The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 8

Approximate Power and Sample Size Calculations with the Benjamini-Hochberg Method

José António Ferreira, *AMC - University of Amsterdam*
Aeilko H. Zwinderman, *University of Amsterdam*

Recommended Citation:

Ferreira, José António and Zwinderman, Aeilko H. (2006) "Approximate Power and Sample Size Calculations with the Benjamini-Hochberg Method," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 8.

DOI: 10.2202/1557-4679.1018

Approximate Power and Sample Size Calculations with the Benjamini-Hochberg Method

José António Ferreira and Aeilko H. Zwinderman

Abstract

We provide a method for calculating the sample size required to attain a given average power (the ratio of rejected hypotheses to the number of false hypotheses) and a given false discovery rate (the number of incorrect rejections divided by the number of rejections) in adaptive versions of the Benjamini-Hochberg method of multiple testing. The method works in an asymptotic sense as the number of hypotheses grows to infinity and under quite general conditions, and it requires data from a pilot study. The consistency of the method follows from several results in classical areas of nonparametric statistics developed in a new context of "weak" dependence.

KEYWORDS: multiple testing, false discovery rate, density estimation, estimation of distribution functions, power calculation, sample size calculation, analysis of microarray data

Author Notes: Jose Ferreira is funded by 'Nederlandse Hartstichting', the Dutch Heart Foundation, through the grant no. 2003B215. We thank T. Bouezmarni for explaining us part of his work in Bouezmarni and Rolin (2003), and a referee (Reviewer #3) for suggestions which helped improving our presentation.

1 Introduction

Classical multiple testing procedures such as the Bonferroni method are quite powerless in genomic research and other types of research involving large numbers of hypotheses, and statisticians have recently developed alternative methods which turn out to be useful and robust. The most important of these methods seems to be that of Benjamini and Hochberg (1995), which by now has a number of variants and extensions. The Benjamini-Hochberg method applies to the p-values computed from a given set of test statistics and consists of rejecting all hypotheses whose associated p-values fall below a data-dependent threshold in such a way that the *false discovery rate*—the ratio of incorrectly rejected hypotheses to the total number of rejected hypotheses—is kept near a desired, pre-specified level. In its adaptive form, which relies on a consistent estimate of the asymptotic *fraction γ of true (or null) hypotheses*, the method not only controls the false discovery rate as it allows the estimation of *average power*—the proportion of correctly rejected hypotheses out of all false (or alternative) hypotheses. These two properties hold asymptotically as the number of hypotheses tends to infinity, provided certain weak requirements are met; they follow from earlier work in Ferreira and Zwinderman (2006b) and will be reviewed and slightly extended below in Section 2.

Naturally, the possibility of estimating power brings up the question of how to carry out sample size calculations, which is of great interest in controlled experiments. In genomics, for example, where researchers try to discover which genes contribute to a given physiological process through so-called microarray experiments, it is often important as part of the planning of an experiment to know roughly how many microarray replicates are required in order to detect a given proportion of genes. The purpose of this article is to show how this question can be formulated and answered within the Benjamini-Hochberg approach to multiple testing provided a consistent estimator of γ is available and certain conditions on the sequence of p-values are fulfilled. The basic ideas and calculations are contained in Subsection 2.2. Strictly speaking, we restrict ourselves to considering a ‘location model’ in which the distributions of the test statistics belong to a location family, but this covers the most common practical applications involving t-statistics and the like. Needless to say, the practical application of our method requires data from some sort of pilot study. [Alternatively, one might play with equation (2.17) below by varying distributions and parameters, but we think that this approach is not very useful in practice.]

The consistency of the adaptive Benjamini-Hochberg procedure and of the associated method of sample size calculation depends on the consistency of em-

pirical distribution functions, density estimators, and certain deconvolution estimators, which in turn depend on the properties of the sequence of p-values. Accordingly, much of the paper is devoted to developing weak dependence conditions under which consistency is possible and to proving that consistency does hold. More specifically, in Section 3 we consider two conditions on sequences of random variables similar in spirit to the mixing conditions of time series analysis and extreme value theory and which, besides being much more general than the latter, do not presuppose any kind of time or spatial ordering in the index of the variables; these conditions are then used to prove convergence results on the three types of estimators mentioned above. The simple framework of Section 3 builds on earlier work of Genovese and Wasserman (2002), Storey et al. (2004) and Ferreira and Zwinderman (2006b); coupled with the Benjamini-Hochberg method as presented in Section 2, it provides a basis for much of the statistical analysis involved in multiple testing problems with very large numbers of hypotheses and, we think, a quick route for further progress in the field.

In Section 4 we apply the convergence results of Section 3 to obtain a ‘Storey-type’ non-parametric estimator of γ based on density estimators (following an approach already considered by Langaas *et al.* (2005)). In Section 5 we use the result on deconvolution estimators to prove the consistency of a ‘semi-parametric’ estimator of γ and of what is essentially the distribution function of the p-values associated to the false hypotheses, from which follows the consistency of our method of sample size calculation.

Finally, in Section 6 we discuss briefly the practical implementation and interpretation of the method.

One perhaps interesting conclusion that follows from our work is that the new statistical problem of testing a very large number of hypotheses can actually be solved with a combination of rather classical methods in nonparametric estimation, even if these need to be suitably adapted; and since such methods work best with large samples, one may well say that in the Benjamini-Hochberg approach to multiple testing the ‘high-dimensionality’ of the problems it addresses is actually a blessing rather than a curse.

Solutions to the problem of calculating sample size in microarray experiments have also been proposed by Lee and Whitmore (2002), Pan, Lin, and Le (2002), Müller et al. (2004), Pawitan et al. (2005), and Tibshirani (2005), but our method is more in line with the Benjamini-Hochberg approach to multiple testing and, we think, simpler and more realistic; it is the only method we know of which provides an explicit equation for calculating sample size on the basis of a couple of theoretical objects whose empirical counterparts can be estimated in practice.

2 Key results on the Benjamini-Hochberg method

Throughout, $X_1, X_2, \dots$ are a sequence of random variables (r.v.'s) on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in $[0, 1]$. We write $n_0 = \gamma_n n$, $n_1 = (1 - \gamma_n)n$, where $n_0, n_1, n \in \mathbb{N}$, $\gamma_n = n_0/n \rightarrow \gamma$ as $n \rightarrow \infty$, and γ is a fixed number in $(0, 1)$, and put $X'_j = X_j$ for $j = 1, \dots, n_0$, and $X''_j = X_{n_0+j}$ for $j = 1, \dots, n_1$. F will always denote the distribution function (d.f.) of a standard uniform r.v. Unless otherwise stated, all limits are taken as $n \rightarrow \infty$.

We shall think of $X'_1, \dots, X'_{n_0}$ and $X''_1, \dots, X''_{n_1}$ as the p-values computed from a set of test statistics, the first n_0 of which ‘under true/null hypotheses’ and the remaining n_1 ‘under false/alternative hypotheses’, and of γ as the asymptotic *proportion of true/null hypotheses*; accordingly, we may think of the p-values as a mixture of uniform and non-uniform r.v.’s with mixture coefficient γ . More precisely, we introduce the following assumptions, which seem practically meaningful and yield useful results about the relevant quantities in the Benjamini-Hochberg method:

$$F_n(u) := \frac{1}{n_0} \sum_{j=1}^{n_0} 1_{\{X'_j \leq u\}} \xrightarrow{P} F(u) \quad (2.1)$$

and

$$G_n(u) := \frac{1}{n_1} \sum_{j=1}^{n_1} 1_{\{X''_j \leq u\}} \xrightarrow{P} G(u) \quad (2.2)$$

uniformly in u , where $G \neq F$ is some d.f. on $[0, 1]$. From (2.1) and (2.2) follows in particular that $\sup_u |H_n(u) - H(u)| \xrightarrow{P} 0$, where H_n is the empirical d.f. of $\{X_1, \dots, X_n\}$ and

$$H = \gamma F + (1 - \gamma)G. \quad (2.3)$$

Except where explicitly stated, (2.1) and (2.2) will be assumed to hold uniformly throughout the paper; in Section 3 we study conditions that seem very general and relevant in genomic research and which guarantee versions of (2.1) and (2.2). Further conditions will be introduced later on.

The Benjamini-Hochberg procedure consists of fixing $q \in [0, 1]$ and of rejecting the hypotheses whose p-values fall below the random threshold $X_{R_n(q):n}$ defined by

$$R_n(q) := \max \left\{ i : X_{i:n} \leq q \frac{i}{n} \right\},$$

where $0 \leq X_{1:n} \leq \dots \leq X_{n:n} \leq 1$ denote the order statistics of the sample of p-values. The r.v. $R_n(q)$ is called the *number of rejections*. If $S_n(q)$ denotes the number of p-values below $X_{R_n(q):n}$ which belong to the set $\{X'_1, \dots, X'_{n_0}\}$ —the number of incorrectly rejected hypotheses—then the r.v. $\Pi_{1,n}(q) = S_n(q)/(R_n(q) \vee 1)$ gives the proportion of incorrect rejections out of all rejections, and its expected value is called *false discovery rate*. The r.v. $\Pi_{2,n}(q) = (R_n(q) - S_n(q))/n_1$, the proportion of rejected p-values from $\{X''_1, \dots, X''_{n_1}\}$, is called the proportion of detected false hypotheses and its expected value is called *average power*. When they exist, the limits of $E\Pi_{1,n}(q)$ and $E\Pi_{2,n}(q)$ will sometimes be referred to by false discovery rate and average power, respectively.

The main question in the Benjamini-Hochberg method is whether the proportion of rejections converges to a constant. Under conditions like (2.1) and (2.2), it has been shown by Genovese and Wasserman (2002), Storey et al. (2004) and Ferreira and Zwinderman (2006b), roughly in increasing order of generality, that

$$\frac{R_n(q)}{n} \rightarrow^P \rho(q, \gamma) = \psi_q^* \left(\frac{q(1-\gamma)}{(1-q)} \right), \quad (2.4)$$

where

$$\psi_q^*(v) = \min\{u \in [0, 1] : \psi_q(u) \leq 1/v\}, \quad v \geq 0, \quad (2.5)$$

$$\psi_q(u) = \sup_{qu \leq t \leq q} \frac{G(t) - t}{t}, \quad u \in [0, 1],$$

whenever ψ_q does not assume the value $\frac{(1-q)}{q(1-\gamma)}$ over an interval. A special case that will be needed later is this: If G is concave and its right-hand derivative at 0 is $G'_+(0) > (1-q\gamma)/[q(1-\gamma)]$, then $q\rho(q, \gamma)$ is the unique positive u such that $G(u) = u(1-q\gamma)/[q(1-\gamma)]$ (see Example ii. in Section 3 of Ferreira and Zwinderman (2006b)). For example, if Φ is the standard normal d.f. then the G of Subsection 2.2 is concave and such that $G'_+(0) = \infty$, so that $q\rho(q, \gamma)$ can be computed as indicated for all $q, \gamma \in (0, 1)$. We observe, however, that the result holds more generally: if Φ is a Student d.f. then the G of Subsection 2.2 is concave except for a small right neighbourhood of 0 and $G'_+(0) = 1$, but $q\rho(q, \gamma)$ is computed in exactly the same way, even though it will not be > 0 if q is chosen small enough.

If (2.4) holds then (see Section 3 of Ferreira and Zwinderman (2006b))

$$\Pi_{1,n}(q) \rightarrow^P q\gamma \quad (2.6)$$

and

$$\Pi_{2,n}(q) \rightarrow^P \rho(q, \gamma) \frac{(1-q\gamma)}{(1-\gamma)}. \quad (2.7)$$

Thus, as long as the proportion of rejections converges, both the proportion of incorrect rejections out of all rejections and the proportion of correct rejections converge, the first to the false discovery rate and the second to average power.

If γ and G are known, one can obviously compute everything of interest from (2.6) and (2.7), namely the asymptotic false discovery rate and average power that result from a particular choice of q , and thus decide on which value of q to take. In most practical problems, however, γ and G need to be estimated. If $\hat{\gamma}_n$ is a consistent estimator of γ (i.e. such that $\hat{\gamma}_n \rightarrow^P \gamma$) then $\hat{H}_n := \hat{\gamma}_n F_n + (1 - \hat{\gamma}_n)G_n \rightarrow^P \gamma F + (1 - \gamma)G$ uniformly, and therefore

$$\hat{G}_n := \frac{\hat{H}_n - \hat{\gamma}_n F}{1 - \hat{\gamma}_n} \rightarrow^P G \quad (2.8)$$

uniformly. Thus, to each consistent estimator of γ there corresponds a consistent estimator of G , and the whole problem of finding estimates of the right-hand sides of (2.6) and (2.7) can be solved by estimating γ .

2.1 Adaptive versions

Observe that when γ is known (2.6) and (2.7) tell us that in order to guarantee an approximate false discovery rate of $\delta \in (0, 1)$ and at the same time achieve maximum asymptotic average power one should take $q = \delta/\gamma$. This suggests that if $\hat{\gamma}_n$ is not too bad an estimator of γ then one might be better off applying the Benjamini-Hochberg procedure with $\delta/\hat{\gamma}_n$ in place of q . By an *adaptive version of the Benjamini-Hochberg method* we mean the Benjamini-Hochberg method with q replaced by $q_n := \delta/\hat{\gamma}_n$, where $\hat{\gamma}_n$ is some estimator of γ and $\delta \in (0, \gamma)$ stands for the false discovery rate one is ready to tolerate. Adaptive versions were considered first by Benjamini and Hochberg (2000) and then by Storey (2002).

As Theorem 4.1 in Ferreira and Zwinderman (2006b) illustrates, adaptive versions of the Benjamini-Hochberg method improve upon the original one even if $\hat{\gamma}_n$ is biased. Moreover, if $\hat{\gamma}_n$ is consistent then so is the adaptive Benjamini-Hochberg procedure in the sense that

$$\frac{R_n(q_n)}{n} \rightarrow^P \rho\left(\frac{\delta}{\gamma}, \gamma\right), \quad \Pi_{1,n}(q_n) \rightarrow^P \delta, \quad (2.9)$$

and

$$\Pi_{2,n}(q_n) \rightarrow^P \rho\left(\frac{\delta}{\gamma}, \gamma\right) \frac{(1 - \delta)}{(1 - \gamma)} \quad (2.10)$$

provided (2.4) holds with $q = \delta/\gamma$.

The proof of this result is practically the same as that of Theorem 4.1 in Ferreira and Zwiderman (2006b) and will not be given here. What matters to us is its practical consequences: If $\hat{\gamma}_n$ is consistent, then in a given application of the adaptive Benjamini-Hochberg procedure $\Pi_{1,n}(q_n)$ will be close to δ , while

$$\frac{R_n(q_n)}{n} \frac{(1 - \delta)}{(1 - \hat{\gamma}_n)}$$

will be close to average power.

We emphasize the fact that one should choose $\delta < \gamma$ in the above. Fixing the false discovery rate at $\delta = \gamma$ would amount to the rejection of *all* the p-values, which though naturally yielding a power of 1 would imply a relatively high rate of false discoveries.

2.2 Power and sample size in location models

It is clear from the above that in many practical problems the Benjamini-Hochberg method allows us to discover many false hypotheses without incurring more than a specified false discovery rate, and to estimate the proportion of true hypotheses as well as the d.f. of the p-values corresponding to the false hypotheses. The latter, however, depends on the power of the test statistics, hence on the sample size used. In order to carry out power calculations pertaining to future applications of the method one would therefore need to specify G for each sample size. Although an *a priori* specification of G based on intuition and past experience is always possible, it is hard to imagine that the typical practitioner will possess the conviction to infer the graph of a function corresponding to a given sample size from the graph of a function corresponding to the smaller sample size of a pilot study. On the other hand, one can often make assumptions about the data and the test statistics used that lead to a simple expression for G as a function of sample size. We shall study here what is perhaps the simplest example of such an approach, in which the d.f.'s of the test statistics belong to a location family.

We consider *test statistics* $T_1, \dots, T_n$ such that T_j has d.f. $t \rightarrow \Phi(t - \theta_j \sqrt{N})$, $j = 1, \dots, n$, where Φ is the d.f. of a symmetric, continuous r.v. whose characteristic function has no zeros, $\theta_1, \dots, \theta_n$ are parameters, and N is a parameter that plays the role of sample size. To represent the situation in which only the last $n_1 = n - n_0$ genes are ‘differentially expressed’ we let $\theta_j = 0$ for $j = 1, \dots, n_0$ and $\theta_j \neq 0$ for $j = n_0 + 1, \dots, n$, and we write $T'_j = T_j$ for $j = 1, \dots, n_0$, $T''_j = T_{n_0+j}$ for $j = 1, \dots, n_1$.

Example: For concreteness one may think of Φ as the standard normal d.f., of T_j as a normal r.v. with mean $\theta_j\sqrt{N}$ and variance 1, and of θ_j as the ‘size effect’ associated with the j -th gene, say. To see that this model is realistic in many situations, suppose that each $\bar{D}_j$ is a sum of random variables that is approximately normal with mean $N\mu_j$ and variance $N\sigma_j^2$, and let $N\hat{\sigma}_{j,N}^2$ be an estimate of its variance (as in the case of the paired and two-sample t-tests). If $T_j = \bar{D}_j/\sqrt{N\hat{\sigma}_{j,N}^2}$ and $\hat{\sigma}_{j,N} \approx \sigma_j$ then T_j is approximately normally distributed with mean $\theta_j\sqrt{N}$ and variance 1, where $\theta_j := \mu_j/\sigma_j$. The possibility of using a general d.f. Φ in place of a normal d.f. allows extra flexibility (e.g. heavier tails or variance $\neq 1$) and can be exploited in a sensitivity study. $\square$

Because the T_j' 's are symmetric, it makes sense to define the p-values $X_1, \dots, X_n$ associated with $T_1, \dots, T_n$ as $X_j(\omega) = P\{\omega' : |T_j(\omega')| > |T_j(\omega)|\} = 2[1 - \Phi(|T_j(\omega)|)]$, $\omega \in \Omega$. Consequently, the d.f. of X_j is given by

$$\Gamma_{\theta_j}(u) = 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{u}{2}\right) - \theta_j\sqrt{N}\right) + \Phi\left(-\Phi^{-1}\left(1 - \frac{u}{2}\right) - \theta_j\sqrt{N}\right); \quad (2.11)$$

in particular, $X'_1, \dots, X'_{n_0}$ are standard uniform r.v.'s.

Denoting by Λ_n the empirical d.f. constructed from $\{\theta_{n_0+1}, \dots, \theta_n\}$ and putting $\mathbf{G}_n(u) = \int \Gamma_{\theta}(u) d\Lambda_n(\theta)$, we see that the d.f. of a randomly chosen element of $\{X_1, \dots, X_n\}$ is

$$\mathbf{H}_n = \gamma_n F + (1 - \gamma_n) \mathbf{G}_n.$$

Clearly, $\mathbf{H}_n(u)$ is also the expected value of $H_n(u)$:

$$\begin{aligned} E H_n(u) &= \gamma_n \frac{1}{n_0} \sum_{j=1}^{n_0} P(X'_j \leq u) + (1 - \gamma_n) \frac{1}{n_1} \sum_{j=1}^{n_1} P(X''_j \leq u) \\ &= \gamma_n F(u) + (1 - \gamma_n) \frac{1}{n_1} \sum_{j=n_0+1}^n \Gamma_{\theta_j}(u) \\ &= \gamma_n F(u) + (1 - \gamma_n) \int \Gamma_{\theta}(u) d\Lambda_n(\theta) = \mathbf{H}_n(u). \end{aligned}$$

Similarly, the d.f. of a randomly chosen element of $\{T_1, \dots, T_n\}$,

$$\mathbf{M}_n(t) = \gamma_n \Phi(t) + (1 - \gamma_n) \int \Phi(t - \theta\sqrt{N}) d\Lambda_n(\theta),$$

and the empirical d.f. of $T_1, \dots, T_n$,

$$M_n = \gamma_n K_n + (1 - \gamma_n) L_n, \quad (2.12)$$

where $K_n(t) = n_0^{-1} \sum_{j=1}^{n_0} 1_{\{T'_j \leq t\}}$ and $L_n(t) = n_1^{-1} \sum_{j=1}^{n_1} 1_{\{T''_j \leq t\}}$, satisfy $EM_n(t) = \mathbf{M}_n$.

In addition to (2.1) and (2.2), another assumption that will be made throughout the paper is that

$$M_n(t) \rightarrow^P M(t) \quad (2.13)$$

uniformly in t . Since (2.1) holds if and only if $K_n \rightarrow^P \Phi$ uniformly (as can be seen on using the symmetry of the T'_j 's), it follows from (2.13) that $L_n \rightarrow^P L$ uniformly as well, where L is some d.f. And since¹

$${}^P \lim_{n \rightarrow \infty} L_n(t) = \lim_{n \rightarrow \infty} EL_n = \lim_{n \rightarrow \infty} \int \Phi(t - \theta\sqrt{N}) d\Lambda_n(\theta) = L(t),$$

and the integral here is a convolution, we can use characteristic functions and the continuity theorem to finally conclude that Λ_n converges weakly to some d.f. Λ ,

$$G(u) = \int \Gamma_\theta(u) d\Lambda(\theta), \quad (2.14)$$

and

$$L(t) = \int \Phi(t - \theta\sqrt{N}) d\Lambda(\theta). \quad (2.15)$$

While (2.15) suggests that if Φ is given and γ can be estimated then one may estimate Λ on the basis of a sample of test statistics—through a so-called deconvolution estimator—(2.14) shows that if Φ is given and an estimate of Λ is available then we have an approximate formula for G which depends in a simple way on the sample size N . One practical consequence of this result is thus that given a choice of Φ and estimates of γ and Λ obtained from an earlier, typically smaller experiment, one can get an estimate of G as a function of N (by inserting the estimate of Λ into (2.14)) and use it to compute approximately the sample size N required to achieve a given asymptotic average power with the Benjamini-Hochberg procedure in a future experiment.

In order to derive a formula for sample size calculation we assume for simplicity that Φ is the standard normal d.f.² Observe first that for the adaptive Benjamini-Hochberg method to have an asymptotic false discovery rate of δ and an asymptotic average power of $1 - \beta \in (0, 1)$ in the sense of (2.9) and (2.10) it is necessary that

$$\frac{\delta}{\gamma} \rho\left(\frac{\delta}{\gamma}, \gamma\right) = (1 - \beta) \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)}. \quad (2.16)$$

¹We denote by ${}^P \lim$ the limit in probability.

²The calculations, however, apply without change to any Φ for which $(\delta/\gamma)\rho(\delta/\gamma, \gamma)$ can be computed as indicated after (2.5).

Secondly we observe that Γ_θ is concave on $[0, 1]$ for each θ . (This can be checked, for example, by computing its derivative and writing it in terms of $\phi = \Phi'$ and Φ^{-1} —as incidentally we do at the beginning of Subsection 5.1—and then setting $t = \Phi^{-1}(1 - u/2)$ and showing that the resulting function of t has positive derivative for all $t > 0$.) Consequently, G is concave and, as mentioned in Section 2, the quantity $q\rho(q, \gamma)$ is the unique $u > 0$ such that $G(u) = u(1 - q\gamma)/[q(1 - \gamma)]$. Taking $q = \delta/\gamma$ and replacing u by the right-hand side of (2.16) in this equation, and using (2.11) and (2.14) to write out the expression of G , we see that *the sample size yielding an asymptotic power of at least $1 - \beta$ is obtained by rounding up the smallest number N such that*

$$\begin{aligned} 1 - \int \Phi \left(\Phi^{-1} \left(1 - \frac{(1 - \beta)}{2} \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)} \right) - \theta\sqrt{N} \right) d\Lambda(\theta) \\ + \int \Phi \left(-\Phi^{-1} \left(1 - \frac{(1 - \beta)}{2} \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)} \right) - \theta\sqrt{N} \right) d\Lambda(\theta) = 1 - \beta. \end{aligned} \quad (2.17)$$

Recall that the adaptive Benjamini-Hochberg method requires $\delta < \gamma$, which ensures that

$$\frac{\delta(1 - \gamma)}{\gamma(1 - \delta)} < 1 \quad \text{and} \quad \Phi^{-1} \left(1 - \frac{(1 - \beta)}{2} \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)} \right) > 0.$$

If Λ is not concentrated at zero then the left-hand side of (2.17) tends to one as $N \rightarrow \infty$, so it is always possible to find N yielding an arbitrarily large power. On the other hand, by what we have said at the end of the previous subsection, if δ is close enough to γ the left-hand side of (2.17) will be too close to 1 for a solution $N > 0$ to exist.

Example: If Φ is the standard normal d.f. and Λ is the d.f. of a normal r.v. with mean μ and variance τ^2 , then (2.17) is

$$1 - \Phi \left(\frac{t_\beta - \sqrt{N}\mu}{\sqrt{1 + N\tau^2}} \right) + \Phi \left(-\frac{t_\beta + \sqrt{N}\mu}{\sqrt{1 + N\tau^2}} \right) = 1 - \beta$$

where $t_\beta := \Phi^{-1} \left(1 - \frac{(1 - \beta)}{2} \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)} \right)$. Table 1 below provides an illustration of how sample size depends on the few parameters involved in this relatively simple case: it gives the values of sample size $\lceil N \rceil$ (the solutions N to this last equation rounded up to the nearest integer) corresponding to $1 - \beta = 0.9$ and several choices of γ , δ , τ , μ . The two top panels allow the comparison of sample sizes corresponding to ‘similar’ false discovery rates—namely $(1 - \gamma)/2$ —but different

γ	δ	τ	μ	$\lceil N \rceil$	γ	δ	τ	μ	$\lceil N \rceil$
0.9	0.05	0.5	0.5	<u>730</u>	0.8	0.10	0.5	0.5	471
			1	<u>63</u>				1	43
			0.5	384				0.5	248
		1	1	183			1	1	118
			0.5	<u>116</u>				0.5	75
0.8	0.05	0.5	1	<u>96</u>	0.8	0.025	0.5	1	62
			0.5	594				0.5	714
			1	53				1	62
		1	0.5	312			1	0.5	376
			1	149				1	179
0.7	0.05	2	0.5	94	0.7	0.025	2	0.5	114
			1	78				1	94

values of γ ; the second top panel and the two bottom panels illustrate how sample size varies with the false discovery rate when $\gamma = 0.8$.

As expected, within each panel the sample size increases as μ decreases; on the other hand, the effect of μ depends very much on τ : if τ —which quantifies the variability of the sequence $\{\theta_j\}$ —is small then the effect of μ is huge, but otherwise it is tiny (compare the underlined figures in the first panel). $\square$

In practice, equation (2.17) needs to be solved for given $1 - \beta$ and δ , and with γ and Λ replaced by appropriate estimates. The rest of the paper is mostly devoted to studying methods of estimating γ and Λ which apply under general conditions.

Remarks: i. The type of model we have considered here is widely used in analyses of microarray data as a means of detecting differences in gene expression levels on the basis of two-sample experiments; see for instance pp. 138-139 of McLachlan et al. (2004) and references therein.

One possible weakness of our method is the requirement that θ_j be a constant (rather than a r.v.), which in our first example amounts to the condition that $\sigma_{j,N}$ be close to σ_j . More realistic would be perhaps to let θ_j be random. On the other hand, intuition tells us that if, conditionally on θ_j , the statistic T_j has (approximately) d.f. $t \rightarrow \Phi(t - \theta_j\sqrt{N})$ then the method still works in an approximate sense, the only difference being that the d.f. Λ_n is interpreted as random (but is still assumed to converge to the non-random d.f. Λ); that being the case, the method is robust against a certain degree of randomness in the θ_j 's.

ii. Some authors measure the efficacy of the Benjamini-Hochberg method in terms of the proportion of incorrect non-rejections among non-rejections,

$$\Pi_{3,n}(q) = 1 - \frac{n_0 - S_n(q)}{(n - R_n(q)) \vee 1},$$

whose expected value we call *false non-discovery rate*; this concept was introduced by Genovese and Wasserman (2002). By Corollary 3.3 and a variant of Theorem 4.1 of Ferreira and Zwinderman (2006b), it can be shown that the analogue of (2.10) is

$$1 - \Pi_{3,n}(q_n) \xrightarrow{P} \frac{\gamma - \delta \rho\left(\frac{\delta}{\gamma}, \gamma\right)}{1 - \rho\left(\frac{\delta}{\gamma}, \gamma\right)}.$$

If in our calculations we require the limit in probability of $1 - \Pi_{3,n}(q_n)$, rather than asymptotic average power, to attain a given value of $1 - \beta \in (0, 1)$, we get

$$\begin{aligned} 1 - \int \Phi\left(\Phi^{-1}\left(1 - \frac{1}{2} \frac{\delta(1 - \beta - \gamma)}{\gamma(1 - \beta - \delta)}\right) - \theta\sqrt{N}\right) d\Lambda(\theta) \\ + \int \Phi\left(-\Phi^{-1}\left(1 - \frac{1}{2} \frac{\delta(1 - \beta - \gamma)}{\gamma(1 - \beta - \delta)}\right) - \theta\sqrt{N}\right) d\Lambda(\theta) = 1 - \beta. \end{aligned}$$

instead of (2.17).

iii. The calculation we propose does not involve the specification of a degree of discrepancy (e.g. difference in means) to be detected with a given probability, but rather the specification of the proportion of true positives one would like to attain. This may seem at odds with standard sample size calculations; however, the difference becomes very natural once we observe that while in standard testing procedures one rejects a hypothesis because a test statistic has (say) too large a value, the Benjamini-Hochberg method rejects subsets of hypotheses on the basis of the *relative magnitudes* of the test statistics; in fact, we have seen that in the Benjamini-Hochberg method the proportion of incorrect rejections out of all rejections is practically a constant, the false discovery rate, irrespective of the typical magnitude of the p-values. Another difference between our method and standard sample size calculations is that the power we wish to attain is actually a proportion—the proportion of false hypotheses detected—rather than the probability of detecting a given discrepancy. Again, the distinction is natural since the proportion of incorrect rejections out of all rejections is practically equal to the false discovery rate rather than to the probability of committing a type I error or an analogue thereof. $\square$

3 Basic assumptions and results

Throughout this section we let $\xi_1, \xi_2, \dots$ be a sequence of r.v.'s with d.f.'s $\Psi^{(1)}, \Psi^{(2)}, \dots$, and let Ψ be another d.f. Since for each x

$$\Psi_n(x) := \frac{1}{n} \sum_{j=1}^n 1_{\{\xi_j \leq x\}} \xrightarrow{P} \Psi(x) \quad (3.1)$$

if and only if $E\Psi_n(x) \rightarrow \Psi(x)$ and $E(\Psi_n(x)^2) \rightarrow \Psi(x)^2$, we see that (3.1) is equivalent to

$$\frac{1}{n} \sum_{j=1}^n \Psi^{(j)}(x) \rightarrow \Psi(x) \quad \text{and} \quad \frac{1}{n^2} \sum_{j,k=1}^n P(\xi_j \leq x, \xi_k \leq x) \rightarrow \Psi(x)^2. \quad (3.2)$$

Moreover, if the convergence in (3.2) is uniform in x then so is the convergence in (3.1), and conversely.

The two conditions in (3.2) are very important for our purposes because they represent the weakest assumptions under which the Benjamini-Hochberg method works in an asymptotic sense. Although the first is transparent and relatively easy to check for specific models, the second is more difficult to deal with because it involves the sum of bivariate d.f.'s, so it seems useful to introduce a stronger condition—similar to but weaker than the mixing conditions of ergodic theory—which can replace it. The condition we shall introduce below has an intuitive interpretation in terms of how many pairs of variables in a model are allowed to have a given degree of dependence and seems to be useful in model building; moreover, it serves to motivate a second, slightly stronger condition that we need to introduce in order to deal with characteristic functions.

For each x and $j = 1, \dots, n$ and each sequence $\{\alpha_{j,n}(x) : 1 \leq j \leq n\}$ of numbers in $[0, 1]$ write

$$d_{j,k,n}(x) = |P(\xi_j \leq x, \xi_k \leq x) - P(\xi_j \leq x)P(\xi_k \leq x)|, \quad j \neq k,$$

$$N_{j,n}(x) = \#\{k \leq n : d_{j,k,n}(x) > \alpha_{j,n}(x)\}.$$

We say that $\xi_1, \xi_2, \dots$ satisfy condition **D** if $\{\alpha_{j,n}(x) : 1 \leq j \leq n\}$ can be chosen in such a way that

$$r_{1,n} := \sup_x \max_{1 \leq j \leq n} \frac{N_{j,n}(x)}{n} \quad \text{and} \quad r_{2,n} := \sup_x \max_{1 \leq j \leq n} \alpha_{j,n}(x) \quad (3.3)$$

tend to zero as $n \rightarrow \infty$.

That this implies the second condition in (3.2) uniformly in x is seen from $\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{j \neq k}^n P(\xi_j \leq x)P(\xi_k \leq x) = \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{j=1}^n P(\xi_j \leq x) \right)^2$ and the inequalities

$$\begin{aligned} & \left| \frac{1}{n^2} \sum_{j \neq k}^n P(\xi_j \leq x, \xi_k \leq x) - \frac{1}{n^2} \sum_{j \neq k}^n P(\xi_j \leq x)P(\xi_k \leq x) \right| \leq \\ & \frac{1}{4n^2} \sum_{j=1}^n \sum_{k: d_{j,k,n}(x) > \alpha_{j,n}(x)} + \frac{1}{n^2} \sum_{j=1}^n \sum_{k: 0 < d_{j,k,n}(x) \leq \alpha_{j,n}(x)} \alpha_{j,n}(x) \leq \\ & \frac{1}{4n} \sum_{j=1}^n \frac{N_{j,n}(x)}{n} + \frac{1}{n} \sum_{j=1}^n \left(1 - \frac{N_{j,n}(x)}{n} \right) \alpha_{j,n}(x), \end{aligned}$$

the right-most side of which goes to zero if (3.3) is assumed.

Since no order relation between the indices of the variables is presupposed, the above definition seems particularly useful in applications to genomic data; specifically, (3.3) could be conceptually justified by assuming, for example, that for each gene there exist a small number—relative to the supposedly large number of genes considered—of genes related to it. Storey (2003) has provided some arguments as to why such type of dependence structure could be assumed to hold in analyses of gene expression data.

In order to prove the consistency of a certain deconvolution estimator we shall need to show later on that

$$\max_{s,t} |E[\{\varphi_{\Psi_n}(s) - \varphi_{E\Psi_n}(s)\}\{\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t)\}]| \xrightarrow{P} 0, \quad (3.4)$$

where φ_{Ψ_n} and $\varphi_{E\Psi_n}$ are the characteristic functions of Ψ_n and $E\Psi_n$. Although one might be tempted to use (3.1) alone to conclude (3.4) (by adapting the proof of Theorem 1 in Csörgö (1981), for example) this seems to require assumptions on the tails of Ψ , which would be somewhat artificial in our case; instead, we opt for strengthening condition **D**:

For each $j = 1, \dots, n$ and each sequence $\{\alpha'_{j,n} : 1 \leq j \leq n\}$ of numbers in $[0, 1]$, put

$$d'_{j,k,n} = \sup_{A,B \in \mathcal{B}(\mathbb{R})} |P(\xi_j \in A, \xi_k \in B) - P(\xi_j \in A)P(\xi_k \in B)|, \quad j \neq k,$$

$$N'_{j,n} = \#\{k \leq n : d'_{j,k,n} > \alpha'_{j,n}\}.$$

We say that $\xi_1, \xi_2, \dots$ satisfy condition **D'** if $\{\alpha'_{j,n} : 1 \leq j \leq n\}$ can be chosen so that

$$r'_{1,n} := \max_{1 \leq j \leq n} \frac{N'_{j,n}}{n} \rightarrow 0 \quad \text{and} \quad r'_{2,n} := \max_{1 \leq j \leq n} \alpha'_{j,n} \rightarrow 0. \quad (3.5)$$

That (3.5) implies (3.4) follows from the fact that

$$\left| E[e^{is\xi_j} e^{it\xi_k}] - E[e^{is\xi_j}] E[e^{it\xi_k}] \right| \leq 16 \alpha'_{j,n} \quad \text{whenever } j \neq k, \quad d'_{j,k,n} \leq \alpha'_{j,n},$$

whose proof is the same as that of Lemma 2 on p. 365 of Billingsley's (1995) book, and (with obvious notation) from the inequalities

$$\begin{aligned} \max_{s,t} \left| E[\{\varphi_{\Psi_n}(s) - \varphi_{E\Psi_n}(s)\} \{\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t)\}] \right| &\leq \\ \frac{1}{n} + \max_{s,t} \frac{1}{n^2} \sum_{j \neq k}^n \left| E[e^{is\xi_j} e^{it\xi_k}] - \varphi_{\Psi(j)}(s) \varphi_{\Psi(k)}(t) \right| &\leq \\ \frac{1}{n} + \frac{1}{n^2} \sum_{j=1}^n \sum_{k: d'_{j,k,n} > \alpha'_{j,n}} + \frac{16}{n^2} \sum_{j=1}^n \sum_{k: 0 < d'_{j,k,n} \leq \alpha'_{j,n}} \alpha'_{j,n} &\leq \quad (3.6) \\ \frac{1}{n} + \frac{1}{n} \sum_{j=1}^n \frac{N'_{j,n}}{n} + \frac{16}{n} \sum_{j=1}^n \left(1 - \frac{N'_{j,n}}{n} \right) \alpha'_{j,n} &\leq \\ \frac{1}{n} + r'_{1,n} + 16r'_{2,n}. \end{aligned}$$

In the following three subsections we prove certain convergence results for distribution and density functions based on $\mathbf{D}$ and $\mathbf{D}'$. Before proceeding, however, it may be of interest to consider two examples in which these conditions can be verified and compared with the more usual strong mixing condition; the second example is particularly illuminating as regards the gap between the second condition in (3.2) and condition $\mathbf{D}$.

Examples i. Suppose that for $j, k \in \mathbb{N}$ the r.v.'s ξ_j and ξ_k have densities $h_{1,j}$ and $h_{1,k}$ and a joint density $h_{2,j,k}$ such that for fixed j

$$\beta_{j,k} := \int |h_{2,j,k}(u, v) - h_{1,j}(u)h_{1,k}(v)| du dv \rightarrow 0$$

as $k \rightarrow \infty$. For instance, $h_{2,j,k}$ could be a bivariate normal density with arbitrary means, variances σ_j^2 and σ_k^2 bounded away from zero and correlation coefficient $\rho_{j,k}$ tending to zero as $k \rightarrow \infty$ for each j (as follows by a minor variant of Scheffé's theorem). Since then

$$\sup_{A, B \in \mathcal{B}(\mathbb{R})} |P(\xi_j \in A, \xi_k \in B) - P(\xi_j \in A)P(\xi_k \in B)| \leq \beta_{j,k} \rightarrow 0$$

as $k \rightarrow \infty$ for each j , we can fix a sequence $\alpha_{j,n} \equiv \alpha_{j,n}(x)$ (independent of x) tending to 0 for each j and restrict the number of variables ξ_k among $\{\xi_1, \dots, \xi_n\} \setminus \{\xi_j\}$ that satisfy $\beta_{j,k} > \alpha_{j,n}$ to $o(n)$ so that both $\mathbf{D}$ and $\mathbf{D}'$ are satisfied.

In the special case of normal variables, an application of the normal comparison lemma as given in Leadbetter *et al.* (1983) yields a bound for the term

$d_{j,k,n}(x)$ involved in **D**: $\sup_x |d_{j,k,n}(x)| \leq c |\sigma_j \sigma_k \rho_{jk}| (1 - \rho_{jk}^2)^{-1/2}$ where $c > 0$ is a constant. Given a covariance function, one can use this inequality to choose an $\alpha_n \equiv \alpha_{j,n}(x)$ (independent of x and j) yielding convenient bounds on $N_{j,n}(x)$, and hence on $r_{1,n}$. For a concrete calculation let $\{\xi_j\}$ be the so-called ‘fractional Brownian noise’ (the sequence of increments of fractional Brownian motion at integer times), a stationary sequence whose correlation function satisfies $\rho_{jk} \sim c' |j - k|^{-2(1-H)}$ as $|j - k| \rightarrow \infty$, where $H \in (0, 1)$ and $c' > 0$ are constants; if $\alpha_n \rightarrow 0$ and $n \alpha_n^{1/(2(1-H))} \rightarrow \infty$ then $N_{j,n}(x) \leq c'' \alpha_n^{-1/(2(1-H))} \forall x, j$ for some $c'' > 0$, and it follows that $r_{1,n} \rightarrow 0$ and that **D** and **D'** hold.

ii. Suppose $\xi_1, \xi_2, \dots$ is an irreducible Markov chain with countable state space and n -step transition probabilities $\{p_{xy}^{(n)}\}$. As is well-known, if the chain is aperiodic and has a stationary probability distribution $\{\pi_y\}$ then

$$\lim_{n \rightarrow \infty} p_{xy}^{(n)} = \pi_y = \lim_{n \rightarrow \infty} P(\xi_n = y) \quad \forall x, y, \quad (3.7)$$

which obviously implies (3.2). [In this case (3.1) is also true a.s. by the ergodic theorem for Markov chains.] In fact, (3.7) implies a property that is generally stronger: for each k ,

$$\beta_{j,n} := \sup_{A,B} |P(A \cap B) - P(A)P(B)| \rightarrow 0 \quad (3.8)$$

as $n \rightarrow \infty$, where the supremum is taken over sets A and B measurable with respect to $\{\xi_1, \dots, \xi_j\}$ and $\{\xi_{j+n}, \xi_{j+n+1}, \dots\}$, respectively. The implication follows from the fact that in the case of Markov chains the numbers $\beta_{j,n}$ in (3.8) can be computed by taking A measurable $\{\xi_j\}$ and B measurable $\{\xi_{n+j}\}$ (as indicated in Bradley (2005), Section 3) and then by applying the Scheffé and bounded convergence theorems. If the convergence is uniform in j —e.g. if the chain is stationary—then (3.8) is the strong mixing condition (e.g. Bradley (2005)).

To show that (3.8) in turn implies **D'** we observe that if $0 < a_n \rightarrow \infty$, $a_n/n \rightarrow 0$, and $\alpha_{j,n} = \max_{|k-j| \geq a_n} \beta_{j \wedge k, |k-j|}$, then

$$\begin{aligned} \alpha_{j,n} &\leq \max \left\{ \max_{1 \leq k < j \wedge j-k \geq a_n} \beta_{k, j-k}, \max_{k-j \geq a_n} \beta_{j, k-j} \right\} \\ &\leq \max \left\{ \max_{1 \leq k < j \wedge l \geq a_n} \beta_{k, l}, \max_{l \geq a_n} \beta_{j, l} \right\} \\ &\leq \max \left\{ \max_{l \geq a_n} \beta_{1, l}, \dots, \max_{l \geq a_n} \beta_{j-1, l}, \max_{l \geq a_n} \beta_{j, l} \right\} \rightarrow 0 \end{aligned}$$

and

$$\frac{N'_{j,n}}{n} = \frac{\#\{k \leq n : d'_{j,k,n} > \alpha_{j,n}\}}{n} \leq \frac{2a_n}{n} \rightarrow 0$$

for each j . More generally, this shows that $\mathbf{D}'$ is weaker than strong mixing.

Now suppose that the chain has period $d \geq 1$. From Chapter XV.9 of Feller (1968) we know that to each state x and each recurrent state y there corresponds an integer $b \in [0, d-1]$ such that

$$\lim_{n \rightarrow \infty} p_{xy}^{(nd+b)} = d \pi_y$$

where $\{\pi_y\}$ is a probability distribution. From this follows that if x and y are recurrent states then

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n P(\xi_j \leq x) &\rightarrow \sum_{x' \leq x} \pi_{x'} \quad \text{and} \\ \frac{1}{n^2} \sum_{j,k=1}^n P(\xi_j \leq x, \xi_k \leq y) &\rightarrow \sum_{x' \leq x} \pi_{x'} \sum_{y' \leq y} \pi_{y'}, \end{aligned}$$

This analogue of (3.7) involving Cesàro convergence is enough to imply (3.2) but not (3.8), since it does not generally imply strong mixing (Bradley (2005), Theorem 3.2). $\square$

3.1 Convergence of empirical d.f.'s

We first prove a result about the rate at which Ψ_n converges to Ψ under condition $\mathbf{D}$; it will provide us with conditions which guarantee that the Benjamini-Hochberg method works as described in Section 2 and will be used in practically everything that follows.

We shall obviously need to assume

$$r_{3,n} := \sup_x \left| \frac{1}{n} \sum_{j=1}^n \Psi^{(j)}(x) - \Psi(x) \right|^2 \rightarrow 0. \quad (3.9)$$

For convenience we write $r_{0,n} := 1/n$ and

$$C_n := \left(\log \log n \sum_{j=0}^3 r_{j,n} \right)^{-1/3} \quad (3.10)$$

Proposition 3.1. *Assume that $\xi_1, \xi_2, \dots$ is a sequence of r.v.'s satisfying (3.9) and condition **D**, and that Ψ is continuous and strictly increasing. Then*

$$C_n \sup_x |\Psi_n(x) - \Psi(x)| \xrightarrow{P} 0. \quad (3.11)$$

Proof. Since $\sup_x |\Psi_n(x) - \Psi(x)| = \sup_x |\Psi_n \circ \Psi^{-1}(x) - x|$ and $\Psi(\xi_1), \Psi(\xi_2), \dots$ also satisfy (3.9) and **D** as long as Ψ is continuous and strictly increasing in the union of the supports of $\xi_1, \xi_2, \dots$ (it need not be strictly increasing in $\mathbb{R}$), we may assume that Ψ is standard uniform.

As before, we have

$$\begin{aligned} E \left[\left(\Psi_n(x) - \frac{1}{n} \sum_{j=1}^n \Psi^{(j)}(x) \right)^2 \right] &= E \left[\frac{1}{n^2} \sum_{j=1}^n (1_{\{\xi_j \leq x\}} - \Psi^{(j)}(x))^2 \right] + \\ &E \left[\frac{1}{n^2} \sum_{j \neq k} (1_{\{\xi_j \leq x\}} - \Psi^{(j)}(x)) (1_{\{\xi_k \leq x\}} - \Psi^{(k)}(x)) \right] \leq \\ &\frac{1}{n} + \frac{1}{n^2} \sum_{j \neq k} [P(\xi_j \leq x, \xi_k \leq x) - P(\xi_j \leq x)P(\xi_k \leq x)] \leq \\ &\frac{1}{n} + \frac{1}{4n} \sum_{j=1}^n \frac{N_{j,n}(x)}{n} + \frac{1}{n} \sum_{j=1}^n \left(1 - \frac{N_{j,n}(x)}{n} \right) \alpha_{j,n} \leq r_{0,n} + r_{1,n} + r_{2,n}. \end{aligned}$$

Thus

$$\begin{aligned} E^{\frac{1}{2}}(|\Psi_n(x) - \Psi(x)|^2) &\leq E^{\frac{1}{2}} \left[\left(\Psi_n(x) - \frac{1}{n} \sum_{j=1}^n \Psi^{(j)}(x) \right)^2 \right] + r_{3,n}^{\frac{1}{2}} \\ &\leq (r_{0,n} + r_{1,n} + r_{2,n})^{\frac{1}{2}} + r_{3,n}^{\frac{1}{2}} \\ &\leq \sum_{j=0}^3 r_{j,n}^{\frac{1}{2}} \end{aligned} \quad (3.12)$$

for all x .

Let $k_n \in \mathbb{N}$ and put $x_{j,n} = j/k_n$, $j = 1, \dots, k_n$. By assumption, $\Psi(x_{j,n}) - \Psi(x_{j-1,n}) = (x_{j,n} - x_{j-1,n}) = 1/k_n$. If $x \in [x_{j-1,n}, x_{j,n}]$ then

$$\Psi_n(x_{j-1,n}) - \Psi(x_{j-1,n}) - \frac{1}{k_n} \leq \Psi_n(x) - \Psi(x) \leq \Psi_n(x_{j,n}) - \Psi(x_{j,n}) + \frac{1}{k_n}$$

whence

$$\sup_x |\Psi_n(x) - \Psi(x)| \leq \max_{j=0,1,\dots,k_n} |\Psi_n(x_{j,n}) - \Psi(x_{j,n})| + \frac{1}{k_n}.$$

Together with (3.12) this gives

$$\begin{aligned} P(a_n \sup_x |\Psi_n(x) - \Psi(x)| > \delta) &\leq \\ k_n \max_{j=1, \dots, k_n} P\left(|\Psi_n(x_{j-1,n}) - \Psi(x_{j-1,n})| > \frac{\delta - Ca_n/k_n}{a_n}\right) &\leq \\ \frac{a_n^2 k_n}{(\delta - Ca_n/k_n)^2} \max_{j=1, \dots, k_n} E\left[|\Psi_n(x_{j-1,n}) - \Psi(x_{j-1,n})|^2\right] &\leq \\ \left(\frac{a_n k_n^{1/2} \sum_{j=0}^3 r_{j,n}^{1/2}}{\delta - Ca_n/k_n}\right)^2 \end{aligned}$$

where $\{a_n\}$ is to be chosen. It follows that $a_n \sup_x |\Psi_n(x) - \Psi(x)| \xrightarrow{P} 0$ if

$$\frac{a_n}{k_n} = o(1), \quad \text{and} \quad k_n r_{j,n} a_n^2 = o(1) \quad \forall j,$$

which will be the case if $a_n = k_n(\log \log n)^{-1/3}$ and $k_n \propto \left(\sum_{j=0}^3 r_{j,n}\right)^{-\frac{1}{3}}$. $\square$

Remarks: i. When $\xi_1, \xi_2, \dots$ are i.i.d. we have $r_{j,n} = 0$ for $j = 1, 2, 3$, and then (3.11) is a very poor relative of the law of the iterated logarithm (namely $\limsup_{n \rightarrow \infty} (2n/\log \log n)^{1/2} \sup_x |\Psi_n(x) - \Psi(x)| \leq 1$ a.s.); our lack of depth reflects the generality we are aiming at, which allows no more than the computation of second moments.

ii. As mentioned at the beginning of this section, (3.11) also holds with

$$r_{4,n} := \sup_x \left| \frac{1}{n^2} \sum_{j \neq k}^n P(\xi_j \leq x, \xi_k \leq x) - P(\xi_j \leq x)P(\xi_k \leq x) \right|$$

in place of $r_{1,n} + r_{2,n}$ and with the requirement that $r_{3,n}, r_{4,n} \rightarrow 0$ in place of the condition that $\xi_1, \xi_2, \dots$ satisfy **D**. $\square$

3.2 Convergence of density estimators

Assume that the d.f. Ψ of (3.1) is concentrated on $[0, 1]$ and has a density ψ . One of our problems later on will consist of estimating $\psi(1-)$ through a non-parametric kernel density estimator as a means for estimating γ . Because standard Rosenblatt kernel estimators tend to be problematic when it comes to estimating compactly supported densities near the boundary points of their support (e.g. S. Chen (1999) and references therein), we consider using *beta kernel density estimators*. These were introduced by S. Chen (1999) and shown to converge to ψ in various ways in the i.i.d. case by Bouezmarni and Rolin (2003).

Here we shall prove a convergence result under the conditions of Proposition 3.1. There are other estimators known to obviate problems at boundary points which could also be used; see the recent approach of Langaas *et al.* (2005) to the problem of estimating γ .

The beta kernel sequence of estimators of ψ is defined for $x \in [0, 1]$ by

$$\hat{\psi}_n(x) = \frac{1}{n} \sum_{j=1}^n K(\xi_j, \alpha_n(x), \beta_n(x)) = \int K(y, \alpha_n(x), \beta_n(x)) d\Psi_n(y), \quad (3.13)$$

where $\alpha_n(x) = x/a_n + 1$, $\beta_n(x) = (1-x)/a_n + 1$, $\{a_n\}$ is a positive sequence converging to zero as $n \rightarrow \infty$, and $K(y, \alpha, \beta) = y^{\alpha-1}(1-y)^{\beta-1}/B(\alpha, \beta)$, $y \in [0, 1]$, $\alpha, \beta > 0$, is the beta density with parameters α and β . Recall the definition of C_n in (3.10).

Proposition 3.2. *Under the conditions of Proposition 3.1, if $a_n \rightarrow 0$ and $C_n a_n^{1/2} \rightarrow \infty$ then $\hat{\psi}_n(x) \xrightarrow{P} \psi(x)$ whenever ψ is continuous at $x \in (0, 1)$. If, moreover, ψ is continuous on $(0, 1)$, right-continuous at 0 and left-continuous at 1, and if $C_n a_n \rightarrow \infty$, then $\sup_{x \in [0, 1]} |\hat{\psi}_n(x) - \psi(x)| \xrightarrow{P} 0$.*

Proof. Denoting by $B_{x,n}$ a beta r.v. with parameters $x/a_n + 1$ and $(1-x)/a_n + 1$, we have

$$\begin{aligned} |\hat{\psi}_n(x) - \psi(x)| &\leq |\hat{\psi}_n(x) - E\hat{\psi}_n(x)| + |E\hat{\psi}_n(x) - \psi(x)| \\ &\leq \left| \int K(y, \alpha_n(x), \beta_n(x)) d(\Psi_n - E\Psi_n)(y) \right| + \\ &\quad \left| \int K(y, \alpha_n(x), \beta_n(x)) d(E\Psi_n - \Psi)(y) \right| + \\ &\quad \left| \int K(y, \alpha_n(x), \beta_n(x)) d\Psi(y) - \psi(x) \right| \quad (3.14) \\ &= \left| \int \Psi_n(y) - E\Psi_n(y) dK(y, \alpha_n(x), \beta_n(x)) \right| + \\ &\quad \left| \int E\Psi_n(y) - \Psi(y) dK(y, \alpha_n(x), \beta_n(x)) \right| + \\ &\quad |E\psi(B_{x,n}) - \psi(x)|. \end{aligned}$$

As $a_n \rightarrow 0$, $EB_{x,n} = (x + a_n)/(1 + 2a_n) \rightarrow x$ and $\text{Var}B_{x,n} = a_n(x + a_n)(1 - x + a_n)/[(1 + 2a_n)^2(1 + 3a_n)] \rightarrow 0$, whence $B_{x,n} \xrightarrow{P} x$ and $E\psi(B_{x,n}) \rightarrow \psi(x)$ whenever ψ is continuous at x . Furthermore, since $0 < B_{x,n} < 1$ a.s., we also have $E\psi(B_{0,n}) \rightarrow \psi(0+)$ and $E\psi(B_{1,n}) \rightarrow \psi(1-)$. If ψ is continuous on $(0, 1)$,

right-continuous at 0 and left-continuous at 1, then $E\psi(B_{x,n}) \rightarrow \psi(x)$ uniformly in x by Lemma 3.1 of Bouezmarni and Rolin (2003). It remains to treat the first two terms appearing after the equality sign in (3.14).

Under the assumptions of Proposition 3.1,

$$\begin{aligned} & \left| \int \Psi_n(y) - E\Psi_n(y) dK(y, \alpha_n(x), \beta_n(x)) \right| \leq \\ & \sup_y |\Psi_n(y) - \Psi(y)| V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) = \\ & C_n^{-1} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) o_P(1); \end{aligned}$$

and since

$$\begin{aligned} \left| \int E\Psi_n(y) - \Psi(y) dK(y, \alpha_n(x), \beta_n(x)) \right| & \leq r_{3,n}^{1/2} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) \\ & \leq C_n^{-1} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))), \end{aligned}$$

it is enough to show that $\{a_n\}$ can be chosen in such a way that

$$C_n^{-1} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) \quad (3.15)$$

remains bounded.

Noting that $y \rightarrow K(y, \alpha_n(x), \beta_n(x))$ increases on $[0, x]$ and decreases on $[x, 1]$ and using Stirling's formula, we see that for $x \in (0, 1)$

$$\begin{aligned} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) &= \int_0^x dK(y, \alpha_n(x), \beta_n(x)) - \int_x^1 dK(y, \alpha_n(x), \beta_n(x)) \\ &= 2 \frac{x^{x/a_n} (1-x)^{(1-x)/a_n}}{B(x/a_n + 1, (1-x)/a_n + 1)} \\ &\sim \sqrt{\frac{2}{\pi}} \frac{x^{-1/2} (1-x)^{-1/2}}{a_n^{1/2}} = O(a_n^{-1/2}). \end{aligned}$$

Thus (3.15) tends to 0 for each $x \in (0, 1)$ if $C_n a_n^{1/2} \rightarrow \infty$, which proves the first statement of the proposition.

On the other hand, we also have

$$\begin{aligned} V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) &= \int_0^1 |K'(t, \alpha_n(x), \beta_n(x))| dt \\ &= \frac{(a_n+1)}{a_n} \int_0^1 |K(t, \alpha_n(x) - 1, \beta_n(x)) - K(t, \alpha_n(x), \beta_n(x) - 1)| dt \leq 2 \frac{(a_n+1)}{a_n} \end{aligned}$$

for $0 < x < 1$, and

$$V_{[0,1]}(K(\cdot, \alpha_n(x), \beta_n(x))) = \left| \int dK(y, \alpha_n(x), \beta_n(x)) \right| = \frac{(a_n+1)}{a_n}$$

for $x = 0$ or $x = 1$. Thus (3.15) tends to 0 uniformly in x if $C_n a_n \rightarrow \infty$. $\square$

Remarks i. In the i.i.d. case it is enough to require $na_n/\log \log n \rightarrow \infty$, which is close to the usual (that of Rosenblatt's estimators) rate of decrease imposed on $\{a_n\}$, to conclude $\hat{\psi}_n(x) \rightarrow \psi(x)$ a.s. for each $x \in (0, 1)$, as follows from the law of the iterated logarithm on replacing C_n by $(n^{-1} \log \log n)^{-1/2}$ in (3.15). Both the pointwise convergence at the boundary points and uniform convergence require the more stringent rate $na_n^2/\log \log n \rightarrow \infty$.

ii. In the i.i.d. case Bouezmarni and Rolin (2003) have shown that, under some smoothness conditions on ψ , $E \int |\hat{\psi}_n - \psi| = O(n^{-2/5})$ when $a_n = O(n^{-2/5})$, similarly to what happens with standard kernel density estimates. $\square$

3.3 Convergence of deconvolution estimators

Now assume that the density of Ψ is given by

$$\psi(x) = \int \phi(x - y) d\Lambda(y), \quad (3.16)$$

where Λ is some d.f. and ϕ is the density of a d.f. Φ whose characteristic function φ_Φ has no zeros. The deconvolution problem is the problem of estimating Λ , or its density function $\lambda = \Lambda'$, on the basis of the sequence $\xi_1, \xi_2, \dots$ and the knowledge of ϕ ; it has been well studied in the literature (see for instance Fan (1991) and Masry (1993) and earlier references cited by these authors), but not under the conditions we are interested in. Here we shall prove the pointwise consistency, in probability, of an estimator λ_n of λ and of an associated estimator of Ψ , under the assumption that $\xi_1, \xi_2, \dots$ satisfy (3.1), condition **D'**, and

$$E\Psi_n(x) = \frac{1}{n} \sum_{j=1}^n \Psi^{(j)}(x) = \frac{1}{n} \sum_{j=1}^n \Phi(x - y_j) = \int \Phi(x - y) d\Lambda_n(y), \quad (3.17)$$

where $\Lambda_n(y) = \sum_{j=1}^n 1_{(-\infty, y]}(y_j)$ and $y_1, y_2, \dots$ are a given sequence of numbers. Together, (3.1), (3.16) and (3.17) imply that Λ_n converges weakly to Λ , and uniformly if Λ is continuous.

Observe that (3.17) is a very natural assumption in the context of the location models of Section 2.2.

A widely used sequence of estimators of $\lambda(x)$ is

$$\lambda_n(x) = \frac{1}{2\pi} \int e^{-itx} \frac{K^*(a_n t) \varphi_{\Psi_n}(t)}{\varphi_\Phi(t)} dt, \quad (3.18)$$

where K^* is a real Fourier transform vanishing outside $[-1, 1]$ of a density function K (the kernel) of bounded variation, φ_{Ψ_n} is the characteristic function of

Ψ_n , and $\{a_n\}$ is a positive sequence converging to zero.³ The rationale behind the definition of λ_n is that the kernel density estimator of ψ defined by $\psi_n(x) = a_n^{-1} \int K((x-y)/a_n) d\Psi_n(y)$ has Fourier transform $t \rightarrow K^*(a_n t) \varphi_{\Psi_n}(t)$, while by assumption the Fourier transform of ψ is $\varphi_{\Phi} \varphi_{\Lambda}$; equating the two, solving for φ_{Λ} and applying the inversion formula for Fourier transforms to the result yields (3.18).

As Fan (1991), we consider two cases: In the so-called ‘super-smooth’ case the characteristic function of Φ is assumed to decrease at most exponentially:

$$|\varphi_{\Phi}(t)| \geq c_0^{-1} e^{-c_1 |t|^{c_2}} |t|^{-c_3} \quad (3.19)$$

for all t such that $|t| \geq C$, where $C > 0$ is fixed, $c_0, c_1, c_2 > 0$ and $c_3 \in \mathbb{R}$ (this includes the normal distribution and is known to yield rather slow rates of convergence). In the ‘ordinary smooth case’, $|\varphi_{\Phi}|$ is bounded below by a polynomial:

$$|\varphi_{\Phi}(t)| \geq d_0^{-1} |t|^{-d_1} \quad (3.20)$$

for all t such that $|t| \geq C$, where $C > 0$ and $d_0, d_1 > 0$. For convenience we put

$$C'_n := \left(\frac{1}{n} + r'_{1,n} + 16r'_{2,n} \right)^{-1/2} \quad (3.21)$$

Proposition 3.3. *Suppose $\xi_1, \xi_2, \dots$ satisfy $\mathbf{D}'$, (3.1), (3.16) and (3.17). Under the above stated conditions on Φ and K we can choose $\{a_n\}$ such that $\lambda_n(x) \rightarrow^P \lambda(x)$ at every continuity point x of λ . Specifically, $\{a_n\}$ can be chosen as*

$$a_n = \left(\frac{c}{c_1} \log(C'_n) \right)^{-\frac{1}{c_2}} \quad \text{for some } c \in (0, 1) \quad (3.22)$$

if (3.19) holds, and as

$$a_n = C'_n \frac{d}{2d_1+1} \quad \text{for some } d \in (0, 1/2) \quad (3.23)$$

if (3.20) holds. Moreover, if λ is Lipschitz continuous and $\int |u|K(u)du < \infty$, we have in each case,

$$\int f(x) \lambda_n(x) dx \rightarrow^P \int f(x) \lambda(x) dx \quad (3.24)$$

for every continuous, compactly supported function f .

³For example, one may take $K^*(t) = (1 - |t|)1_{[-1,1]}(t)$, the Fourier transform of the density $K(x) = \pi^{-1}(1 - \cos x)x^{-2}$, or $K^*(t) = 2(1 - |t|)^3 1_{[1/2,1]}(|t|) + (1 - 6t^2 + 6|t|^3)1_{[-1/2,1/2]}(t)$, the Fourier transform of $K(x) = (3/(8\pi))(\sin(x/4)/(x/4))^4$ (Wand (1998)).

Proof. We write

$$\begin{aligned}
 2\pi |\lambda_n(x) - \lambda(x)| &\leq \left| \int e^{-itx} \frac{K^*(a_nt)}{\varphi_\Phi(t)} \{\varphi_{\Psi_n}(t) - \varphi_\Phi(t)\varphi_\Lambda(t)\} dt \right| + \\
 &\quad \left| \int e^{-itx} \{K^*(a_nt)\varphi_\Lambda(t) - \varphi_\Lambda(t)\} dt \right| \\
 &\leq \left| \int e^{-itx} \frac{K^*(a_nt)}{\varphi_\Phi(t)} \{\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t)\} dt \right| + \quad (3.25) \\
 &\quad \left| \int e^{-itx} \frac{K^*(a_nt)}{\varphi_\Phi(t)} \{\varphi_{E\Psi_n}(t) - \varphi_\Psi(t)\} dt \right| \\
 &\quad \left| \int e^{-itx} \{K^*(a_nt)\varphi_\Lambda(t) - \varphi_\Lambda(t)\} dt \right|
 \end{aligned}$$

and study the convergence to 0 of the three last terms in (3.25).

The third term satisfies

$$\begin{aligned}
 \left| \int e^{-itx} \{K^*(a_nt)\varphi_\Lambda(t) - \varphi_\Lambda(t)\} dt \right| &\propto \left| \frac{1}{a_n} \int K\left(\frac{x-y}{a_n}\right) \lambda(y) dy - \lambda(x) \right| \\
 &\leq \int K(u) |\lambda(x - a_n u) - \lambda(x)| du,
 \end{aligned}$$

and the last integral tends to 0 if λ is continuous at x and $a_n \rightarrow 0$. If, in addition, λ is Lipschitz continuous and K has first order moment, we have

$$\int K(u) |\lambda(x - a_n u) - \lambda(x)| du \leq a_n c \int |u| K(u) du \quad \forall x$$

for some $c > 0$, and can thus conclude that the convergence of the last term in (3.25) holds uniformly in x . As will be seen below, the other two terms in (3.25) can be bounded by quantities which do not depend on x ; consequently, under the additional assumptions on λ , we have $\lambda_n \rightarrow^P \lambda$ uniformly, and (3.24) follows.

Taking the expectation and applying (3.6) to the square of the first term we get

$$\begin{aligned}
 E \left| \int e^{-itx} \frac{K^*(a_nt)}{\varphi_\Phi(t)} \{\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t)\} dt \right|^2 &\leq \\
 \int \int e^{-i(s+t)x} \frac{K^*(a_ns)}{\varphi_\Phi(s)} \frac{K^*(a_nt)}{\varphi_\Phi(t)} E[(\varphi_{\Psi_n}(s) - \varphi_{E\Psi_n}(s))(\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t))] ds dt &\leq \\
 \left(\int \left| \frac{K^*(a_ns)}{\varphi_\Phi(s)} \right| ds \right)^2 \max_{s,t} E[\{\varphi_{\Psi_n}(s) - \varphi_{E\Psi_n}(s)\} \{\varphi_{\Psi_n}(t) - \varphi_{E\Psi_n}(t)\}] &\leq \\
 \left(\int \left| \frac{K^*(a_ns)}{\varphi_\Phi(s)} \right| ds \right)^2 \left(\frac{1}{n} + r'_{1,n} + 16r'_{2,n} \right). &
 \end{aligned}$$

In the super-smooth case,

$$\begin{aligned} \left| \int \frac{K^*(a_n s)}{\varphi_\Phi(s)} ds \right| &\leq \frac{1}{a_n} \int_{-1}^1 \left| \frac{K^*(t)}{\varphi_\Phi(t/a_n)} \right| dt \\ &\leq \frac{1}{a_n \min_{|t/a_n| \leq C} |\varphi_\Phi(t/a_n)|} \int_{-a_n C}^{a_n C} |K^*(t)| dt + \\ &\quad \frac{1}{a_n \min_{C \leq |t/a_n| \leq 1/a_n} |\varphi_\Phi(t/a_n)|} \int |K^*(t)| dt \\ &\leq O(1) + O\left(c_0 \exp\{c_1 a_n^{-c_2}\} \left(\frac{1}{a_n}\right)^{1+c_3}\right) \end{aligned}$$

for large n by the continuity of K^* and (3.19), so consistency requires us to choose $\{a_n\}$ such that

$$\exp\{c_1 a_n^{-c_2}\} \left(\frac{1}{a_n}\right)^{1+c_3} \left(\frac{1}{n} + r'_{1,n} + 16r'_{2,n}\right)^{1/2} \rightarrow 0,$$

which will be the case if (3.22) holds. Similarly, in the ordinary smooth case $\left| \int K^*(a_n s)/\varphi_\Phi(s) ds \right| \rightarrow 0$ if $\{a_n\}$ is chosen according to (3.23).

As to the remaining term, (3.17) gives

$$\begin{aligned} \left| \int e^{-itx} \frac{K^*(a_n t)}{\varphi_\Phi(t)} \{\varphi_{E\Psi_n}(t) - \varphi_\Psi(t)\} dt \right| &= \left| \int e^{-itx} K^*(a_n t) \{\varphi_{\Lambda_n}(t) - \varphi_\Lambda(t)\} dt \right| \propto \\ \left| \int K\left(\frac{x-y}{a_n}\right) d\Lambda_n(y) - \int K\left(\frac{x-y}{a_n}\right) d\Lambda(y) \right| &= \left| \int \Lambda_n(y) - \Lambda(y) dK\left(\frac{x-y}{a_n}\right) \right| = \\ \left| \int \Lambda_n(x - a_n u) - \Lambda(x - a_n u) dK(u) \right| &\leq \sup_x |\Lambda_n(x) - \Lambda(x)| V_{\mathbb{R}}(K) \rightarrow 0 \end{aligned}$$

as $a_n \rightarrow 0$ by the remark following (3.17) and the fact that K is of bounded variation. $\square$

Remarks: i. Suppose λ is bounded and uniformly continuous. Then the last term in (3.25) can be bounded by

$$\int K(u) |\lambda(x - a_n u) - \lambda(x)| du \leq \int K(u) \max_{x': |x' - x''| \leq a_n u} |\lambda(x') - \lambda(x'')| du$$

for all x ; and since the integral on the right here tends to zero by the bounded convergence theorem, we conclude, as in the proof, that $\lambda_n \xrightarrow{P} \lambda$ uniformly. Thus (3.24) also holds if λ is bounded and uniformly continuous. This variant will be used in Section 5.

ii. Our proof also gives (by means of the inequality $(x + y)^2 \leq 2(x^2 + y^2)$ applied twice to (3.25)) the mean square, pointwise consistency of λ_n , and in the i.i.d. case will yield the uniform rates of Fan (1991) if his extra assumptions about λ and K are added. Our statements are of course weaker than those of Masry (1993), but one must bear in mind the much greater simplicity of our assumptions and the wider range of applicability of Proposition 3.3. $\square$

4 Non-parametric estimators of γ and G

We now return to the setting of Section 2 and begin by studying a non-parametric estimator of γ and associated estimators of G and $g := G'$. The estimator of γ is based on the density estimators of Subsection 3.2 and, just like the estimator of Langaas *et al.* (2005), can be seen as a generalization—and, at least asymptotically, as an improved version—of Storey’s estimator (Storey (2002)).

One problem with ‘Storey-type’ estimators, and with our estimator in particular, is that they are consistent only if $G'(1-) = 0$, and, when biased, *overestimate* γ (see below). Since overestimating γ implies overestimating power and hence underestimating the sample size required in order to attain a given power, it is important to look for other estimators that are consistent under more general conditions. A consistent, ‘semi-parametric’ estimator of γ based on the location models of Subsection 2.2 will be introduced later on in Subsection 5.2. However, given the difficulty in finding fully non-parametric estimators in the case $G'(1-) > 0$, Storey-type estimators are always useful as first approximations and also as means of checking semi-parametric estimators whenever these are available; that is why we consider them here.

Suppose G has a density g with $g(1-) < \infty$, so that H has a density $h = \gamma + (1 - \gamma)g$ on $[0, 1]$. Let $\hat{h}_n$ be an estimator of h such that $\hat{h}_n(1) \rightarrow h(1-)$. For instance, $\hat{h}_n$ can be the beta kernel density estimator of Subsection 3.2:

$$\hat{h}_n(u) = \frac{1}{n} \sum_{j=1}^n \frac{X_j^{u/a_n} (1 - X_j)^{(1-u)/a_n}}{B(u/a_n + 1, (1-u)/a_n + 1)}, \quad u \in [0, 1], \quad (4.1)$$

where $a_n \rightarrow 0$ and $C_n a_n \rightarrow \infty$. Then

$$\hat{h}_n(1) \xrightarrow{P} h(1-) = \gamma + (1 - \gamma)g(1-),$$

whence

$$\hat{\gamma}_n := \hat{h}_n(1) \quad (4.2)$$

is a consistent estimator of γ if and only if $h(1-) = \gamma$, or equivalently if and only if $g(1-) = 0$. Together with Proposition 3.2, this shows

Theorem 4.1. Suppose $X_1, X_2, \dots$ satisfy the conditions of Proposition 3.1 with H in place of Ψ and C_n as the analogue of (3.10), and that $\sup_u |F_n(u) - F(u)| \rightarrow^P 0$, from which follows that $\sup_u |G_n(u) - G(u)| \rightarrow^P 0$ as well. Suppose further that $G' = g$ exists and is continuous on $[c, 1)$ for some $c \geq 0$ and left continuous at 1. If $a_n \rightarrow 0$ and $C_n a_n \rightarrow \infty$, then $\hat{\gamma}_n = \hat{h}_n(1)$, where $\hat{h}_n$ is given by (4.1), is a consistent estimator of γ if and only if $g(1-) = 0$; moreover, $\sup_{u \in [c, 1]} |\hat{h}_n(u) - h(u)| \rightarrow^P 0$, where $h = H'$ on $[c, 1)$.

The consideration of subintervals $[c, 1)$ of $[0, 1]$ accounts for a situation often suggested in practical applications: h is continuous on $(0, 1]$ but unbounded (see the histogram in Fig. 1 of Storey and Tibshirani (2003) or Fig. 5.2 of MacLachlan et al. (2004)); in this case we actually have $h_n(0) \rightarrow^P \infty$ by Theorem 4.1 of Bouezmarni and Rolin (2003).

Besides γ , in applications one also wants to estimate G and g . Clearly, with $\hat{\gamma}_n$ given by (4.2) the $\hat{G}_n$ of (2.8) is consistent if and only if $g(1-) = 0$. Similarly, Theorem 4.1 implies that if $g(1-) = 0$ and h is continuous on $[c, 1]$ then $\sup_{u \in [c, 1]} |\hat{g}_n(u) - g(u)| \rightarrow^P 0$, where $\hat{g}_n = (\hat{h}_n - \hat{\gamma}_n)/(1 - \hat{\gamma}_n)$.

As always, from an asymptotic point of view there are plenty of safe choices of a_n , but in practice a less than perfect choice of smoothing parameter may cause too much bias or too much variance in h_n . Examining the graphs of $\hat{G}_n$ and $\hat{g}_n$ and the associated values of $\hat{\gamma}_n$ that result from different choices of a_n may serve as a guide to the choice of the smoothing parameter. Indeed, $\hat{G}_n$ is not a d.f. and $\hat{g}_n$ is not a density function ($\hat{G}_n$ will typically decrease over small intervals, and $\hat{g}_n$ need not integrate to 1), and by varying $\hat{\gamma}_n$ (through a_n) somewhat one can assess the plausibility of different estimates of γ with reference to the ‘regularity’ of $\hat{G}_n$ and $\hat{g}_n$.

As to the bias, it is clear that, asymptotically, $\hat{\gamma}_n$ overestimates γ by $(1 - \gamma)g(1-)$. The variant of Storey’s estimator proposed by Ferreira and Zwinderman (2006b) is

$$\tilde{\gamma}_n(u) := \min_{0 \leq t \leq u} \frac{1 - H_n(t)}{1 - t}, \quad (4.3)$$

where $u \in (0, 1)$; asymptotically, this exceeds γ by

$$(1 - \gamma)\kappa(u), \quad \text{where} \quad \kappa(u) = \min_{0 \leq t \leq u} \frac{1 - G(t)}{1 - t}.$$

Thus the asymptotic bias of $\hat{\gamma}_n$ is smaller than that of $\tilde{\gamma}_n(u)$ if and only if

$$g(1-) < \kappa(u). \quad (4.4)$$

Although (4.4) does not generally hold, it does if for instance g is decreasing, because in that case $(1 - G(t))/(1 - t)$ decreases in t (the derivative of this function is negative if and only if $1 - G(t) < g(t)(1 - t)$). Consequently, $\hat{\gamma}_n$ has, asymptotically, a smaller bias than $\tilde{\gamma}_n(u)$ if G has a decreasing density. In this sense, $\hat{\gamma}_n$ represents an improvement upon Storey's estimator.⁴

Remark: Storey's estimator and $\tilde{\gamma}_n(u)$ can be seen as examples of (4.2) with h_n as the histogram estimator with a 'bin width' of $1 - u$. Storey (2002) also proposed a method for 'optimally' choosing u , which can therefore be seen as a method of choosing the bin width in the histogram estimate of h . $\square$

Though apparently realistic in many applications, the condition $g(1-) = 0$ does not generally hold in the case of the location models of Subsection 2.1. It is therefore of some interest to know something about the magnitude of the bias of $\hat{\gamma}_n$ in that case. In the next subsection we restrict ourselves to location models in which Φ is standard normal.

4.1 The asymptotic bias of $\hat{\gamma}_n$ in location models when Φ is standard normal

Writing $\phi = \Phi'$ and assuming that $\phi(x) > 0$ for all $x \in (0, \Phi^{-1}(1))$, we have by (2.14) that

$$\begin{aligned} g(u) &= \int \frac{d}{du} \Gamma_\theta(u) d\Lambda(\theta) \\ &= \frac{1}{2} \int \frac{\phi\left(\Phi^{-1}(1 - u/2) - \theta\sqrt{N}\right) + \phi\left(-\Phi^{-1}(1 - u/2) - \theta\sqrt{N}\right)}{\phi(\Phi^{-1}(1 - u/2))} d\Lambda(\theta) \end{aligned}$$

hence

$$g(1-) = \int \frac{\phi(-\theta\sqrt{N})}{\phi(0)} d\Lambda(\theta) > 0 \quad (4.5)$$

unless Λ and Φ have no common points of increase. As observed in Subsection 2.2, if Φ is the standard normal d.f. then G is concave, and it follows that $\tilde{\gamma}_n$ has a bigger asymptotic bias than $\hat{\gamma}_n$.

To see how large the bias of $\hat{\gamma}_n$ typically is suppose first that Λ is the d.f. of a normal r.v. Θ with mean μ and variance τ^2 , as in the example towards the

⁴What we have been calling Storey's estimator is $\tilde{\gamma}_n(u) := (1 - H_n(u))/(1 - u)$, where $0 < u < 1$; asymptotically, its bias is at least as large as that of $\tilde{\gamma}_n(u)$.

end of Subsection 2.2: a simple calculation yields

$$g(1-) = \sqrt{\frac{1}{1+\tau^2 N}} e^{-\frac{1}{2} \frac{\mu^2}{1/N+\tau^2}} \leq \frac{1}{\tau\sqrt{N}} e^{-\frac{1}{2} \frac{\mu^2}{1/N+\tau^2}} \sim \frac{1}{\tau\sqrt{N}} e^{-\frac{1}{2} \frac{\mu^2}{\tau^2}},$$

which shows that the bias of $\hat{\gamma}_n$ decreases with $\tau\sqrt{N}$ and with $|\mu|/\tau$ (the reciprocal of the coefficient of variation of Θ), and that the bias of $\hat{\gamma}_n$ is of order $1/\sqrt{N}$ as $N \rightarrow \infty$.

More generally, if Λ has a density λ then

$$g(1-) = \int e^{-\frac{1}{2}(\theta\sqrt{N})^2} \lambda(\theta) d\theta = \sqrt{\frac{2\pi}{N}} f_{\Theta+N(0,1/N)}(0),$$

where $f_{\Theta+N(0,1/N)}$ is the density of the sum of Θ and a normal random variable with mean zero and variance $1/N$, which implies that the bias of $\hat{\gamma}_n$ decreases as the mass of Λ moves away from zero. If in addition λ is bounded in a neighbourhood of the origin, it follows that

$$g(1-) \sim \lambda(0) \sqrt{\frac{2\pi}{N}} \quad \text{as } N \rightarrow \infty,$$

which shows that asymptotically the bias of $\hat{\gamma}_n$ is of order $1/\sqrt{N}$ and decreases with the amount of mass Λ puts near zero.

5 Main results for location models

We have seen in Subsection 2.2 that one way of computing approximately the sample size required to achieve an asymptotic power of $1-\beta$ with the Benjamini-Hochberg method given an asymptotic false discovery rate of δ is to solve equation (2.17) for N with γ and Λ replaced by estimates from an earlier study. The choice of Φ is a separate problem, and non-parametric estimators of γ have already been discussed. Here we will show how Λ can be estimated if a consistent estimator of γ is available, and then introduce a consistent estimator of γ . Our setting is that of Subsection 2.2, and we use the result of Subsection 3.3; we assume that Λ has a density λ , and that Φ and the kernel K used below satisfy the conditions stated in Proposition 3.3.

For notational simplicity we take $N = 1$; in order to get an estimator of $\lambda(\theta)$ for general N we need only multiply the right-hand side of (5.3) below by $\sqrt{N}$ and replace the θ in the integrand by $\theta\sqrt{N}$.

We begin by showing that

$$\lambda_n(\theta) := \frac{1}{2\pi} \int e^{-it\theta} \frac{K^*(a_n t) \varphi_{L_n}(t)}{\varphi_{\Phi}(t)} dt, \quad (5.1)$$

where

$$L_n(t) := \frac{M_n(t) - \gamma_n \Phi(t)}{1 - \gamma_n} \quad (5.2)$$

and $\{a_n\}$ is a certain positive sequence tending to 0, converges to $\lambda(\theta)$ whenever λ is continuous at θ .

We know from Subsection 2.2 that

$$EL_n(t) = \frac{EM_n(t) - \gamma_n \Phi(t)}{1 - \gamma_n} = \int \Phi(t - \theta) d\Lambda_n(\theta),$$

$$L_n(t) \xrightarrow{P} L(t) = \int \Phi(t - \theta) d\Lambda(\theta) \quad \text{and} \quad l(t) = \int \phi(t - \theta) d\Lambda(\theta),$$

as long as L has a density l , Λ_n converges to Λ , and $T_1, T_2, \dots$ satisfy condition **D'** (so that the conclusion of Proposition 3.1 applies to M_n). Thus—see (3.16) and (3.17)—we are practically in the situation of Subsection 3.3, the only difference being that L_n is not an empirical d.f. However, to show that Proposition 3.3 applies with $\Psi_n = L_n$ and $\Psi = L$ (with an appropriate choice of $\{a_n\}$) we need only to check that the three last terms in (3.25) go to zero in this case as well. The third term is independent of Ψ_n . The first term involves the difference

$$\varphi_{\Psi_n}(s) - \varphi_{E\Psi_n}(s) \equiv \varphi_{L_n}(s) - \varphi_{EL_n}(s) = \frac{\varphi_{M_n}(s) - \varphi_{EM_n}(s)}{1 - \gamma_n};$$

since by our assumptions on $T_1, T_2, \dots$ (3.6) applies to the numerator of the right-hand side here, the bound on the first term is exactly as before except for the division by $1 - \gamma_n$. Finally, the middle term involves the difference

$$\varphi_{E\Psi_n}(t) - \varphi_{\Psi}(t) \equiv \varphi_{EL_n}(t) - \varphi_L(t) = \varphi_{\Phi}(t) (\varphi_{\Lambda_n}(t) - \varphi_{\Lambda}(t)),$$

hence it tends to zero exactly as before.

This proves $\lambda_n \xrightarrow{P} \lambda$ at continuity points of λ . In order to define an estimator of Λ we now replace γ_n by a consistent estimator $\hat{\gamma}_n$ in (5.2) and L_n by $\hat{L}_n$ in (5.1):

$$\hat{\lambda}_n(\theta) := \frac{1}{2\pi} \int e^{-it\theta} \frac{K^*(a_n t) \varphi_{\hat{L}_n}(t)}{\varphi_{\Phi}(t)} dt, \quad (5.3)$$

where

$$\hat{L}_n(t) := \frac{M_n(t) - \hat{\gamma}_n \Phi(t)}{1 - \hat{\gamma}_n}. \quad (5.4)$$

Since

$$\begin{aligned}\hat{\lambda}_n(\theta) - \lambda_n(\theta) &= \left(\frac{1}{1 - \hat{\gamma}_n} - \frac{1}{1 - \gamma_n} \right) \frac{1}{2\pi} \int e^{-it\theta} \frac{K^*(a_nt) \varphi_{M_n}(t)}{\varphi_\Phi(t)} dt - \\ &\quad \left(\frac{\hat{\gamma}_n}{1 - \hat{\gamma}_n} - \frac{\gamma_n}{1 - \gamma_n} \right) K\left(\frac{\theta}{a_n}\right),\end{aligned}$$

we have $\hat{\lambda}_n - \lambda_n \rightarrow^P 0$ whenever $\hat{\gamma}_n - \gamma_n \rightarrow^P 0$, i.e., whenever $\hat{\gamma}_n$ is consistent. This proves

Theorem 5.1. *Suppose $T_1, T_2, \dots$ satisfy condition **D'**, so that in particular the conclusion of Proposition 3.1 holds with C'_n in place of C_n and the M_n of (2.12) in place of Ψ_n , and that the Λ of (2.14)-(2.15) has a density λ . Let $\{a_n\}$ satisfy (3.22) or (3.23) according to whether φ_Φ satisfies (3.19) or (3.20). If $\hat{\gamma}_n$ is a consistent estimator of γ , then the estimator of $\hat{\lambda}_n(\theta)$ defined by (5.3) and (5.4) is consistent; if moreover λ is continuous and its support is contained in a bounded interval I , then*

$$\hat{G}_n(u) := \int_I \Gamma_\theta(u) \hat{\lambda}_n(\theta) d\theta \rightarrow^P \int \Gamma_\theta(u) \lambda(\theta) d\theta = G(u)$$

uniformly in u .

Observe that it is the last statement (which follows by a special case of Remark i. to Proposition 3.3) that allows us to estimate the asymptotic average power as a function of N ; recall the remarks following (2.14) and the reasoning that led to (2.17).

The requirement that λ have compact support, which derives from the restrictions imposed in Proposition 3.3, is somewhat unpleasant; in practice, one has to choose I and hope that $\int_I \Gamma_\theta(u) \lambda(\theta) d\theta$ is sufficiently close to $G(u)$.

5.1 A consistent estimator of γ

Let us write $\|f\|_p = (\int |f|^p d\nu)^{1/p}$ for $f \in L^p(\nu)$, where ν is a probability measure on $\mathbb{R}$ and $p \geq 1$. To define a consistent estimator of γ we put

$$D_n(\gamma') = \|\hat{L}_n(\cdot, \gamma') - \check{L}_n(\cdot, \gamma')\|_p, \quad \gamma' \in [0, 1], \quad (5.5)$$

where

$$\hat{L}_n(t, \gamma') = \frac{M_n(t) - \gamma' \Phi(t)}{1 - \gamma'},$$

$$\hat{\lambda}_n(\theta, \gamma') := \frac{1}{2\pi} \int e^{-it\theta} \frac{K^*(a_n t) \varphi_{\hat{L}_n(\cdot, \gamma')}(t)}{\varphi_{\Phi}(t)} dt,$$

$$\check{L}_n(t, \gamma') = \int_I \Phi(t - \theta) \hat{\lambda}_n(\theta, \gamma') d\theta,$$

and I is the bounded interval of Theorem 5.1. Then we take $\hat{\gamma}_n$ to be one value of $\gamma' \in [0, 1]$ which minimizes $D_n(\gamma')$; it will be shown that, asymptotically, such γ' is unique.

We can avoid indeterminacies in the calculation of the norm in the case where $\gamma' = 1$ by, for example, interpreting $\check{L}_n(t, \gamma')$ as $\min\{\check{L}_n(t, \gamma'), 1\}$, say, so that $D_n(1) = \lim_{\gamma' \uparrow 1} \|M_n - \gamma' \Phi - (1 - \gamma')\|_p / (1 - \gamma') = \infty$.

Theorem 5.2. *Suppose $T_1, T_2, \dots$ satisfy condition $\mathbf{D}'$, so that in particular the conclusion of Proposition 3.1 holds with C'_n in place of C_n and the M_n of (2.12) in place of Ψ_n , and that the Λ of (2.14)-(2.15) has a continuous density λ with a bounded interval I as support. Let $\{a_n\}$ satisfy (3.22) or (3.23) according to whether φ_{Φ} satisfies (3.19) or (3.20). If $\hat{\gamma}_n$ is such that $D_n(\hat{\gamma}_n) = \min_{\gamma' \in [0, 1]} D_n(\gamma')$, with D_n defined by (5.5), then $\hat{\gamma}_n \xrightarrow{P} \gamma$.*

Proof. Because, as in the beginning of this section,

$$\hat{L}_n(t, \gamma_n) \xrightarrow{P} \frac{M(t) - \gamma \Phi(t)}{1 - \gamma} = L(t),$$

$$\hat{\lambda}_n(\theta, \gamma_n) \xrightarrow{P} \lambda(\theta) \quad \text{and} \quad \check{L}_n(t, \gamma_n) \xrightarrow{P} L(t),$$

we have $D_n(\gamma_n) \xrightarrow{P} 0$. Thus, by definition of $\hat{\gamma}_n$, $D_n(\hat{\gamma}_n) \leq D_n(\gamma_n) \xrightarrow{P} 0$.

As just pointed out, with the right interpretation of $\check{L}_n(t, \gamma')$ the case ${}^P \limsup_n \gamma_n = 1$ cannot arise.

Suppose $\hat{\gamma}_n \not\xrightarrow{P} \gamma$. Since $\hat{\gamma}_n \in [0, 1]$, there exists a subsequence $\{n_k\}$ such that $\hat{\gamma}_{n_k} \rightarrow \hat{\gamma} \neq \gamma$ a.s. Then, a.s.,

$$\hat{L}_{n_k}(t, \hat{\gamma}_{n_k}) = \frac{M_{n_k}(t) - \hat{\gamma}_{n_k} \Phi(t)}{1 - \hat{\gamma}_{n_k}} \rightarrow \frac{M(t) - \hat{\gamma} \Phi(t)}{1 - \hat{\gamma}},$$

$$\begin{aligned} \hat{\lambda}_{n_k}(\theta, \hat{\gamma}_{n_k}) &= \frac{1}{2\pi} \int e^{-it\theta} K^*(a_n t) \frac{\varphi_{M_{n_k}}(t) - \hat{\gamma}_{n_k} \varphi_{\Phi}(t)}{(1 - \hat{\gamma}_{n_k}) \varphi_{\Phi}(t)} dt \\ &= \frac{1}{2\pi} \int e^{-it\theta} K^*(a_n t) \frac{\varphi_{M_{n_k}}(t) - \gamma_{n_k} \varphi_{\Phi}(t)}{(1 - \gamma_{n_k}) \varphi_{\Phi}(t)} dt \frac{(1 - \gamma_{n_k})}{(1 - \hat{\gamma}_{n_k})} + \\ &\quad \frac{1}{2\pi} \int e^{-it\theta} K^*(a_n t) \frac{\varphi_{\Phi}(t)(\gamma_{n_k} - \hat{\gamma}_{n_k})}{(1 - \hat{\gamma}_{n_k}) \varphi_{\Phi}(t)} dt \end{aligned}$$

$$\begin{aligned} &\rightarrow \lambda(\theta) \frac{(1-\gamma)}{(1-\hat{\gamma})} + \lim_{k \rightarrow \infty} K\left(\frac{\theta}{a_{n_k}}\right) \frac{(\gamma_{n_k} - \hat{\gamma}_{n_k})}{(1-\hat{\gamma}_{n_k})} \\ &= \lambda(\theta) \frac{(1-\gamma)}{(1-\hat{\gamma})} \end{aligned}$$

for all $\theta \neq 0$ (because $K(x)$ as $|x| \rightarrow \infty$), and

$$\check{L}_{n_k}(t, \gamma_{n_k}) = \int_I \Phi(t - \theta) \hat{\lambda}_{n_k}(\theta, \gamma_{n_k}) d\theta \rightarrow L(t) \frac{(1-\gamma)}{(1-\hat{\gamma})}.$$

But then, a.s.,

$$D_{n_k}(\hat{\gamma}_{n_k}) \rightarrow \frac{1}{(1-\hat{\gamma})} \|M - \hat{\gamma}\Phi - (1-\gamma)L\|_p = \frac{|\gamma - \hat{\gamma}|}{(1-\hat{\gamma})} \|\Phi\|_p > 0,$$

contradicting $D_n(\hat{\gamma}_n) \xrightarrow{P} 0$. □

Of course, if the estimator $\hat{\gamma}_n$ used in Theorem 5.1 coincides with the estimator of Theorem 5.2, then the corresponding estimators of λ coincide as well; but in Theorem 5.1 any other consistent estimator will do.

6 Discussion: practical implementation and interpretation

The application of the method to sample size determination on the basis of data from a microarray array experiment is illustrated in a companion ‘applied’ article by the authors (Ferreira and Zwinderman (2006a)). Here we shall discuss some of the issues involved in a generic application and the interpretation of γ and λ in practice.

In the first place, Φ is assumed known, which in particular determines the computation of the p-values. In some cases Φ is assumed to be the Student or the normal distribution, but as the work of Chen, Klebanov and Yakovlev (2005) suggests these need not always be appropriate choices. In other cases the p-values can be computed using resampling methods (e.g. Pollard and van der Laan (2005)), which may turn out to be useful in specifying Φ . Whether or not a ‘fully nonparametric’ method of sample size calculation that dispenses the knowledge of Φ is possible remains an open question to us.

Having opted for a given Φ , an estimate of γ may be obtained from the estimator $\hat{\gamma}_n$ of Subsection 5.2 and eventually compared with estimates obtained from the non-parametric estimator of Section 4. Then the $\hat{\lambda}_n$ of (5.3) and the

$\hat{G}_n$ of Theorem 5.1 can be computed by numerical integration, and to carry out a sample size determination one replaces $d\Lambda(\theta)$ by $\hat{\lambda}_n(\theta)d\theta$ in equation (2.17) (which—see (2.14)—is really an equation involving $\hat{G}_n$) and solves (approximately) the latter for N .

We have assumed throughout that γ is a number between 0 and 1. In practice this implies that (say) the number of genes engaged in a given physiological process is very large. This assumption seems to be correct in studies of complex traits diseases (e.g. heart disease, diabetes); to mention just one example, the data analyzed by Allison et al. (2002) indicates that almost 30% of genes considered are differentially expressive in the two conditions studied.

Finally, two words about λ , the limiting density of ‘effect sizes’. Our method does not explicitly account for the many sources of variation present in microarray experiments, and this may seem paradoxical. The explanation, however, is very simple: all the experimental conditions and sources of variation pertaining to the pilot experiment are embodied by λ (or by its estimate), and in order to carry out a sample size calculation one assumes that the same λ will capture the conditions of any another experiment. Thus, in our approach it is implicitly assumed that the future experiment for which the sample size is being calculated will follow the standards set by the pilot experiment (hence in particular use the same technology, the same laboratory techniques, etc.). Since the idea of experiments ‘performed under identical conditions’ is one of the basic paradigms of applied statistics, there is hardly anything new here; but this interpretation of λ can be used to check the consistency of two or more microarray experiments addressing the same biological question and possibly performed at different laboratories and/or based on different technologies.

Secondly, the existence of Λ follows from assumptions (2.1), (2.2) and (2.13). To the numbers $\theta_1, \theta_2, \dots$ one associates a d.f. (in the sense of Analysis: a non-decreasing, right-continuous function, which in this case happens to be a probability d.f.) Λ_n , and under those assumptions we have $\Lambda_n(x) \rightarrow \Lambda(x)$ at each continuity point x of Λ . Although Λ is a d.f., the effect sizes $\theta_1, \theta_2, \dots$ need not be regarded as random; in fact, insofar Λ is a deterministic function, *as a set* they remain essentially constant from experiment to experiment.

References

- Allison, D., Gadbury, G., Heo, M., Fernandez, J., Lee, C., Prolla, T., et al. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*(39), 1-20.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*(57), 289-300.
- Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*(25), 60-83.
- Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley.
- Bouezmarni, T., and Rolin, J. (2003). Consistency of beta kernel density function estimators. *Canadian Journal of Statistics*, 31, 89-98.
- Bradley, R. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2, 107-144.
- Chen, L., Klebanov, L., and Yakovlev, A. (2005). Normality of gene expression revisited. *Technical Report*. (Department of Biostatistics and Computational Biology, University of Rochester Medical Center, USA)
- Chen, S. (1999). Beta kernel estimators for density functions. *Computational Statistics and Data Analysis*, 31, 131-145.
- Csörgö, S. (1981). Limit behaviour of the empirical characteristic function. *The Annals of Statistics*, 9(1), 130-144.
- Fan, J. (1991). On the optimal rates of convergence of nonparametric deconvolution estimators. *The Annals of Statistics*, 19(3), 1257-1272.
- Feller, W. (1968). *An Introduction to Probability Theory and its Applications* (3 ed., Vol. I). Wiley.
- Ferreira, J., and Zwinderman, A. (2006a). Approximate sample size calculations with microarray data: an illustration. *Statistical Applications in Genetics and Molecular Biology*. (In press.)
- Ferreira, J., and Zwinderman, A. (2006b). On the Benjamini-Hochberg method. *The Annals of Statistics*, Vol. 34(4).
- Genovese, C., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Stat. Soc. Ser. B*(64), 499-517.

- Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Stat. Soc. Ser. B*, 67(4), 555-572.
- Leadbetter, M., Lindgren, G., and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag.
- Lee, M., and Whitmore, G. (2002). Power and sample size for microarray studies. *Statistics in Medicine*(21), 3543-3570.
- Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Processes and their Applications*, 47, 53-74.
- McLachlan, G., Do, K.-A., and Ambroise, C. (2004). *Analysing Microarray Gene Expression Data*. Wiley.
- Muller, P., Parmigiani, G., Robert, C., and Rosseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Stat. Assoc.*, 99(468), 990-1001.
- Pan, W., Lin, J., and Le, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology*, 3(5).
- Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rates, sensitivity and sample size calculation for microarray studies. *Bioinformatics*, 21(13), 3017-3024.
- Pollard, K., and van der Laan, M. (2005). Resampling-based multiple testing with asymptotic strong control of type I error. (Preprint, Division of Biostatistics, University of California, Berkeley, U.S.A.)
- Storey, J. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B*(64), 479-498.
- Storey, J. (2003). Comment on ‘Resampling-based multiple testing for DNA microarray data analysis’, by Y. Ge, S. Dudoit and T. Speed. *Test*(12), 52-60.
- Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Stat. Soc. Ser. B*(66), 187-205.
- Storey, J., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100(16), 9440-9445.

- Tibshirani, R. (2005). *A simple method for assessing sample sizes in microarray experiments* (Technical report). Stanford University. (Available at <http://www-stat.stanford.edu/tibs/research.html>)
- Wand, M. (1998). Finite sample performance of deconvolving density estimators. *Statistics and Probability Letters*(37), 131-139.