the mean values of Z in the two groups (p-value $< 10^{-10}$).

Hereafter, two Benjamini-Hochberg procedures, with a control of the FDR at level $\alpha = 0.05$, are compared:

- $\mathsf{BH}_{\mathsf{ss}}$, the usual single-sampling method based on the p-values p_k of the t-tests,
- $\mathsf{BH}_{\mathsf{ds}}$, the method based on the p-values $\tilde{p}_k$ of the double-sampling t-statistics integrating Z. In fact, only 202 gene expressions, which squared intra-group correlation with Z is larger than 0.25 are subject to the double-sampling

correction while the remaining test statistics are left unchanged. The results of a Principal Component Analysis performed on the gene expressions show the correlation structure in this set of genes (see Figure 5).

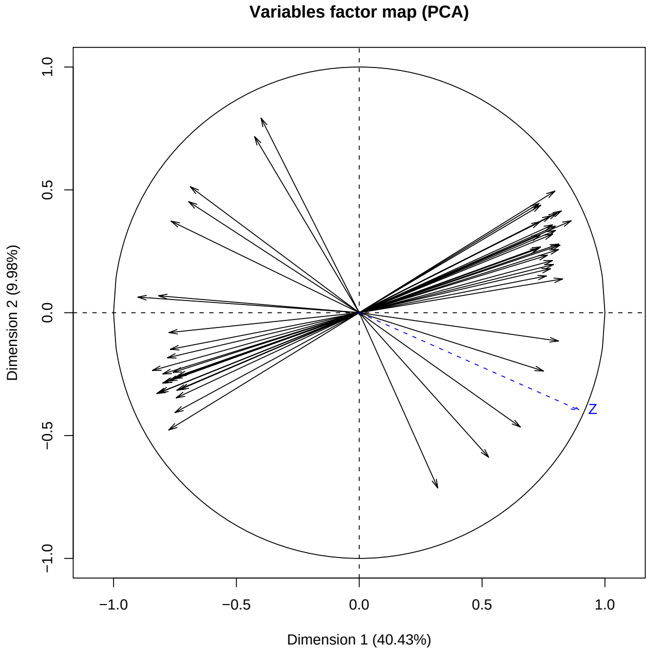

Figure 5: Principal Components Analysis on the set of selected gene expressions (for convenience, only the variables for which the correlation with the 2D factor map is higher than 0.6 are represented).

Figure 6 displays the scatterplot of p_k versus $\tilde{p}_k$ for these 202 gene expressions and point out some disagreements between the single-sampling and the double-sampling approaches. As pointed out by Causeur and Dhorne (1998), a statistical property of the double-sampling inference can help interpreting such disagreements. Indeed, the numerator of the test statistics defined by expression (3.2) is the difference between the double-sampling estimators of $\mu_1^{(k)}$ and $\mu_2^{(k)}$. Furthermore, these estimators are also the intra-group means of the N values of $Y^{(k)}$ obtained by appending the $n^{(k)}$ observed values and the $N - n^{(k)}$ predicted values from Z by the analysis of covariance model fitted on the small sample. The double-sampling test-statistics is therefore similar to a classical t-statistics derived on the whole sample with imputed values. However, the variance that appears in the denominator of expression (3.2) accounts for the dispersion due to the imputations of $N - n^{(k)}$ values.

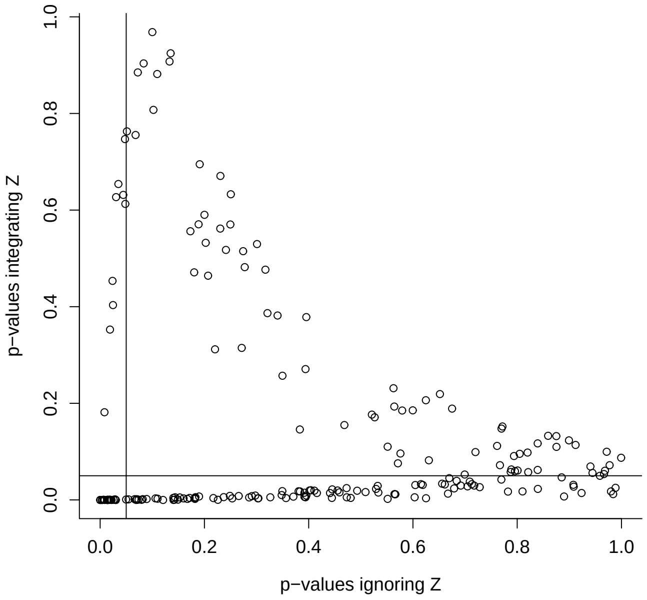

Figure 6: Plot of the single-sampling versus the double-sampling p-values for the set of gene expressions correlated to Z.

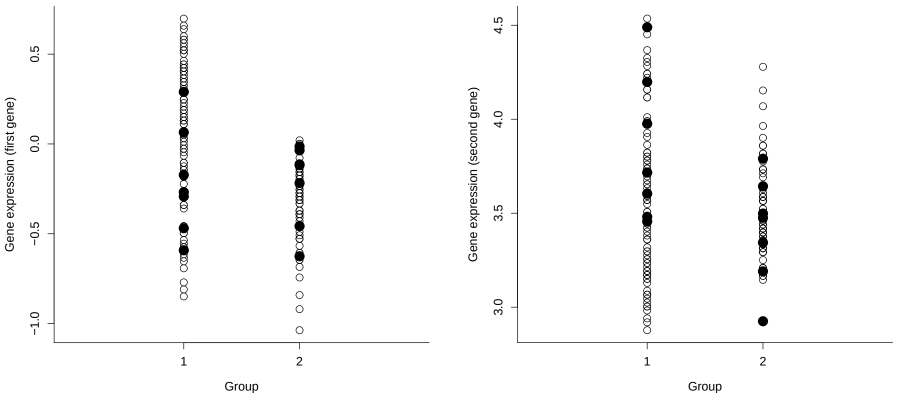

Figure 7: Two examples of disagreements between the single-sampling approach (test based on the observed values identified by black dots) and the double-sampling strategy (test based on both observed and predicted values identified by white dots)

In an illustrative purpose, the results for two gene expressions, say $Y^{(1)}$ and $Y^{(2)}$, are detailed here. For $Y^{(1)}$, $H_0^{(1)}$ is rejected with the double-sampling method (p-value = 0.007) whereas it is not with the t-test (p-value = 0.89), and for $Y^{(2)}$ the situation is opposite (p-values of 0.035 vs. 0.65). For both genes, Figure 7 plots the observed values of the responses together with the imputed values. The plots reveal that the 7 observed values in each group are not representative of the expected distribution regarding their pH, which explains the large difference between the p-values.

Finally, Table 1 gives the 2 × 2 table summarizing the number of genes declared as differentially observed with both methods. In this particular case, increasing the power of the test by use of the pH have resulted in declaring more genes as differentially expressed.

Table 1: Number of positive and negative genes in the selected set of genes for both methods

	$\mathsf{BH_{ss}}$: single-sampling tests		
$\mathsf{BH_{ds}}$: double-sampling tests	Positive	Negative	Sum
Positive	8	58	66
Negative	0	3376	3376

6. Concluding Remarks

In the present paper, a new method for simultaneous tests of a large number of hypotheses is presented and shown to improve the power of existing methods by integrating external information. Enhancements are made possible by an auxiliary variable which measurements are available together with the other variables on a small sample and also on additional items. First, a variable-specific double-sampling test statistics is proposed and its small-sample distribution is provided. This enables precise calculations of the power function for a given intra-group correlation. By analogy with the Benjamini-Hochberg procedure in the single-sampling case, a procedure for multiple tests is deduced in a double-sampling scheme. For a relevant choice of an auxiliary variable, as highly correlated with the responses as possible, the type II error rate is shown to be reduced. Moreover, integrating external information also stabilizes the dispersion of the error rates.

For convenience, it has been chosen to focus on the case of only one covariate while Causeur and Husson (2007) provide the statistical tools to extend the present results to a multivariate situation. Although this extension is not necessary to understand how auxiliary information can be used to improve the power of the tests, it is however necessary for most applications of multiple testing since it allows a more comprehensive approach of high dimensional data. It also raises the issue of the selection of covariates in order to give insight to the joint distribution of the test statistics, at least within blocks of variables. This idea is probably not far from the suggestion made by von Heydebreck *et al.* (2004), in the context of the analysis of microarray experiments, of a preliminary filtering of the genes by use of metadata before the actual multiple testing method is applied.

The small-sample distribution of the double-sampling test statistics that is given above gives numerical tools to study the impact of the intra-condition correlation on the distribution of the error rates. However, further mathematical developments are needed to derive closed-form expressions for the moments of these error rates. First, this would assess the control of the False Discovery Rate in a double-sampling framework. Moreover, this would yield tools to derive the optimal allocation for the numbers of observations of the responses and the covariates for a target expected non-discovery rate.

Acknowledgements

The authors are grateful to Marie Damon, from the Unité Mixte de Recherches Systèmes d'Élevage, Nutrition Animale et Humaine of the Institut National de la Recherche Agronomique, who provided the microarray data used in the last section of this paper.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* B **57**, 289-300.

Benjamini, Y., Krieger, A. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491-507.

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

Breslow, N., McNeney, B. and Wellner, J. (2003). Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling. *Annals of Statistics* **31**, 1110-1139.

Causeur, D. (2005). Optimal sampling from concomitant variables for regression problems. *Journal of Statistical Planning and Inference* **128**, 289-301.

Causeur, D. and Dhorne, T. (1998). Finite-sample properties of a multivari- ate extension of double-regression. *Biometrics* **54**, 1591-1601.

Causeur, D. and Husson, F. (2007). Asymptotic distribution of double-sampling tests for general linear hypotheses. *Statistics* **42** 115-125.

Cochran, W. (1963). *Sampling Techniques*, 2nd edition. Wiley.

Conniffe, D. (1985). Estimating regression equations with common explanatory variables but unequal numbers of observations. *Journal of Econometrics* **27**, 179-196.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71-103.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the fdr procedure. *Journal of the Royal Statistical Society* B **64**, 499-518.

Kendziorski, C., Newton, M., Lan, H. and Gould, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899-3914.

Li, S., Bigler, J., Lampe, J., Potter, J. and Feng, Z. (2005). FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine* **24**, 2267-2280.

LAonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31-46.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society* B **67**, 411-426.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, Iss. 1, Article 3.

Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society* B **66**, 187-205.

von Heydebreck, A., Huber, W. and Gentleman, R. (2004). Differential expression with the bioconductor project. Bioconductor Project Working Papers.

Received August 14, 2007; accepted October 4, 2007.

Maela Kloareg
IRMAR UMR 6625 CNRS
Laboratoire de Mathématiques Appliquées Agrocampus Rennes
CS 84215, 65 rue de St-Brieuc
35042 Rennes cedex, France.
Maela.Kloareg@agrocampus-ouest.fr

David Causeur
IRMAR UMR 6625 CNRS
Laboratoire de Mathématiques Appliquées Agrocampus Rennes
CS 84215, 65 rue de St-Brieuc
35042 Rennes cedex, France.
David.Causeur@agrocampus-ouest.fr