Sequential Design for Microarray Experiments

Gilles DURRIEU and Laurent BRIOLLAIS

A critical aspect in the design of microarray studies is the determination of the sample size necessary to declare genes differentially expressed across different experimental conditions. In this article, we propose a sequential approach where the decision to stop the experiment depends on the accumulated microarray data. The study could stop whenever sufficient data have been accumulated to identify gene expression changes across several experimental conditions. The gene expression response is modeled by a robust linear regression model. We then construct a sequential confidence interval for the intercept of this model, which represents the median gene expression at a given experimental condition. We derive the stopping rule of the experiment for both continuous and discrete sequential approaches and give the asymptotic properties of the stopping variable. We demonstrate the desirable properties of our sequential approach, both theoretically and numerically. In our application to a study of hormone responsive breast cancer cell lines, we estimated the stopping variable for the sample size determination to be smaller than the actual sample size available to conduct the experiment. This means that we can obtain an accurate assessment of differential gene expression without compromising the cost and size of the study. Altogether, we anticipate that this approach could have an important contribution to microarray studies by improving the usual experimental designs and methods of analysis.

KEY WORDS: Dose-response; Gene expression; Robust regression; Sample size.

1. INTRODUCTION

Microarray chips and other high-throughput technologies have changed radically the nature of genomics assays by allowing the simultaneous screening of thousands of genes in a single experiment. Statistical inference about change patterns in microarrays has led to the competing applications of numerous statistical approaches and computing algorithms for several problems: class comparison, class discovery, and class prediction Simon (2003). A particular aspect of these experiments, which has received less attention, is the planning of the experimental design that should enable experimenters to have efficient and valid inference about expression profiling. Particular designs of interest for class comparison problems include two- and multiclass experiments, where the classes are the experimental conditions; for example, different types of tumor tissue, stages of tumor progression, doses of an exposure, or time-points. There have been a number of articles discussing the general issues related to experimental designs (for reviews, see Churchill 2002; McShane, Shih,, and Michalowska 2003), the paradigm being to use fixed-sample size plans. The theory of experimental design can be used to optimize the information gained from an experiment (Kerr and Churchill 2001a,b; Glonek and Solomon 2004) and to determine the number of arrays required to conduct the experiment. However, this leads generally to very crude sample size estimates, whose determination depends on quantities such as the variability of gene expression across the different experimental conditions and the false discovery rate (FDR) (Benjamini and Hochberg 1995) that are generally unknown before the study (Pan, Lin, and Le 2002; Zien, Fluck, Zimmer, and Lengauer 2003; Gadbury et al. 2004; Pawitan, Michiels, Koscielny, Gusnanto, and Ploner 2005; Muller, Parmigiani, Robert, and Rousseau 2004; Dobbin and Simon 2005). Therefore, there is no guarantee that the study will have sufficient efficiency to detect the expression profiles of interest. To address this problem, Warnes and Liu (2005) developed a procedure to estimate sample size that uses an estimate of the standard deviation for each gene based on control samples from existing studies. Tibshirani (2006) proposed a method for assessing sample sizes based on a permutation-based analysis of pilot data, which avoids strong parametric assumptions and allows prior information about the required quantities. The authors emphasized a two-stage approach but it remains unclear how the data from the pilot and main studies should be analyzed.

A more formal multistage strategy that does not rely on the distribution of primary measurements from other studies, or from a pilot study, is the sequential approach. In clinical trials, sequential designs are well known to offer numerous ethical and economic advantages due to the possibility of early stopping either for futility or obvious advantage of a treatment (Jennison and Turnbull 2000). Indeed, the sample size, which is unknown at the start of the study period, is determined subsequently in part by the nature of the sequentially accumulating data. Such approaches could be particularly beneficial in the design of efficient microarray experiments for several reasons. First, many biologists use sequential designs implicitly by collecting an initial sample (a pilot study for example) and by performing a first analysis. In some experiments, additional samples will be collected and further analyses will be carried out, but without proper adjustment for the multiple analyses. Sequential approaches are also known to require less observation than fixed-sample size approaches to reach the same conclusion. This will be very advantageous for microarray studies both because of their cost (DNA chips retail for about $1,000 each) and the difficulties to obtaining sample material. Finally, determining an appropriate sample size to conduct a microarray experiment is often impossible and sequential approaches give a more rational solution by accumulating information to provide an estimate of the sample size, an estimate that can be given with its precision (because it is a random variable). Despite these obvious advantages, sequential designs have received very little attention in the field of microarray studies and more generally in the field of genomics.

Gilles Durrieu is Associate Professor, Department of Statistics, University of Bordeaux 1, CNRS, UMR 5805 EPOC, 33120 Arcachon, France (E-mail: *g.durrieu@epoc.u-bordeaux1.fr*). Laurent Briollais is Research Scientist, Samuel Lunenfeld Research Institute, Toronto, ON M5G 1X5, Canada (E-mail: *laurent@mshri.on.ca*). This work was supported by the Discovery Grant 93270-04 from the Natural Sciences and Engineering Research Council of Canada and the Centre National de la Recherche Scientifique. The authors thank Dr. Cyr Emile M'lan, Prof. Jaromir Antoch, Prof. Jean-Marc Deshouillers, Prof. Jana Jurečková, Dr. Shelley Bull, and Dr. Chi Yip Ho for their contributions to our manuscript.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Applications and Case Studies
DOI 10.1198/jasa.2009.0135

To the best of our knowledge, only two articles discussed their application in the context of microarray studies. The first one introduced a sequential procedure for classification problems (Fu, Dougherty, Mallick, and Carroll 2005). It provides stopping criteria that ensure with a certain level of confidence that at stopping the misclassification probability of any future subject into a particular experimental group will be smaller than a predetermined threshold. Although interesting, this approach does not apply directly to class comparison problems. The second article deals with class comparison problems and proposed the construction of nonparametric prediction intervals to identify differentially expressed genes (Gibbons et al. 2005). A fixed number of control samples (normal tissues) is first obtained, followed by sequential collection of the experimental samples (tumor tissues), one by one or by groups. The control samples are then used to construct a predictive interval for the mean gene expression or other summary statistic of the experimental samples. The procedure stops when the probability that this summary statistic calculated for one specific gene (or a group of genes) in the experimental samples is outside the prediction limits with high probability. This approach is not truly sequential because the probability of being outside the predictive bounds is determined for different sizes of the experimental samples but does not depend on the accumulated information on gene expression data. To our knowledge, neither of these approaches has been applied in real situations.

In this article, we propose a formal theory for sequential microarray design in the context of class comparison problems. Section 2 describes the main novelties: that the decision to stop the experiment really depends on the data that are accumulating, and that the statistical properties of the stopping rule proposed are thoroughly studied both theoretically and by simulation. Our approach builds on the general linear regression framework, which is a very common strategy for microarray studies. Besides the sequential aspect, we use nonparametric regression to provide robustness against outliers and distribution-free inference, which is particularly advantageous in small sample size microarray problems. The principle of our approach is the construction of confidence interval (CI) for the model parameter, corresponding to a particular level of gene expression. The case of correlated observations is also discussed and treated by adjustment of the data. The definition of the stopping rule of the sequential procedure for a given precision is presented in Section 3. We develop both the case of continuous and discrete monitoring and the asymptotic properties of the stopping variable. In Section 4, we analyze gene expression profiles at various doses of estrogen exposure in breast cancer cell lines. Additional finite properties of our approach, in particular the use of continuous versus discrete sampling, and the potential bias due to batch effects are studied in Section 5 by simulation. Finally, Section 6 concludes with a general discussion of our approach and its feasibility and applicability in practice.

2. METHODS

2.1 The Linear Model Framework

Linear models have become a classical statistical framework for the analysis of two-channels cDNA arrays, especially with the development of the statistical package LIMMA in *R* (Smyth 2004). For example, Kerr et al. (2001a, b) used linear models and the analysis of variance (ANOVA) to estimate expression differences and assess the variability of their estimates. They assumed a fixed effects model for the log gene intensity, which accounts for the variability due to the array, dye effect, treatment (or experimental condition), and gene effect. Differentially expressed genes are those that exhibit significant treatment by gene interactions, whereas normalization is effected by including a dye term in the model. The random error of the model represents variations due to unknown sources and is typically assumed to be normally distributed. Other authors have proposed linear models on a gene-by-gene basis, with a separate error for each gene, but still including normalization as part of their linear model (Jin et al. 2001; Wolfinger et al. 2001). Their models also include a random effect for the arrays. These modeling approaches have raised several questions, in part because the inference about gene effects is performed jointly with the normalization process and the estimation of multiple other effects (Yang and Speed 2002). In contrast, other linear models have been proposed for the log-ratios of intensity from experiments in which normalization has been carried out separately for each slide, typically using a nonlinear adjustment, which could not be captured in a linear model (Yang and Speed 2002). The only terms that are included in these models relate to mRNA samples and their treatment. The general form of the model is

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta}+\varepsilon_n, \tag{1}$$

where for all $n \geq 1$, $\boldsymbol{Y} \equiv \boldsymbol{Y}_n = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n)'$ is the $\log_2$ ratio of gene expression from all slides, $\boldsymbol{X} \equiv \boldsymbol{X}_n \equiv \boldsymbol{X}_{n\times p}$ is a known $n \times p$ design matrix of experimental conditions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of unknown parameters to be estimated, and $\varepsilon \equiv \varepsilon_n \equiv (\varepsilon_1, \ldots, \varepsilon_n)'$is the error term. The principle of this model, in general, is to compare the means of the log gene expression distribution under different conditions (treatments). If ε follows a normal $N\,(0,\boldsymbol{I})$ distribution, then ordinary least squares provide the maximum likelihood estimator of $\boldsymbol{\beta}$. However, it is clear from microarray data that the normality assumptions are often violated for interesting genes. Unusual probes and outlying probe level measurements often occur to upset normality. To avoid the distributional assumptions and to protect against outlying measurements, we propose a robust inferential method for model (1) by using quantile regression.

An overview of robust statistics literature can be found in Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Jurečková and Sen (1996), and Huber (2003). Our approach is based on regression quantile estimators of $\boldsymbol{\beta}$ (Koenker and D'Orey 1987; Dodge and Jurečková 1995). Instead of focusing on the changes in the mean gene expression, the quantile regression approach allows one to test whether there is a change in the τth quantile of y for any given $\tau \in (0,1)$. When the conditional distributions of y are nonGaussian, the mean might not be the best summary, and a change in distributions may not be detected.

Inference for linear quantile regression models has become a subject of intense investigation in the past years. Any solution of the following minimization problem ($\boldsymbol{x}'_i$denotes the ith row of the matrix $\boldsymbol{X}$)

$$\widehat{\boldsymbol{\beta}}(\theta) \equiv \widehat{\boldsymbol{\beta}}^n(\theta) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta(Y_i - \boldsymbol{x}'_i\boldsymbol{\beta}) \tag{2}$$

is called a *θ-regression quantile* with $\rho_\theta(x) = x(\theta - \mathbb{I}(x < 0))$ where $\mathbb{I}(\mathcal{P})$ takes the value 1 or 0 depending on whether the condition $\mathcal{P}$ is satisfied or not. Here, we used the L^1 – norm estimator, also known as least absolute deviation (LAD) estimator, obtained by taking $\theta = 1/2$, chosen as an alternative to least squares estimators in the presence of heavy-tailed distribution. It is known that this estimator performs better in the presence of a heavy-tailed distribution. Following Jurečková (1984), under regularity conditions, we have as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\beta}_1^{\,n}(\theta) - \beta_1\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{q^2(\theta)}{4}\right), \qquad (3)$$

where $q(\theta) = 1/f\,(Q(\theta))$ is the quantile density function with $Q(\theta)$ denoting the quantile function. The asymptotic variance of $\widehat{\boldsymbol{\beta}}^{\,n}(\theta)$ is unknown because it depends on the distribution function f of the error term in the model, which itself is unknown. So, we estimate this asymptotic variance using the regression invariant and scale equivariant kernel-type estimator of $q(\theta)$ in model (1) based on the regression quantiles. This estimator of $q(\theta)$, for $0 < \theta < 1$, is defined by

$$\widehat{Z}_n(\theta) = \frac{1}{\nu_n^2}\int_0^1 \widehat{\beta}_1^{\,n}(w)\,k\left(\frac{\theta - w}{\nu_n}\right)dw,$$

where $(\nu_n)_{n\,\geq\,1}$ is a sequence of properly chosen bandwidths and k denotes the kernel function. The quality of a density estimate is now widely recognized to be primarily determined by the choice of bandwidth. The asymptotic behavior of this estimator has been studied by Dodge and Jurečková (1995). It was proven that under regular conditions, as $n \to \infty$,

$$\sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) = O_p(1) \text{ and } \sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, q^2(\theta)\,\overline{K}), \qquad (4)$$

where $\overline{K} = \int K^2(x)\,dx$ with $K(x) = \int_{-\infty}^{x} k(y)\,dy$. Because of the choice of LAD estimator, θ is fixed at 0.5.

2.2 Confidence Interval of the Model Intercept

To construct the expected CI of given bounded length, we estimate $q(1/2)/2$ by

$$\widehat{W}_n(1/2) = \frac{\widehat{Z}_n(1/2)}{2}.$$

Using (4) with $\theta = 1/2$ we have as $n \to \infty$,

$$\widehat{W}_n(1/2) \xrightarrow{P} \frac{1}{2f(0)}.$$

So (3) and Slutsky's theorem imply

$$\frac{\sqrt{n}\,\left(\widehat{\beta}_1^{\,n}(1/2) - \beta_1\right)}{\widehat{W}_n(1/2)} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \text{ as } n \to \infty.$$

Hence, a $(1 - \alpha)$% CI for β_1 is given by

$$I_n = \left[\widehat{\beta}_1^{\,n}(1/2) - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\widehat{W}_n(1/2), \widehat{\beta}_1^{\,n}(1/2) + \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\widehat{W}_n(1/2)\right]. \qquad (5)$$

2.3 Nonindependent Observations

In Section 2.1, we assumed that the errors are iid. When it is not the case, we can replace the model (1) by

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{h}(\lambda, \boldsymbol{X}_n\boldsymbol{\beta})\varepsilon_n, \qquad (6)$$

where ε_n are iid and independent of X_n and $\boldsymbol{h}\,(\boldsymbol{\lambda}, \boldsymbol{X}_n\boldsymbol{\beta})$ models the correlations between the error terms, which can depend on the known variables and the new parameter $\boldsymbol{\lambda}$ (Maronna, Martin, and Yohai 2006). For example, in our application $\boldsymbol{\lambda}$ represents the correlation between gene expressions within the same replicate. The same regression model (1) can then be applied to the following transformed variables

$$\boldsymbol{Y}_n' = \boldsymbol{Y}_n\boldsymbol{h}^{-1}(\lambda, \boldsymbol{X}_n\boldsymbol{\beta}) \quad\text{and}\quad \boldsymbol{X}_n' = \boldsymbol{X}_n\boldsymbol{h}^{-1}(\lambda, \boldsymbol{X}_n\boldsymbol{\beta}). \qquad (7)$$

An example of function h is given in our application. In the following, we will assume we are working with the transformed variables if there is evidence for correlation between the observations (for example, observations within the same technical or biological replicate).

3. SEQUENTIAL APPROACH

The goal of our analyses is to detect a significant change in gene expression across various experimental conditions. From the asymptotic distribution of the intercept given in Equation (4), it is straightforward to see that genes that are differentially expressed will have β_1 significantly different from 0. Therefore, a sequential procedure can be designed such that experiment stops whenever gene expression changes resulting from exposure to several conditions have been measured with a desired precision. We describe more formally this concept based on the work of several authors (Jurečková 1991; Jurečková and Sen 1996; Hušková 1994 among others) in the following section. Our contribution consists in the use of a robust estimator of linear regression parameters without assuming a distribution for the error term.

3.1 Stopping Rule for Continuous Monitoring

We construct a CI I_n for the first component of the linear model based on a robust estimator of β_1 that satisfies

$$L_n \leq 2d \quad\text{and}\quad P_F(I_n \ni \beta_1) \geq 1 - \alpha, \qquad (8)$$

where L_n denotes the length of I_n. Starting with an initial sample size n_0, the experiment stops when the number of replicates N_d is the smallest $n > n_0$ such that the length $L(N_d) < 2\ d$.

The length L_n of the interval I_n in (5) satisfies

$$\sqrt{n}L_n = 2\,z_{1-\alpha/2}\widehat{W}_n(1/2) \xrightarrow{P} \frac{z_{1-\alpha/2}}{f(0)} \quad \text{as } n \to \infty. \qquad (9)$$

So, for a given sample size n and for a given α, the length of the CI I_n is a random variable. The sequential procedure consists of adding one new observation at a time (i.e., one technical replicate at the high dose in our experiment) until the stopping rule $L_n \leq 2d$ ($d > 0$ fixed) is satisfied.

Comparing I_n to $I_n^* = [\widehat{\beta}_1^{\,n}(1/2) - d, \widehat{\beta}_1^{\,n}(1/2) + d]$ with fixed $d > 0$ using (5), our stopping rule can be expressed as

$$n \geq \frac{z_{1-\alpha/2}^2\widehat{W}_n^2(1/2)}{d^2}.$$

To avoid an erroneous determination of $\widehat{W}_n^2(1/2)$, we must choose an initial sample size n_0 that is sufficiently large. Then, the stopping variable N_d satisfying the previous conditions on n can be defined by

$$N_d = \min\left\{ n \geq n_0 \middle| n \geq \frac{z_{1-\alpha/2}^2 \widehat{W}_n^2(1/2)}{d^2} \right\}, \quad d>0. \tag{10}$$

If we have a given sufficiently large sequence of observations, then N_d would be a nondecreasing function of d. The monotonicity of N_d follows directly from the definition of N_d.

3.2 Stopping Rule for Discrete Monitoring: Group Sequential Approach

For discrete monitoring (group sequential approach), the $(1-\alpha)$ CI for β_1 at the kth analysis is given by

$$I_{mk} = \left[\hat{\beta}_1^{mk}(1/2) - \frac{C_k(\alpha)}{\sqrt{mk}}\,\widehat{W}_{mk}(1/2), \hat{\beta}_1^{mk}(1/2) + \frac{C_k(\alpha)}{\sqrt{mk}}\,\widehat{W}_{mk}(1/2)\right]. \tag{11}$$

$$I_{mk} = \left[\hat{\beta}_1^{mk}(1/2) - d_k, \hat{\beta}_1^{mk}(1/2) + d_k\right], \tag{12}$$

where m is the size of each group (assumed equal) and $C_k(\alpha)$ is the critical value for the kth analysis.

Then, the new stopping variable N_{d_k} for the group sequential approach can be defined by

$$N_{d_k} = mk = \min\left\{ n \geq n_0 \middle| n \geq \frac{C_k(\alpha)^2 \widehat{W}_n^2(1/2)}{d_k^2} \right\}, \quad d_k > 0. \tag{13}$$

In group sequential methods, the critical values $C_k(\alpha)$ can be determined by using an α-spending function (Jennison and Turnbull, 2000, chap. 7). The principle is to partition the Type I error into K probabilities $\pi_1, \pi_2, \ldots \pi_K$, where K is the number of analyses performed, which sum to α. Thus, π_k represents the probability of stopping at analysis k to reject H_0 when this hypothesis is true, also termed the error spent at stage k. It can be determined using an error spending function, which satisifies $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. The value $f(t)$ indicates the cumulative Type I error that is spent when a fraction t of the maximum anticipated information has been obtained, denoted $I_{\max}$. When this maximum information has been reached, the sequential procedure will stop, either accepting or rejecting the null hypothesis.

The Type I error allocated to each analysis is

$$\pi_1 = f(I_1/I_{\max}), \quad \pi_k = f(I_k/I_{\max}) - f(I_{k-1}/I_{\max}), \quad k = 2, 3, \ldots, K, \tag{14}$$

where I_k is the amount of information at analysis k. Without lack of generality, we took $I_k = mk$ and $I_{\max} = mK$.

The critical values need to satisfy the constraint

$$Pr\left\{|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1,} |Z_k| \geq c_k | H_0\right\} = \pi_k, \tag{15}$$

where $|Z_k|$ is the standardized test statistic at the kth analysis. In our case, the Z_k statistics correspond to the $\hat{\beta}_1^{mk}(1/2)$ divided by its standard deviation. The critical values can be determined numerically assuming a given α – *spending function.* In the following, we consider O'Brien and Fleming test (1979). Another form of this approach has been proposed by Lan and DeMets (Lan and DeMets 1983) with the following spending function

$$f(t) = \min\{2 - 2\phi(z_{1-\alpha/2}/\sqrt{t}), \alpha\} \tag{16}$$

where $t = I_k/I_{\max}$ and ϕ is the standard normal distribution.

3.3 Asymptotic Properties of the Stopping Variable

Our first result shows that our construction of N_d fulfills the requirements (8). We first state our main result and the proofs are given in the Appendix. These properties are satisfied for the continuous and discrete cases and our notation later refers to the continuous case.

Theorem 1. Under regularity conditions in the distribution function, the sequence of matrices $\mathbf{X}_\mathrm{n}$ and the kernel estimator, we have

1. $E_F(N_d) \to +\infty$ as $d \to 0_+$.
2. $P_F(N_d < +\infty) = 1$ for any $d > 0$.
3. $N_d \big/ n_d \xrightarrow{P} 1$ as $d \to 0_+$, where $n_d = z_{1-\alpha/2}^2 \sigma^2 / d^2$ and $\sigma = \sigma_{1/2} = 1/(2f(0))$.
4. $P_F(I_{N_d} \ni \beta_1) \geq 1 - \alpha$ for $\alpha \in (0,1)$ fixed as $d \to 0_+$.

The next theorem concerns the asymptotic behavior of the stopping variable.

Theorem 2. Let $\eta \in (1/4, 1/3)$ and ν_n be such that $\nu_n = n^{2\eta-1}$. Under regularity conditions we have as $d \to 0_+$:

$$d^{1-2\eta}\left(\sqrt{N_d} - \frac{z_{1-\alpha/2}\,\sigma}{d}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^{2(1-2\eta)}\, z_{1-\alpha/2}^{2(1-2\eta)} \int K^2(x)\,dx\right).$$

Proofs of Theorems 1 and 2 require an intermediate result covering essentially the problem of uniform continuity in probability (Anscombe condition, Anscombe 1952) of the least-absolute regression estimator and the asymptotic kernel variance estimator. The proof of Theorem 2 is given in the Appendix. To avoid the presentation of lengthy mathematical results, the proof of the Anscombe condition is given as a technical report posted online at the JASA website.

Proof of Theorem 1. (1) It follows by construction that N_d increases as d decreases; this together with the fact that $L_n > 0$ with probability 1 implies that

$$\forall m \geq 0, \quad \lim_{d \to 0_+} P_F\{N_d < m\} \leq \lim_{d \to 0_+} P_F\{L_{N_d} < 2d\} = 0,$$

and thus $N_d \xrightarrow{P} \infty$ as $d \to 0_+$. Therefore, using the monotone convergence theorem we have

$$\lim_{d \to 0} E_F(N_d) = E_F\left(\lim_{d \to 0} N_d\right) = \infty.$$

(2) For every $n \geq n_0$ we have

$$P_F\{N_d > n\} \le P_F\left\{n < d^{-2} z^2_{1-\alpha/2} \widehat{W}^2_n(1/2)\right\}.$$

From (4) we deduce that $\widehat{W}_n(1/2)/n \xrightarrow{p} 0$ as n → ∞, hence

$$P_F\{N_d = \infty\} = \lim_{n\to\infty} P_F\{N_d > n\} \le \lim_{n\to\infty} P_F\{n < d^{-2} z^2_{1-\alpha/2} \times \widehat{W}^2_n(1/2)\} = 0.$$

Therefore,

$$P_F\{N_d = \infty\} = 0 \text{ and } P_F\{N_d < +\infty\} = 1 \text{ for every } d > 0.$$

(3) Note that with probability 1, we have

$$\frac{z^2_{1-\alpha/2}\widehat{W}^2_{N_d}(1/2)}{d^2} \le N_d < \max\left(n_0 + 1, \frac{z^2_{1-\alpha/2}\widehat{W}^2_{N_d-1}(1/2)}{d^2} + 1\right).$$

Recalling that $n_d = z^2_{1-\alpha/2}\sigma^2/d^2$, we obtain

$$\frac{\widehat{W}^2_{N_d}(1/2)}{\sigma^2} < \frac{N_d}{n_d} < \max\left(\frac{n_0}{n_d}, \frac{\widehat{W}^2_{N_d-1}(1/2)}{\sigma^2}\right) + \frac{1}{n_d}. \qquad (17)$$

Moreover, using the fact that

$$\widehat{W}_n(1/2) \xrightarrow{P} \sigma \text{ as } n \to \infty, \quad N_d \xrightarrow{P} \infty \text{ as } d \to 0_+$$

and uniform continuity in probability of $\widehat{W}_n(1/2)$, we obtain

$$\frac{\widehat{W}_{N_d}(1/2)}{\sigma} \xrightarrow{P} 1 \text{ and } \frac{\widehat{W}_{N_d-1}(1/2)}{\sigma} \xrightarrow{P} 1 \text{ as } d \to 0_+. \qquad (18)$$

By (18), as $d \to 0_+$, we have

$$\max\left(\frac{n_0}{n_d}, \frac{\widehat{W}^2_{N_d-1}(1/2)}{\sigma^2}\right) \to \max(0, 1) = 1.$$

Finally, by (17) we obtain the expected result, i.e.,

$$\frac{N_d}{n_d} \xrightarrow{P} 1 \text{ as } d \to 0_+.$$

(4) Moreover, because $\widehat{\beta}^n_1$ is asymptotically normal and uniformly continuous in probability, we obtain,

$$\frac{\sqrt{N_d}\left(\widehat{\beta}_{N_d}(1/2) - \beta_1\right)}{\widehat{W}_{N_d}(1/2)} = \left(\frac{\sigma}{\widehat{W}_{N_d}(1/2)}\right)\left(\frac{N_d}{n_d}\right)^{1/2} \times \left(\frac{\sqrt{n_d}\left(\widehat{\beta}_{N_d}(1/2) - \beta_1\right)}{\sigma}\right)$$

$$= \left(\frac{\sigma}{\widehat{W}_{N_d}(1/2)}\right)\left(\frac{N_d}{n_d}\right)^{1/2}\left\{\frac{\sqrt{n_d}\left(\widehat{\beta}_{n_d}(1/2) - \beta_1\right)}{\sigma} + \frac{\sqrt{n_d}\left(\widehat{\beta}_{N_d}(1/2) - \widehat{\beta}_{n_d}(1/2)\right)}{\sigma}\right\} \xrightarrow{\mathcal{D}}$$

$\mathcal{N}(0, 1)$ as $d \to 0_+$.

As d → 0_+, we have

$$\lim_{d\to 0_+} P_F\left\{\left|\widehat{\beta}_{N_d}(1/2) - \beta_1\right| \le z_{1-\alpha/2}\widehat{W}_{N_d}(1/2)N_d^{-1/2}\right\} = 1 - \alpha. \qquad (19)$$

By (19) and (10) we finally have

$$\lim_{d\to 0_+} P_F\left\{\widehat{\beta}_{N_d}(1/2) - d \le \beta_1 \le \widehat{\beta}_{N_d}(1/2) + d\right\} \ge 1 - \alpha .$$

4. THE MICROARRAY EXPERIMENT

Recently, Lobenhofer et al. (2004) illustrated the interest of model-based approaches for microarray analysis in a study that tried to identify genes that respond to estrogen treatment and evaluated more particularly the doses of estrogen capable of inducing a transcriptional response in breast cancer cell lines. In terms of our methodology, this dataset possesses several major points of interest. First, we believe that their fixed sample size design is not optimal and the differentially expressed genes could have been detected using fewer replicates. Second, the design can be analyzed using discrete monitoring, the group can be a technical replicate ($m = 2$) or a biological replicate ($m = 4$). Because of the small sample size of this experiment, we decided to use the technical replicate as the group. Third, the observations can be correlated within each technical and biological replicate. Finally, this study was among the rare ones to clearly establish a dose-response effect in a gene expression experiment and therefore, the variable dose has a real scientific interest and needs to be modeled appropriately.

In this study, the gene expression levels of a hormone responsive breast cancer cell line (MCF-7) are measured after stimulation with various concentrations of estrogen, above (high-dose effect) and below (low-dose effect) normal physiologic levels and compared with the corresponding levels of expression in control samples (Lobenhofer et al. 2004). The data were downloaded from the website *http://dir.niehs.nih.gov/microarray/datasets/home-pub.htm.*

The cell lines were treated with estrogen at four concentrations (10^{-8} M, 10^{-10} M, high doses) and (10^{-13} M, 10^{-15} M, low doses) or with concentration-matched ethanol solvent (control samples). The RNA sample from each estrogen treatment and its corresponding control were compared using custom-made cDNA microarrays, ToxChip version 1.0, in a dye-swap fashion. Each chip had 1,920 clones double-printed in two subarrays, so each clone was duplicated on the array. The dye swap and the duplicated spots lead to four technical replicates for each biological replicate. There were two biological replicates, so eight measurements at each dose. In this analysis, we focused on seven genes that were validated by real-time polymerase chain reaction (RT-PCR). Five genes were confirmed upregulated and two were downregulated. These genes are known to be involved in different cancer pathways. Because the threshold dose can be observed below the level 10^{-8} M, we just included in our analysis the 2 low doses and only 1 high dose (10^{-10} M). Thus, there were a total of 24 measurements available for each gene (3 doses × 8 measurements).

The data were preprocessed before analysis as described in Lobenhofer et al. (2004). In brief, the gridding quality of microarray images was checked for misalignment using the gridcheck function in the R/maanova package. Spots with higher background intensity than foreground were removed from the analysis. The normalization process included an intensity-lowess transformation and channel-mean centering of the data.

The data were then transformed to remove the possible correlation between the observations belonging to the same biological replicate, following the principle described in Section 2.3. To estimate the function $\boldsymbol{h}$ in (6, 7), we proceeded as follows using the pooled data of the seven genes. First, we estimated the residuals from the quantile regression model specified in (1), including the dose and dye effects as explanatory variables. Then, we estimated the variance-covariance matrix of the residuals using a linear model with a compound symmetry correlation structure defined as a block-diagonal matrix in which each block is 4 × 4 with 1 for the diagonal entries and λ for the off-diagonal entries.

Each block of dimension (4 × 4) represents a biological replicate and we assumed the same correlation structure for each gene as well as the same correlation between two technical replicates. To analyze the data, we fitted the regression model (1) with the dose and dye effects as covariables using quantile regression on the transformed response and covariates. Then, we computed a robust CI I_n of fixed length (Sections 2–3) for the model intercept. The initial sample size n_0 was fixed at 16, which corresponds to the number of measurements at the two low doses. Then, we determine sequentially the number of groups of technical replicates at the high dose, where each group had two observations (duplicated spots). The experimental process stops for N_{d_k} the smallest $n > n_0$ is such that the length of I_n ($L_n < 2\,d$). The results are presented in Table 1 for $\alpha = 0.001$, the value chosen by Lobenhofer et al. (2004) to adjust for multiple comparisons in their study.

For each gene, the precision parameter d_k ($k = 1, \ldots, 4$) is determined by the constraint on the stopping variable $N_{d_k} = C_k^2(\alpha)\widehat{W}_n^2(1/2)/d_k^2$. It is lower than 24 and can take only the following values 16, 18, 20, 22, or 24. The estimator $\widehat{W}_n^2(1/2)$ denotes the kernel regression estimator of the standard deviation of the intercept parameter.

These results indicate the minimal number of technical replicates N_{d_k} required to estimate the gene expression level with a fixed precision. We observe that for a relatively small precision parameter (d_k values), the stopping value remains lower than 24 (the maximum number of technical replicates available for the 3 doses). Genes 1, 2, 3, 4, and 5 were significant upregulated differentially expressed ($p < 0.001$) and genes 6 and 7 are downregulated differentially expressed. These seven genes were confirmed to be differencially expressed by quantitative PCR. In conclusion, our analysis showed that the experimenter could have stopped the experiment after collecting only 22 replicates and would have reached the same conclusion as in Lobenhofer et al. (2004).

Table 1. Results from group sequential discrete monitoring microarray experminent

Label	Name	d	N_d	$\widehat{\beta}_1^{N_d}$	$SD\ (\widehat{\beta}_1^{N_d})$	99.9% CI
Gene 1	SDF1	0.375	22	1.175	0.283	(0.976, 1.374)
Gene 2	MYB	0.410	20	0.971	0.457	(0.578,1.364)
Gene 3	CDC28	0.243	18	0.668	0.111	(0.582,0.755)
Gene 4	LRP8	0.480	22	0.479	0.547	(0.094,0.864)
Gene 5	CDC25A	0.265	18	0.081	0.042	(0.048,0.113)
Gene 6	TMP3	0.215	18	−0.369	0.257	(−0.569,−0.169)
Gene 7	PRKCZ	0.226	20	−0.445	0.260	(−0.669,−0.221)

NOTE: d, precision parameter; N_d, stopping variable estimator; $\widehat{\beta}_1^{N_d}$, LAD intercept estimator; $SD(\widehat{\beta}_1^{N_d})$, kernel regression estimator of the standard deviation of $\widehat{\beta}_1^{N_d}$; CI, 99.9% confidence interval of the intercept.

5. SIMULATION STUDY

To better evaluate the performance of our sequential method, we conducted a small simulation study. Our goal was to assess more specifically the sensitivity of our sequential approach to three factors: (1) the use of continuous versus discrete sequential procedures; (2) the distribution of the error term in the model; (3) the presence of a batch effect. The design of the simulation study mimicked the experimental design used in our application. The X and ϵ were generated independently, where X follows under discrete monitoring a discrete uniform distribution taking the values 0, 3, and 5 and under continuous monitoring a uniform distribution with values belonging to the interval (0, 5). In the discrete monitoring, each group included sequentially had a size of eight observations and the group was added at one specific dose. We also considered the possibility of a batch effect, the value of which changed whenever a new group was added (discrete monitoring) or when eight observations were successively added (continuous monitoring). The data were generated from the following linear model:

$$Y = \beta_1 + \beta_2\text{Dose} + \beta_3\text{Batch} + \delta\varepsilon$$

where β_1 is the intercept denoting the gene expression at an initial dose, β_2 is the slope parameter coding for the change in gene expression with dose effect. The terms β_3 and δ correspond to the additive and multiplicative effects of the batch. The source of variation ϵ is a random vector of independent and identically distributed errors with a distribution function, which is either Normal $N\left(0, \sigma_*^2\right)$ or Laplace $L\left(0, \sigma_*^2\right)$. The reason for choosing this latter distribution is to show the robustness of the L^1 estimator in presence of long-tailed error distributions.

The regression parameters (β_1, β_2, and β_3) were fixed respectively to 1, −0.2 and 0.1 and σ_* was set to 0.3 and 0.4, as observed in the application. The precision parameter d was determined such that the theoretical stopping rule value $n_d = z_{1-\alpha/2}^2\sigma^2/d^2$ was equal to 32 or 56, where σ^2 denotes the theoretical asymptotic variance of the β_1 estimator. In all simulations, we used $\nu_n = n^{-\delta}$, with $\delta = 1 - 2\,\eta = 0.42 \in (1/3, 1/2)$. In this situation, the conditions of the theorems giving the limiting behavior of the kernel estimator are satisfied. The Epanechnikov kernel was used but we obtained similar results with the Gaussian kernel. The initial sample size was fixed at $n_0 = 16$ in all simulations and the number of Monte-Carlo simulations was 1,000. Results are presented for the normal error distribution in Table 2 and for Laplace error distribution in Table 3.

Continuous versus Discrete Sequential Procedure With No Batch Effect. We obtained almost unbiased parameter estimates of the regression parameters when the procedure was continuous. The use of a discrete procedure does not alter the

Table 2. Simulation results with normal error in regression model (standard deviations are in parenthesis)

σ_*	d	Batch	n_d	Median N_d	Mean intercept True = 1	Mean slope True = − 0.2	Mean batch True = 0.1	$\sigma = \sigma_{1/2}$	Mean W_{N_d}	CI 99.9%	CIL	CP
						Continuous						
0.3	0.2187	0	32	32	0.9970	−0.1986	—	0.3760	0.3522	(0.7835, 1.2104)	0.4269	1.0000
				(0.7611)	(0.1351)	(0.0485)	—		(0.1083)			
0.3	0.2187	1	32	35	0.9851	−0.2017	0.1070	0.3760	0.3521	(0.7625, 1.2078)	0.4453	1.0000
				(2.9536)	(0.2199)	(0.0678)	(0.1117)		(0.1320)			
0.4	0.2916	0	32	32	0.9991	−0.1995	—	0.5013	0.4756	(0.7177, 1.2806)	0.5629	1.0000
				(0.7738)	(0.1783)	(0.0644)	—		(0.1480)			
0.4	0.2916	1	32	34	0.9994	−0.2017	0.1034	0.5013	0.4715	(0.6972, 1.3016)	0.6044	0.9753
				(1.1264)	(0.3309)	(0.0781)	(0.1414)		(0.1818)			
0.3	0.1653	0	56	48	0.9968	−0.1981	—	0.3760	0.3352	(0.8269, 1.1666)	0.3397	0.9988
				(1.3801)	(0.1159)	(0.0418)	—		(0.1046)			
0.3	0.1653	1	56	55	0.9996	−0.1980	0.0989	0.3760	0.3384	(0.8099, 1.1893)	0.3794	0.9782
				(2.4446)	(0.2198)	(0.0545)	(0.0950)		(0.1349)			
0.4	0.2204	0	56	53	0.9979	−0.1988	—	0.5013	0.4568	(0.7725, 1.2233)	0.4508	0.9988
				(1.4393)	(0.1486)	(0.0539)	—		(0.1413)			
0.4	0.2204	1	56	52	1.0025	−0.1987	0.0981	0.5013	0.4418	(0.7533, 1.2517)	0.4984	0.9855
				(2.2060)	(0.2804)	(0.0733)	(0.1147)		(0.1824)			
						Discrete						
0.3	0.2187	0	32	40	0.9896	−0.1953	—	0.3760	0.3570	(0.7866, 1.1927)	0.4061	0.9876
				(0.8261)	(0.1590)	(0.0752)	—		(0.1062)			
0.3	0.2187	1	32	48	0.9909	−0.1954	0.0994	0.3760	0.3731	(0.7803, 1.2015)	0.4212	0.9517
				(1.2895)	(0.2255)	(0.0820)	(0.0680)		(0.1151)			
0.4	0.2916	0	32	40	0.9913	−0.1989	—	0.5013	0.4711	(0.7201, 1.2624)	0.5423	0.9959
				(0.8871)	(0.2159)	(0.1033)	—		(0.1428)			
0.4	0.2916	1	32	48	0.9749	−0.1933	0.1048	0.5013	0.4971	(0.6950, 1.2548)	0.5598	0.9509
				(1.3299)	(0.3133)	(0.1128)	(0.0927)		(0.1531)			
0.3	0.1653	0	56	56	0.9995	−0.1992	—	0.3760	0.3309	(0.8329, 1.1662)	0.3333	0.9948
				(1.4688)	(0.1363)	(0.0665)	—		(0.1088)			
0.3	0.1653	1	56	72	0.9984	−0.2000	0.1017	0.3760	0.3456	(0.8188, 1.1780)	0.3592	0.9684
				(2.3661)	(0.2074)	(0.0732)	(0.0631)		(0.1145)			
0.4	0.2204	0	56	56	0.9947	−0.1954	—	0.5013	0.4404	(0.7676, 1.2217)	0.4541	0.9886
				(1.4287)	(0.1913)	(0.0913)	—		(0.1396)			
0.4	0.2204	1	56	72	0.9923	−0.1948	0.0974	0.5013	0.4621	(0.7594, 1.2252)	0.4658	0.9606
				(2.2734)	(0.2614)	(0.0946)	(0.0802)		(0.1533)			

NOTE: σ_*, standard deviation of the error term; d, precision parameter; Batch (takes the value 1 and 0 according to whether the batch effect is included or not), $n_d = z_{1-\alpha/2}^2\ \sigma^2/d^2$ is the theoretical stopping variable value in the continuous case and for the discrete case this is the sample size to calculate $I_{\max}$ and the corresponding critical values; $\sigma = \sigma_{1/2} = 1/2 f(0)$ is the asymptotic standard deviation of the intercept regression estimator; W_{N_d}, kernel regression estimator of σ; CI, mean 99.9% confidence interval of the intercept; CIL, length of the CI; CP, coverage probability. The results are based on the simulation study with 1,000 replications.

Table 3. Simulation results with Laplace error in regression model (standard deviations are in parenthesis)

σ_*	d	Batch	n_d	Median N_d	Mean intercept True = 1	Mean slope True = −0.2	Mean batch True = 0.1	$\sigma = \sigma_{1/2}$	Mean W_{N_d}	CI 99.9%	CIL	CP
						Continuous						
0.3	0.1234	0	32	31	0.9948	−0.1968	—	0.2121	0.1967	(0.8819, 1.1078)	0.2260	1.0000
				(0.8613)	(0.0656)	(0.0252)	—		(0.0685)			
0.3	0.1234	1	32	33	0.9874	−0.1969	0.1085	0.2121	0.1938	(0.8626, 1.1122)	0.2496	0.9844
				(0.9950)	(0.1316)	(0.0332)	(0.0620)		(0.0805)			
0.4	0.1645	0	32	31	0.9861	−0.1931	—	0.2828	0.2562	(0.8343, 1.1379)	0.3036	1.0000
				(0.8309)	(0.0937)	(0.0361)	—		(0.0903)			
0.4	0.1645	1	32	33	0.9711	−0.1936	0.1117	0.2828	0.2603	(0.8089, 1.1333)	0.3244	0.9847
				(1.0535)	(0.1671)	(0.0431)	(0.0779)		(0.1083)			
0.3	0.0933	0	56	48	0.9900	−0.1955	—	0.2121	0.1887	(0.8983, 1.0818)	0.1836	0.9990
				(1.3714)	(0.0577)	(0.0225)	—		(0.0649)			
0.3	0.0933	1	56	49	0.9833	−0.1970	0.1098	0.2121	0.1843	(0.8779, 1.0888)	0.2109	0.9702
				(2.4277)	(0.1112)	(0.0313)	(0.0480)		(0.0812)			
0.4	0.1240	0	56	49	0.9882	−0.1943	—	0.2828	0.2507	(0.8671, 1.1094)	0.2423	1.0000
				(1.5013)	(0.0768)	(0.0311)	—		(0.0877)			
0.4	0.1240	1	56	44	0.9858	−0.1958	0.1074	0.2828	0.2431	(0.8527, 1.1190)	0.2663	0.9796
				(2.2752)	(0.1418)	(0.0411)	(0.0684)		(0.1087)			
						Discrete						
0.3	0.1234	0	32	40	0.9934	−0.1958	—	0.2121	0.2020	(0.8870, 1.0997)	0.2127	0.9959
				(0.9797)	(0.0764)	(0.0375)	—		(0.0676)			
0.3	0.1234	1	32	48	0.9842	−0.1946	0.1047	0.2121	0.2111	(0.8703, 1.0981)	0.2278	0.9692
				(1.3594)	(0.1119)	(0.0446)	(0.0357)		(0.0739)			
0.4	0.1645	0	32	40	0.9860	−0.1908	—	0.2828	0.2640	(0.8457, 1.1263)	0.2806	0.9948
				(0.9060)	(0.1019)	(0.0529)	—		(0.0897)			
0.4	0.1645	1	32	48	0.9771	−0.1906	0.1060	0.2828	0.2785	(0.8183, 1.1359)	0.3176	0.9571
				(1.3292)	(0.1555)	(0.0615)	(0.0478)		(0.1015)			
0.3	0.0933	0	56	64	0.9907	−0.1943	—	0.2121	0.1913	(0.9015, 1.0800)	0.1785	0.9907
				(1.8276)	(0.0677)	(0.0340)	—		(0.0679)			
0.3	0.0933	1	56	72	0.9879	−0.1965	0.1042	0.2121	0.2004	(0.8890, 1.0869)	0.1979	0.9638
				(2.4965)	(0.1011)	(0.0397)	(0.0353)		(0.0758)			
0.4	0.1240	0	56	56	0.9846	−0.1917	—	0.2828	0.2467	(0.8678, 1.1013)	0.2335	0.9897
				(1.5957)	(0.0878)	(0.0431)	—		(0.0867)			
0.4	0.1240	1	56	72	0.9746	−0.1904	0.1060	0.2828	0.2685	(0.8399, 1.1092)	0.2693	0.9590
				(2.4501)	(0.1356)	(0.0537)	(0.0479)		(0.0977)			

NOTE: σ_*, standard deviation of the error term; d, precision parameter; batch (takes the value 1 and 0 according to whether the batch effect is included or not); $n_d = z^2_{1-\alpha/2}\sigma^2/d^2$ is the theoretical stopping variable value in the continuous case and for the discrete case this is the sample size to calculate $I_{\max}$ and the corresponding critical values; $\sigma = \sigma_{1/2} = 1/2f(0)$ is the asymptotic standard deviation of the intercept regression estimator; W_{N_d}, kernel regression estimator of σ; CI, mean 99.9% confidence interval of the intercept; CIL, length of the CI; CP, coverage probability. The results are based on the simulation study with 1,000 replications.

parameter estimates but the standard deviation of the estimates is slightly larger in that case. The standard deviation of the intercept, estimated by the kernel procedure, is generally slightly underestimated compared with the theoretical value in both sequential procedures. The sample size required to stop the experiment is larger under the discrete procedure. In some cases, one or two additional groups (of eight observations) are needed to stop the experiment compared with the continuous procedure. The length of the CI is inversely proportional to the stopping rule value and, consequently, it is smaller under the discrete procedure. The coverage probability was very close to the 99.9% nominal level in the continuous procedure but was lower than that in the discrete procedure. This is because more groups are used than expected, and therefore the critical values used in the group sequential procedure were not optimal.

Batch Effect. The parameter estimates (intercept and slope) are not altered by the presence of the batch effect and are still almost unbiased. The additive effect of the batch is also well estimated in the model. The length of the CIs is larger when there is a batch effect, which is expected because the multiplicative effect of the batch increased the variance of the error distribution in the regression model. The batch effect also affected the coverage probabilities, which are smaller than the nominal 99.9% level even in the continuous case. The bias in the coverage probabilities is larger with the discrete sequential procedure.

Normal versus Laplace Error Distribution. The results are very similar with normal and Laplace error terms. The parameter estimate is not very sensitive to the distribution of the error term, which shows the value of a robust estimator.

6. DISCUSSION

Our article establishes a general theoretical framework for sequential approaches in microarray experiments. Our theoretical developments as well as our real and simulated data demonstrate its advantages as compared with fixed-sample size approaches. More particularly, our application showed that we were able to replicate Lobenhofer et al.'s (2004) results and reach the same conclusions with fewer technical replicates. Sequential designs, in contrast to fixed-sample size designs, can achieve efficiency by making just enough measurements to evaluate the experimental endpoints with the desired precision. Numerically, we also proved that our approach is valid in finite samples and is robust to the choice of the error term distribution in the linear model. The use of discrete instead of continuous monitoring did not alter the parameter estimates of the model. However, for small sample size problems ($n_d < 25$), as observed in our application, the use of discrete monitoring slightly increased the sample size needed to stop the experiment. Our sequential approach is therefore slightly liberal and some adjustments of the critical values are required to achieve the correct nominal Type I error rate.

The major question related to this type of approach is its feasibility and applicability in practice. We partly answered this question through our application and small simulations. In the application, we sampled sequentially the technical replicates with two observations each. Sampling groups of biological replicates, each with four observations, would have been more relevant, but was not possible because of the relatively small sample size available for the experiment. The initial sample size was fixed to 16 and therefore the range of possible stopping variables was quite limited. Despite this limitation, our sequential procedure was always able to converge and gave CIs that were relatively precise for this type of problem. The small simulations confirmed the robustness of our approach to outliers and departure from normality. In many biological experiments, especially those involving technologies such as microarrays, data have generally a low signal-to-noise ratio that makes analysis more complex and very sensitive to outliers. Therefore, our approach warrants more general application in microarray studies. Our simulations also demonstrated the accuracy of the procedure to estimating the model parameters and its good behavior in the presence of batch effects.

Besides the statistical properties mentioned previously, it is important to discuss the technical feasibility and applicability of the method. In a recent NCIC (National Cancer Institute of Canada) project, we planned to apply a novel microarray experimental design where four batches of eight microarrays will be sequentially realized with the aim to study tumor progression in prostate cancer patients. The first two batches allow one to perform a first screen of methylation profiles and draw hypotheses regarding the role of several genetic pathways. Each batch corresponds to a particular stage, so the next batches will make precise the role of these pathways in tumor progression. In our simulations, we introduced a method that took direct account of batch effects. Alternatively, one could adjust the data before analysis at the preprocessing step of the experiment (Johnson, Li, and Rabinovic 2007). Because most array technologies are based on comparative hybridization, one can use the control sample for the adjustment. Therefore, we think batch effects are not a major problem to deal with, when present in the experiment.

At this stage of development, our approach can also be applied to find individual genes differentially expressed across several experimental conditions (treatment versus control, different doses, or time-points) and rank them according to their importance for the given class comparison problem. For example, a classical z test can be performed on the intercept and can serve as a criterion to rank the most important genes in the microarray experiment. We also plan an extension of this method to the joint analysis of multiple genes that could be differentially expressed in a coordinated manner. For example, the work from Gibbons et al. (2005) can be easily adapted to our method. We are also planning to integrate the correlation structure directly in the analysis rather than adjusting for it to gain some efficiency.

Altogether, we anticipate that this approach could have a significant contribution to microarray data analysis by improving the usual experimental designs and methods of analysis.

APPENDIX: PROOFS

Before demonstrating the proof of Theorem 2, we provide some auxiliary results on the asymptotic normality of N_d. We now study the asymptotic behavior of the stopping rule N_d. We assume that $\nu_n = n^{2\eta-1}$ for some $\eta \in (1/4, 1/3)$.

Lemma A.1. Let $\eta \in (1/4, 1/3)$ and ν_n be such that $\nu_n = n^{2\eta-1}$. Under regularity conditions, we have as $d \to 0_+$:

(1)

$$N_d^\eta\left(\widehat{W}_{N_d}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\overline{K}\right), \tag{A.1}$$

(2)

$$N_d^\eta\left(\widehat{W}_{N_d-1}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\overline{K}\right), \tag{A.2}$$

(3) for all ε' > 0:

$$P\left\{N_d^\eta\left|\left(\frac{\sqrt{N_d} - \sqrt{N_d - 1}}{\sqrt{\overline{K}}\,\sigma}\right)\frac{d}{z_{1-\alpha/2}}\right| > \epsilon'\right\} \to 0. \tag{A.3}$$

Proof of Lemma A.1. (1) By (4) we obtain

$$\sqrt{n\nu_n}\left(\widehat{W}_n(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\overline{K}\right). \tag{A.4}$$

Taking $\nu_n = n^{2\eta-1}$, we have

$$\sqrt{N_d\nu_{N_d}}\left(\widehat{W}_{N_d}(1/2) - \sigma\right) = \left(\frac{N_d}{z_{1-\alpha/2}^2\sigma^2/d^2}\right)^\eta \left(\frac{z_{1-\alpha/2}^2\sigma^2}{d^2}\right)^\eta \left(\widehat{W}_{N_d}(1/2) - \sigma\right). \tag{A.5}$$

By the third result of Theorem 1 we have, as $d \to 0_+$,

$$\frac{N_d^\eta}{\frac{z_{1-\alpha/2}^{2\eta}\sigma^{2\eta}}{d^{2\eta}}} \xrightarrow{P} 1 \tag{A.6}$$

and by (A.4), as $d \to 0_+$,

$$P\left\{n^\eta\left(\widehat{W}_n(1/2) - \sigma\right) \le y\right\} \to \Phi\left(\frac{y}{\sqrt{\overline{K}}\,\sigma}\right), \quad \forall y \in \mathbb{R}, \tag{A.7}$$

where Φ denotes the cumulative distribution function of the standard normal distribution.In addition to (A.6) and (A.7), the theorem of Anscombe (1952) allows us to conclude that, $\forall y \in \mathbb{R}$,

$$P\left\{\frac{\sigma^{2\eta} z_{1-\alpha/2}^{2\eta}}{d^{2\eta}}\left(\widehat{W}_{N_d}(1/2) - \sigma\right) \le y\right\} \to \Phi\left(\frac{y}{\sqrt{\overline{K}}\,\sigma}\right)$$

as $d \to 0_+$.

Therefore, as $d \to 0_+$, we have

$$\frac{\sigma^{2\eta} z_{1-\alpha/2}^{2\eta}}{d^{2\eta}}\left(\widehat{W}_{N_d}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \overline{K}\sigma^2\right). \tag{A.8}$$

Using (A.6), (A.8), and Slutsky's Theorem on (A.5), we obtain

$$N_d^\eta\left(\widehat{W}_{N_d}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\,\overline{K}\right) \text{ as } d \to 0_+,$$

(2) Relation (A.2) is derived in the same way.

(3) We now prove (A.3). For $x > 1$, we have $0 \le \sqrt{x} - \sqrt{x-1} \le 1/(2\sqrt{x-1})$. We thus have, for all E' > 0,

$$P\left\{N_d^\eta\left(\frac{\sqrt{N_d} - \sqrt{N_d - 1}}{\sqrt{\overline{K}}\,\sigma}\right)\frac{d}{z_{1-\alpha/2}} > \epsilon'\right\} \le P\left\{N_d^\eta\frac{d\,(N_d-1)^{-1/2}}{2\sqrt{\overline{K}}\,z_{1-\alpha/2}\,\sigma} > \epsilon'\right\}.$$

Moreover, as $d \to 0_+$, we have

$$N_d^\eta\frac{d\,(N_d-1)^{-1/2}}{\sqrt{\overline{K}}\,z_{1-\alpha/2}\,\sigma\,2} \sim \frac{N_d^{\eta-1/2}}{2\sqrt{n_d}\sqrt{\overline{K}}}.$$

Consequently, by Theorem 1 (3), we have as $d \to 0_+$,

$$\frac{N_d^{\eta-1/2}}{2\sqrt{n_d}\sqrt{\overline{K}}} \xrightarrow{P} 0.$$

Proof of Theorem 2. One one hand, as $N_d = \min\{n \ge n_0 \mid \widehat{W}_n(1/2) \le \frac{\sqrt{n}\,d}{z_{1-\alpha/2}}\}$, with $d > 0$, we have

$$\widehat{W}_{N_d}(1/2) \le \frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}},$$

and, therefore as $d \to 0_+$,

$$\limsup P\left\{N_d^\eta\left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\} \le \limsup P\left\{N_d^\eta\left(\widehat{W}_{N_d}(1/2) - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\}.$$

By Lemma A.1 (1), we have as $d \to 0_+$,

$$\limsup P\left\{N_d^\eta\left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\} \le \Phi(y). \tag{A.9}$$

On the other hand, according to the definition of the stopping rule N_d, we have

$$\widehat{W}_{N_d-1}(1/2) > \frac{\sqrt{N_d - 1}\,d}{z_{1-\alpha/2}},$$

and therefore, as $d \to 0_+$,

$$\liminf P\left\{N_d^\eta\left(\widehat{W}_{N_d-1}(1/2) - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\} \le \liminf P\left\{N_d^\eta\left(\frac{\sqrt{N_d-1}\,d}{z_{1-\alpha/2}} - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\}.$$

Using Lemma A.1 (2) and Lemma A.1 (3), we obtain, as $d \to 0_+$,

$$\liminf P\left\{N_d^\eta\left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\} \ge \Phi(y). \tag{A.10}$$

Therefore, by (A.9) and (A.10), we have $\forall y \in \mathbb{R}$

$$P\left\{N_d^\eta\left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right)\frac{1}{\sqrt{\overline{K}}\,\sigma} \le y\right\} \to \Phi(y) \text{ as } d \to 0_+.$$

Since $d\sqrt{N_d} \xrightarrow{P} z_{1-\alpha/2}\,\sigma$, the proof is complete.

[Received October 2006. Revised December 2008.]

REFERENCES

Anscombe, F. J. (1952), "Large Sample Theory of Sequential Estimation," *Proceedings of the Cambridge Philosophical Society,* 48, 600–607.

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,* Ser. A (General), 57, 289–300.

Churchill, G. A. (2002), "Fundamentals of Experimental Design for cDNA Microarrays," *Nature Genetics,* 32, 490–495.

Dobbin, K., and Simon, R. (2005), "Sample Size Determination in Microarray Experiments for Class Comparison and Prognostic Classification," *Biostatistics (Oxford, England),* 6, 27–38.

Dodge, Y., and Jurečková, J. (1995), "Estimation of Quantile Density Function Based on Regression Quantiles," *Statistics & Probability Letters,* 23, 73–78.

Fu, W., Dougherty, E., Mallick, B., and Carroll, R. (2005), "How Many Samples Are Needed to Build a Classifier: A General Sequential Approach," *Bioinformatics (Oxford, England),* 21, 63–70.

Gadbury, G. L., Page, G. P., Edwards, J. W., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J., and Allison, D. B. (2004), "Power Analysis and Sample Size Estimation in the Age of High Dimensional Biology: A Parametric Bootstrap Approach and Examples From Microarray Research," *Statistical Methods in Medical Research,* 13, 325–338.

Gibbons, R. D., Bhaumik, D. K., Cox, D. R., Grayson, D. R., Davis, J. M., and Sharma, R. P. (2005), "Sequential Prediction Bounds for Identifying Differentially Expressed Genes in Replicated Microarray Experiments," *Journal of Statistical Planning and Inference,* 129, 19–37.

Glonek, G. F., and Solomon, P. J. (2004), "Factorial and Time Course Designs for cDNA Microarray Experiments," *Biostatistics (Oxford, England),* 5, 89–111.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

Huber, P. J. (2003), *Robust Statistics*, New York: Wiley series in probability and statistics.

Hušková, M. (1994), "Some Sequential Procedures Based on Regression Rank Scores," *Nonparametric Statistics,* 3, 285–298.

Jennison, C., and Turnbull, B. W. (2000), *Group Sequential Methods with Applications to Clinical Trials*, New York: Chapman and Hall.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001), "The Contribution of Sex, Genotype and Age to Transcriptional Variance in *Drosophila melanogaster*," *Nature Genetics,* 29, 389–395.

Johnson, W. E., Li, C., and Rabinovic, A. (2007), "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods," *Biostatistics (Oxford, England),* 8, 118–127.

Jurečková, J. (1984), "Regression Quantiles and Trimmed Least Squares Estimator Under a General Design," *Kybernetika,* 20, 345–357.

Jurečková, J. (1991): "Confidence Sets and Intervals," in *Handbook of Sequential Analysis*, eds. B. K. Ghosh and P. K. Sen, New-York: Marcel Dekker, pp. 269–282.

Jurečková, J., and Sen, P. K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations*, New York: John Wiley.

Kerr, M. K., and Churchill, G. A. (2001a), "Experimental Design for Gene Expression Microarrays," *Biostatistics (Oxford, England),* 2, 183–201.

Kerr, M. K., and Churchill, G. A. (2001b), "Statistical Design and the Analysis of Gene Expression Microarray Data," *Genetical Research,* 77, 123–128.

Koenker, R. W., and D'Orey, V. (1987), "Computing Regression Quantiles," *Journal of the Royal Statistical Society,* Ser. A (General), C49, 383–401.

Lan, G. K. K., and DeMets, D. L. (1983), "Discrete Sequential Boundaries for Clinical Trials," *Biometrika,* 70: 659–663.

Lobenhofer, E. K., Cui, X., Bennett, L., Cable, P. L., Merrick, B. A., Churchill, G. A., and Afshari, C. A. (2004), "Exploration of Low-Dose Estrogen Effects: Identification of No Observed Transcriptional Effect Level (NOTEL)," *Toxicologic Pathology,* 2, 482–492.

Maronna, R. A., Martin, R. M., and Yohai, V. J. (2006), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

McShane, L. M., Shih, J. H., and Michalowska, A. M. (2003), "Statistical Issues in the Design and Analysis of Gene Expression Microarray Studies of Animal Models," *Journal of Mammary Gland Biology and Neoplasia,* 8, 359–374.

Muller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays," *Journal of the American Statistical Association,* 99, 990–1001.

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics,* 35, 549–556.

Pan, W., Lin, J., and Le, C. T. (2002), "How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach," *Genome Biology,* 3, 0022.1–0022.10.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005), "False Discovery Rate, Sensitivity and Sample Size for Microarray Studies," *Bioinformatics (Oxford, England),* 21, 3017–3024.

Simon, R. M. (2003), "Using DNA Microarrays for Diagnostic and Prognostic Prediction," *Expert Review of Molecular Diagnostics,* 3, 587–595.

Smyth, G. K. (2004), "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology,* 3. Article 3.

Tibshirani, R. (2006), "A Simple Method for Assessing Sample Sizes in Microarray Experiments," *BMC Bioinformatics,* 7, 106.

Warnes, G. R., and Liu, P. (2005), "Sample Size Estimation for Microarray Experiments," available at www.bioconductor.org/repository/devel/vignette/ssize.pdf.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001), "Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models," *Journal of Computational Biology,* 8, 625–637.

Yang, Y. H., and Speed, T. P. (2002), "Design Isuues for cDNA Microarray Experiments," *Nature Review Genetics,* 3, 579–588.

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003), "Microarrays: How Many You Need," *Journal of Computational Biology,* 10, 653–667.