1-1-2007

A bayesian approach to estimation and testing in time-course microarray experiments

Claudia Angelini

Daniela De Canditiis

Margherita Mutarelli

Marianna Pensky
University of Central Florida

Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 24

A Bayesian Approach to Estimation and Testing in Time-course Microarray Experiments

Claudia Angelini, *Istituto per le Applicazioni del Calcolo*
Daniela De Canditiis, *Istituto per le Applicazioni del Calcolo*
Margherita Mutarelli, *Lab. Bioinformatica, ISA-CNR; Dip. Patol. Gen., Seconda Universitµa di Napoli, Italy*
Marianna Pensky, *University of Central Florida*

Recommended Citation:
Angelini, Claudia; De Canditiis, Daniela; Mutarelli, Margherita; and Pensky, Marianna (2007) "A Bayesian Approach to Estimation and Testing in Time-course Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*: Vol. 6: Iss. 1, Article 24.
DOI: 10.2202/1544-6115.1299

A Bayesian Approach to Estimation and Testing in Time-course Microarray Experiments

Claudia Angelini, Daniela De Canditiis, Margherita Mutarelli, and Marianna Pensky

Abstract

The objective of the present paper is to develop a truly functional Bayesian method specifically designed for time series microarray data. The method allows one to identify differentially expressed genes in a time-course microarray experiment, to rank them and to estimate their expression profiles. Each gene expression profile is modeled as an expansion over some orthonormal basis, where the coefficients and the number of basis functions are estimated from the data. The proposed procedure deals successfully with various technical difficulties that arise in typical microarray experiments such as a small number of observations, non-uniform sampling intervals and missing or replicated data. The procedure allows one to account for various types of errors and offers a good compromise between nonparametric techniques and techniques based on normality assumptions. In addition, all evaluations are performed using analytic expressions, so the entire procedure requires very small computational effort. The procedure is studied using both simulated and real data, and is compared with competitive recent approaches. Finally, the procedure is applied to a case study of a human breast cancer cell line stimulated with estrogen. We succeeded in finding new significant genes that were not marked in an earlier work on the same dataset.

Author Notes: This research was supported in part by the NSF grants DMS-0505133 and DMS-0652524, the CNR Short Term Mobility grants 2005 and 2006, the PhD program in Computational Biology of the Second University of Napoli and the "Progetto Oncoproteomica" ISS/ISA-CNR. We are grateful to Luigi Cicatiello, Angelo Facchiano and Alessandro Weisz for their constructive comments, to Prof. Rita Carotenuto and Giovanni Ferraiolo for their help in copy-editing the manuscript and two anonymous referees whose valuable comments helped to substantially improve the paper. Also, Marianna Pensky would like to thank Claudia Angelini and Daniela De Canditiis for warm hospitality while visiting Italy to carry out part of this work and Claudia Angelini would like to thank Marianna Pensky for the warm hospitality while visiting Florida.

Erratum

Page 2, paragraph 1, lines 8-10, which reads:
"Recent papers of Park et al. (2003), Conesa et al. (2006), Di Camillo et al. (2005) and the Limma package by Smyth (2005) have similar approaches"

Should read:
"Recent papers of Park et al. (2003), Di Camillo et al. (2005) and the Limma package by Smyth (2005) have similar approaches"

Page 2, paragraph 3, lines 19-23, which reads:
"An increasing interest in studying the dynamic regulation of gene expression led to new developments in the area of analysis of time-course microarray experiments, see e.g. de Hoon et al. (2002), Bar-Joseph et al. (2003), Bar-Joseph (2004), and more recent approaches by Storey et al. (2005) and Tai and Speed (2006)"

Should read:
"An increasing interest in studying the dynamic regulation of gene expression led to new developments in the area of analysis of time-course microarray experiments, see e.g., de Hoon et al. (2002), Bar-Joseph et al. (2003), Bar-Joseph (2004), and more recent approaches by Storey et al. (2005), Conesa et al. (2006) and Tai and Speed (2006)"

Page 8, case 1 of formula (6) has to be read:

$$\hat{\sigma}^{-M_i} e^{-B/2\hat{\sigma}^2}$$

1 Introduction

Gene expression levels in a given cell can be influenced by different factors, namely pharmacological or medical treatments. The response to a given stimulus is usually different for different genes and may depend on time. One of the goals of modern molecular biology is the high-throughput identification of genes associated with a particular treatment or biological process of interest. The recently developed technology of microarrays allows one to simultaneously monitor the expression levels of thousands of genes. Although microarray experiments can be designed to study different factors of interest, in this paper, for simplicity, we consider experiments involving comparisons between two biological conditions (for example, control and treatment) made over the course of time.

In what follows, we consider data consisting of measurements of differences in the expression levels between "treated" and "control" samples of N genes at n time points in the interval $[0, T]$. The objective is first to identify the genes that respond to the treatment and then to estimate the type of response. This experimental setup can be easily realized by the direct hybridization of two samples on cDNA microarrays and then repetition of the hybridization process at different time points after the treatment.

In general, the problem can be formulated as follows. Consider data consisting of the records on N genes. The record on each gene is taken at n time points $t^{(j)}$, $j = 1, .., n$, and for a gene i at a time point $t^{(j)}$ there are $k_i^{(j)}$ records available, making the total number of records for gene i to be

$$M_i = \sum_{j=1}^{n} k_i^{(j)}. \tag{1}$$

Each record can be modeled as a noisy measurement of a function $s_i(t)$ at a time point $t^{(j)}$. The number of time points is relatively small ($n \approx 10$) and very few replications are available at each time point ($k_i^{(j)} = 0, 1, \dots K$ where $K = 1$, 2 or 3) while the number of genes is very large ($N \approx 10,000$). The objective of the analysis is to identify and estimate the curves that are different from the identical zero (i.e., *significant*). Subsequently, the curves may undergo some kind of clustering in order to group genes on the basis of their type of response to the treatment.

Currently, the statistical literature mostly addresses static microarray experiments, see e.g. Efron *et al.* (2001), Lonnstedt and Speed (2002), Dudoit *et al.* (2002), Kerr *et al.* (2002), Ishwaran and Rao (2003), among many others. Although time-series microarray experiments have lately appeared in

literature, there is still a shortage of statistical methods that are designed specifically for time-course experiments. For example, SAM version 3.0 software package (originally proposed by Tusher *et al.* (2001) and later described in Storey *et al.* (2003)) was adapted to handle time-course data by considering the time points as different groups. In a similar manner, the ANOVA approach by Kerr *et al.* (2000) and Wu *et al.* (2003) was applied to time-course experiments by treating the time variable as a particular experimental factor. Recent papers of Park *et al.* (2003), Conesa *et al.* (2006), Di Camillo *et al.* (2005) and the Limma package by Smyth (2005) have similar approaches. The above methods have the shortcoming of applying statistical techniques designed for static data to time-course data, so that the results are invariant under permutation of the time points. The biological temporal structure of data is ignored.

On the other hand, typical microarray experiments present some technical difficulties such as small number of observations (the time series are usually very short and hence asymptotic methods cannot be used), non-uniform sampling intervals and missing or multiple data, that make them unsuitable to classical time-series and signal processing algorithms.

An increasing interest in studying the dynamic regulation of gene expression led to new developments in the area of analysis of time-course microarray experiments, see e.g. de Hoon *et al.* (2002), Bar-Joseph *et al.* (2003), Bar-Joseph (2004), and more recent approaches by Storey *et al.* (2005) and Tai and Speed (2006).

The goal of the present paper is to develop a statistical methodology specifically designed for time-course microarray data with a new, fully Bayesian approach. The method treats records as functional data, thus preserving temporal structure and taking into account the temporal nature of the data. Another advantage is that the number of records M_i for each gene is not required to be equal for all genes, thus avoiding the tiresome problem of missing data. Moreover, no adjustment is necessary even if records for each gene are taken at different time points.

Since the response curve for each gene is relatively simple and only a few measurements for each gene are available, each $s_i(t)$ curve is globally estimated by expanding it over an orthogonal basis (Legendre polynomials or Fourier). Therefore, each function is described by a vector of coefficients. This is, in fact, similar to Storey *et al.* (2005), where each of the response curves is expanded over the polynomial or B-spline basis with the coefficients estimated by the least squares procedure and the number of basis functions used in these expansions is the same for all genes. By contrast, we propose a Bayesian approach for the simultaneous estimation of response curves as well

as for testing their significance and ranking them accordingly. As a result, the technique is more uniform and flexible than that of Storey *et al.* (2005): it allows a different number of basis functions for each curve (which improves the fits), it does not require one to pre-determine the most significant genes to select the dimension of the fit and avoids a somewhat ad-hoc evaluation of the p-values. Also, our method can accommodate various types of error distributions, namely, all scale mixtures of a normal distribution (e.g., normal, Student T and double-exponential) as well as any additional prior information. By avoiding a non-parametric treatment of errors in the model as in Storey *et al.* (2005), we avoid resampling methods, which may be rather formidable to a practitioner. In fact, all the formulae used in Bayesian computations are explicit and easy to implement.

The Tai and Speed (2006) algorithm is also based on the Bayesian paradigm, although our methodology is very different from theirs. Tai and Speed (2006) apply Bayesian techniques directly to the vectors of observations. For this reason, analysis can be performed only if the same number of replicates is present at any time point and the results of analysis are completely independent of time measurements. Also, unlike Storey *et al.* (2005) and the present paper, Tai and Speed (2006) only rank the genes without providing the cut off point to determine which genes are significant.

The method closest to ours is the Bayesian clustering technique of Heard *et al.* (2006), where the gene profiles are also represented by expansions over a certain basis and the normal-inverse gamma prior is imposed on the unknown coefficients. The number of clusters as well as cluster membership are as well treated as random variables leading to a fully Bayesian model, as in our paper. However, the goal of Heard *et al.* (2006) is different from ours, and they do not address many of the issues that we treat in depth (replications, different time points for different genes, etc.). Nevertheless, the philosophical similarity between the two approaches makes the technique of Heard *et al.* (2006) an attractive and natural choice for subsequent clustering of the curves that are found to be significant by our algorithm.

In what follows, we consider data sets containing measurements of the differences in the expression levels between the "treated" and the "control" samples of N genes at n time points in the interval $[0, T]$. This is the so-called "one sample" problem, in contrast to the situation where the expression levels of the "treated" and the "control" samples are recorded separately. If the design points for both samples coincide (as it is required in Tai and Speed (2006)), then our technique can be easily extended to the "two sample" case by calculating the differences of observations. Potentially our technique can be extended to the case when design points of the "treated" and the "control"

samples are completely different. However, this will be a topic for future research.

Although in the present paper we refer to time-course microarray experiments, the method can be applied to any experimental design with a quantitative factor. For example, by replacing the time variable with a dose variable, one can apply our technique to dose-response studies (provided the number of different doses is relatively large).

The rest of the paper is organized as follows. In Section 2 we describe the hierarchical Bayesian model. Sections 2.2 and 2.3 describe modeling the gene expression profiles and errors. Section 2.4 explains how to estimate the gene-dependent parameters. Section 2.5 describes hypothesis testing while Section 2.6 outlines the procedure for estimating the gene profiles. To complete the methodology, Section 2.7 provides the techniques for estimating global parameters. Finally, Section 2.8 summarizes the complete algorithm. In Section 3 the performance of the proposed method is evaluated using simulated and real data and then compared with the recent competitive methods of Storey *et al.* (2005) and Tai and Speed (2006). Section 4 concludes the paper and the Appendix contains the derivation of the formulae in Section 2.

2 Statistical modeling, estimation and testing of gene expression profiles

2.1 The data structure

The experiment that motivated the proposed methodology consisted of a series of two-color cDNA microarrays where the control (untreated) sample was compared with treated samples after various time intervals upon treatment. The expression value of each microarray was the result of a competitive hybridization. The mRNA samples were reverse-transcripted into cDNA, one sample was labeled with a green (Cy3) and the other with a red (Cy5) fluorescent dye, then they were mixed and applied to the microarray. After the cDNA had hybridized, the microarray image was captured using a scanner and the intensities in the two channels were measured. For each spot, the relative expression value is measured as the $\log_2$ red to green fluorescence intensity ratio. The data are assumed to be already pre-processed to remove systematic sources of variation. For a detailed discussion of the normalization procedures for microarray data we refer the reader to e.g. Yang *et al.* (2002), Cui *et al.* (2002), McLachlan *et al.* (2004) or Wit and McClure (2004).

The measurements are taken at n different time points in $[0, T]$ where

the sampling grid $t^{(1)}, t^{(2)}, \ldots, t^{(n)}$ is not necessarily uniformly spaced. For each array, the measurements consist of N normalized $\log_2$-ratios $z_i^{j,k}$, where $i = 1, \ldots, N$, is the gene number, index j corresponds to the time point $t^{(j)}$ and $k = 1, \ldots, k_i^{(j)}$, $k_i^{(j)} \geq 1$, accommodates for possible technical replicates at time $t^{(j)}$. Note that usually, by the structure of the experimental design, $k_i^{(j)}$ are independent of i, i.e. $k_i^{(j)} \equiv k^{(j)}$ and $M = \sum k^{(j)}$. However, since some observations may be missing due to technical errors in the experiment, we let $k_i^{(j)}$ to depend on i.

For each gene i, we assume that evolution in time of its relative expression is governed by a function $s_i(t)$ and each of the measurements involves some measurement error, i.e.

$$z_i^{j,k} = s_i(t^{(j)}) + \zeta_i^{j,k}, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, n, \quad k = 1, \ldots, k_i^{(j)}. \tag{2}$$

The measurement errors $\zeta_i^{j,k}$ are assumed to be i.i.d. with zero mean and finite variance. The function $s_i(t)$ represents the temporal differential expression level of gene i over the interval $[0, T]$ and it is the quantity of interest. In particular, $s_i(t) \equiv 0$ means that gene i is not affected by the treatment while $s_i(t) \not\equiv 0$ indicates that gene i changes its biological response due to the treatment. In this case, the value of $s_i(t)$ is a measure of the effect induced by the stimulus, hence its estimation is of great interest to investigate the underlying biology. Therefore, the objective of the analysis is to identify the genes for which the hypothesis $s_i(t) \equiv 0$ can be rejected (those genes are called *significant*) and to estimate their expression profiles $s_i(t)$.

2.2 Modeling the gene expression profiles

Each function $s_i(t)$ is globally estimated, since the measurements are available only at a few time points. Specifically, we expand each function over some standard orthonormal basis on $[0, T]$

$$s_i(t) = \sum_{l=0}^{L_i} c_i^{(l)} \phi_l(t) \tag{3}$$

and characterize each of them by the vector of its coefficients $\mathbf{c}_i$. In the present paper we use Legendre polynomials or Fourier basis suitably rescaled and normalized in $[0, T]$, but other choices are possible. Due to the fact that functions $s_i(t)$ model a biological system, these functions are continuous, although discontinuities in the first derivatives are allowed. The values of the coefficients $c_i^{(l)}$ and the degrees of the polynomials L_i are estimated from the observations via a Bayesian approach.

We assume that the genes are conditionally independent, so that combination of (2) and (3) yields

$$\mathbf{z}_i = \mathbf{D}_i \mathbf{c}_i + \boldsymbol{\zeta}_i \tag{4}$$

where $\mathbf{z}_i = (z_i^{1,1} \dots z_i^{1,k_1}, \cdots, z_i^{n,1}, \dots z_i^{n,k_n})^T \in R^{M_i}$ is the column vector of all measurements for gene i (see (1)), $\mathbf{c}_i = (c_i^0, \dots, c_i^{L_i})^T \in R^{L_i+1}$ is the column vector of the coefficients of $s_i(t)$ in the chosen basis, $\boldsymbol{\zeta}_i = (\zeta_i^{1,1}, \dots, \zeta_i^{1,k_1}, \cdots, \zeta_i^{n,1}, \dots, \zeta_i^{n,k_n})^T \in R^{M_i}$ is the column vector of random errors and $\mathbf{D}_i$ is the $M_i \times (L_i + 1)$ block design matrix the j-row of which is the block vector $[\phi_0(t_j)\ \phi_1(t_j)\ \dots\ \phi_{L_i}(t_j)]$ replicated k_j^i times.

The proposed model is fully Bayesian, since we treat all parameters either as random variables or as nuisance parameters, thus recovered from data. We assume that given σ^2, the vector of errors $\boldsymbol{\zeta}_i$ is normally distributed $\boldsymbol{\zeta}_i \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{M_i})$, hence

$$\mathbf{z}_i \mid L_i, \mathbf{c}_i, \sigma^2 \quad \sim \quad \mathcal{N}(\mathbf{D}_i \mathbf{c}_i, \sigma^2 \mathbf{I}_{M_i}).$$

On the unknown parameters we elicit the following priors:

$$\begin{aligned} L_i &\sim Pois^*(\lambda, L_{\max}), \text{ Poisson with parameter } \lambda \text{ truncated at } L_{\max}\ ; \\ \mathbf{c}_i \mid L_i, \sigma^2 &\sim \pi_0 \delta(0, \dots, 0) + (1 - \pi_0) \mathcal{N}(0, \sigma^2 \tau_i^2 \mathbf{Q}_i^{-1}). \end{aligned}$$

We choose the truncated Poisson distribution $Pois^*(\lambda, L_{\max})$ to model the number of terms in the expansion (3), because polynomials of very low order may not adequately represent functions $s_i(t)$, while large values of L_i lead to higher variances. Parameter λ is proportional to the average degree of the polynomial and $L_{\max}$ refers to the maximal possible degree. The values of both parameters are treated as known constants. In the simulation study for $n = 11$, we chose $\lambda = 9$ and $L_{\max} = 6$, approximatively corresponding to a prior degree of three of the polynomials in (3). In general, λ and $L_{\max}$ should be chosen considering the number of available time points and the nature of the problem. Anyway, simulations show that the results of estimation and testing are quite robust with respect to the choice of λ and $L_{\max}$.

The prior distribution on the vectors of coefficients $\mathbf{c}_i$ is chosen to be the mixture of a point weight at zero and a multivariate normal density with the covariance matrix $\sigma^2 \tau_i^2 \mathbf{Q}_i^{-1}$. This choice reflects the fact that some of the curves are identical zeros and that positive and negative coefficients $c_i^{(l)}$ are equally likely for the others. Parameter π_0 is the prior probability of the treatment not affecting a gene and it is a global parameter estimated from the data. Matrix $\mathbf{Q}_i$ is a diagonal matrix that accounts for the decay of the

coefficients in the chosen basis. If functions $s_i(t)$ are ν_i times continuously differentiable, then coefficients $c_i^{(l)}$ have polynomial decays $c_i^{(l)} \sim (l+1)^{-\nu_i}$ in both Legendre and Fourier bases, which corresponds to $\mathbf{Q}_i^{ll} = (l+1)^{2\nu_i}$. The choice of the gene-dependent parameter ν_i is difficult, especially considering that the amount of data is usually insufficient for a reliable estimation. However, our extensive simulations show that the method's performance is practically independent of the choice of ν_i, thus we choose a single global parameter ν in our further analysis. Note that if no assumption about smoothness is made, then $\nu = 0$ and $\mathbf{Q}_i = \mathbf{I}$.

2.3 Modeling the errors

The scale of coefficients $\mathbf{c}_i$ can be different for different genes (the more the gene is affected by the treatment, the larger the value of τ_i). We model this by scaling the covariance matrix for the i-th gene by $\sigma^2\tau_i^2$. Parameters τ_i are estimated from observations and parameter σ^2 is assumed to be a random variable

$$\sigma^2 \quad \sim \quad \rho(\sigma^2).$$

The latter choice allows one to account for possibly non-Gaussian errors (quite common in microarray experiments), without sacrificing closed form expressions for estimators and test statistics. In particular, among the possible choices, we consider three types of priors $\rho(\cdot)$:

case 1: $\rho(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$, the point mass at σ_0^2. The marginal distribution of the error is normal.

case 2: $\rho(\sigma^2) = IG(\gamma, b)$, the Inverse Gamma distribution. The marginal distribution of the error is Student T.

case 3: $\rho(\sigma^2) = c_\mu \sigma^{M_i - 1} e^{-\sigma^2\mu/2}$. The marginal distribution of the error is double exponential.

The global hyperparameters, π_0 and the $\rho(\sigma^2)$ specific parameters (σ_0^2 for case 1, γ and b for case 2 and μ for case 3), are estimated from data. Possible strategies how to do this are discussed in Section 2.7. Once the hyperparameters are estimated, a Bayesian analysis is carried out by combining the prior information and the sample information into the posterior distribution.

2.4 Estimation of gene dependent parameters

If the global parameters of the model were known, one could proceed to a gene-by-gene analysis of coefficients $\mathbf{c}_i,\ i = 1, \cdots, N$. In this section, we only provide the final formulae, referring the reader to the Appendix for the calculation details. To deal with different choices of $\rho(\sigma^2)$ we introduce a function

$$F(A,B) = \int_0^\infty \sigma^{-A} e^{-B/2\sigma^2} \rho(\sigma^2) d\sigma^2 \tag{5}$$

that can be explicitly calculated in the three cases discussed above as:

$$F(M_i,B) = \begin{cases} \hat{\sigma}^{M_i} e^{-B/2\hat{\sigma}^2} & \text{in case 1,} \\ \frac{\Gamma(\frac{M_i}{2}+\gamma)}{\Gamma(\gamma)}\ b^{-\frac{M_i}{2}} (1+\frac{B}{2b})^{-(\frac{M_i}{2}+\gamma)} & \text{in case 2,} \\ c_\mu \sqrt{2\pi/\mu}\ e^{-\sqrt{B\mu}} & \text{in case 3.} \end{cases} \tag{6}$$

We also denote

$$g_\lambda(L_i) = \left[\sum_{l=0}^{L_{\max}} (l!)^{-1} \lambda^l e^{-\lambda}\right]^{-1} (L_i!)^{-1} \lambda^{L_i} e^{-\lambda}, \quad L_i = 0, \ldots, L_{\max}, \tag{7}$$

$$H_i(\mathbf{z}_i) = \mathbf{z}_i^T \mathbf{z}_i - \mathbf{z}_i^T \mathbf{D}_i (\mathbf{D}_i^T \mathbf{D}_i + \tau_i^{-2} \mathbf{Q}_i)^{-1} \mathbf{D}_i^T \mathbf{z}_i, \tag{8}$$

$$V(\mathbf{z}_i, L_i, M_i) = |\tau_i^2 \mathbf{D}_i^T \mathbf{D}_i + \mathbf{Q}_i|^{-1/2} \left[(L_i+1)!\right]^\nu\ F(M_i, H_i(\mathbf{z}_i)). \tag{9}$$

Here $g_\lambda(L_i)$ is the pdf of the truncated Poisson distribution. In the following, we suppress dependence on λ, τ_i and ν and the marginal density of $\mathbf{z}_i$ is of the form

$$p(\mathbf{z}_i) = (2\pi)^{-M_i/2} \left[\pi_0 F(M_i, \mathbf{z}_i^T \mathbf{z}_i) + (1-\pi_0) \sum_{L_i=0}^{L_{\max}} g_\lambda(L_i)\, V(\mathbf{z}_i, L_i, M_i)\right]. \tag{10}$$

Expression (10) contains a gene-dependent parameter τ_i estimated by $\hat{\tau}_i = \arg\max_{\tau_i} p(\mathbf{z}_i)$. The posterior pdf of the degree L_i given data $\mathbf{z}_i$ is calculated as

$$p(L_i|\mathbf{z}_i) = (2\pi)^{-M_i/2} g_\lambda(L_i) \left[\pi_0 F(M_i, \mathbf{z}_i^T \mathbf{z}_i) + (1-\pi_0) V(\mathbf{z}_i, L_i, M_i)\right] \Big/ p(\mathbf{z}_i). \tag{11}$$

For each gene i, we estimate L_i either by maximizing the posterior pdf (11) (MAP principle) or by using its posterior mean. After τ_i and L_i are estimated, we replace them with $\hat{\tau}_i$ and $\hat{L}_i$ in all the subsequent calculations.

Hence, the posterior pdf of $\mathbf{c}_i$ given $\mathbf{z}_i$ and $\hat{L}_i$ becomes

$$p(\mathbf{c}_i|\mathbf{z}_i,\hat{L}_i) = [(2\pi)^{\frac{M_i}{2}} p(\mathbf{z}_i|\hat{L}_i)]^{-1} [\ \pi_0 F(M_i,(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i))\delta(\mathbf{0}) + \frac{(1-\pi_0)[(\hat{L}_i+1)!]^\nu}{(2\pi\hat{\tau}_i^2)^{(\hat{L}_i+1)/2}} F(M_i+\hat{L}_i+1,(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i)+\hat{\tau}_i^{-2}\mathbf{c}_i^T\mathbf{Q}_i\mathbf{c}_i)\ ] \tag{12}$$

where $(2\pi)^{\frac{M_i}{2}} p(\mathbf{z}_i|\hat{L}_i) = \pi_0 F(M_i,\mathbf{z}_i^T\mathbf{z}_i) + (1-\pi_0)V(\mathbf{z}_i,L_i,M_i)$ and $\mathbf{0} = (0,\cdots,0)$.

2.5 Identifying the significant genes

Our main goal is now to test the hypotheses $H_{0i}:\ \mathbf{c}_i = 0$ versus $H_{1i}:\ \mathbf{c}_i \neq 0$ for $i = 1,\ldots,N$. For this purpose, we introduce the Bayes factors BF_i, the quotient between the posterior odds ratio and the prior odds ratio (see e.g. Berger (1985))

$$BF_i = \frac{\sqrt{|\mathbf{Q}_i + \hat{\tau}_i^2\mathbf{D}_i^T\mathbf{D}_i|}}{[(\hat{L}_i+1)!]^\nu}\ \frac{F(M_i,\mathbf{z}_i^T\mathbf{z}_i)}{F(M_i,H_i(\mathbf{z}_i))}. \tag{13}$$

Then the posterior probability that $\mathbf{c}_i = \mathbf{0}$ can be expressed as

$$p(\mathbf{c}_i = \mathbf{0}|\mathbf{z}_i,\hat{L}_i) = \pi_0 \Big/ \left[\pi_0 + (1-\pi_0)(BF_i)^{-1}\right]. \tag{14}$$

Note that, although Bayes factors BF_i can be used for independent testing of the null hypotheses H_{0i}, $i = 1,\ldots,N$, the classical Bayesian approach does not account for the multiplicity of comparisons. Nevertheless, when N is large as in microarray experiments, the problem of multiplicity cannot be ignored. To take into account multiplicity and control the familywise error, we apply the Bayesian multiple testing procedure of Abramovich and Angelini (2006). A simple hierarchical prior model is obtained by imposing a prior distribution $\pi(r) > 0$, $r = 0,1,2,\cdots$, on the number r of alternative hypotheses (the number of significant genes in our case). Afterwards, a decision is made by finding the most likely configuration of null and alternative hypotheses.

Of particular interest in the microarray context are the "sparse" priors $\pi(r)$ that force $E(r)$ to be relatively small with respect to N. They allow to model the prior belief that usually a relatively small number on the total of genes are differentially expressed. The number of true alternatives r can be estimated by the global maximum of the posterior likelihood or by the step-up or the step-down procedure (see Abramovich and Angelini (2006)). In this paper, we use the step-up procedure implemented as follows. Bayes factors are ordered so that $BF_{(1)} \leq BF_{(2)} \leq \ldots \leq BF_{(N)}$ and the corresponding hypotheses are

re-indexed. After that, one starts from the most plausible null hypothesis $BF_{(N)}$ and continues accepting the null hypotheses as long as

$$BF_{(i)} > \frac{i}{N-i+1}\ \frac{\pi(i)}{\pi(i-1)}. \tag{15}$$

All the remaining hypotheses are rejected and the corresponding genes are called significant. In a similar manner, one can also implement the step-down procedure. Note that if the prior $\pi(i)$ in (15) is Binomial with parameter α, then both the step-down and step-up procedures provide the same answer as the global maximization procedure.

For the prior $\pi(r)$ in (15) one can choose, for example, a Binomial $B(N,\alpha)$ prior, a truncated Poisson $Poiss^*(\alpha, N)$ prior or any other "sparse" pdf suggested by the biological knowledge. We observe that the smaller the parameter α is, the more sparse the prior $\pi(r)$ is and the fewer genes are chosen a-priori as differentially expressed. For example, if $\pi(r) = B(N,\alpha)$, then (15) becomes

$$BF_{(i)} > \alpha/(1-\alpha),$$

showing clearly that smaller α will imply a stronger control of the multiplicity. A similar reasoning applies to $Poiss^*(\alpha, N)$ prior as well. In principle α can be estimated from data. For simplicity, in our simulations we use the Binomial prior with $\alpha = 1 - \hat{\pi}_0$.

Remark 1. The use of the Bayesian multiple testing procedure of Abramovich and Angelini (2006) allows one to identify significant genes, to rank them and to estimate their expression profiles all in one unified Bayesian paradigm. However, one can use only parts of the above proposed method to select the significant genes or to estimate the expression profile of genes selected using another procedure.

Remark 2. A more accurate evaluation of Bayes factors would be based on averaging over L_i and calculating $p(\mathbf{c}_i = \mathbf{0}|\mathbf{z}_i)$ instead of $p(\mathbf{c}_i = \mathbf{0}|\mathbf{z}_i, \hat{L}_i)$ in (14). However, it would lead to a significant increase in the amount of computations and this is the reason why we replace the true values L_i with the estimators $\hat{L}_i$. In any case, replacing the more reliable average between the models with a plug-in estimator obtained upon maximization of the posterior likelihood is an accepted procedure in Bayesian computations (see e.g. Chipman et al. (2001) or Burnham and Anderson (2002)).

2.6 Estimation of gene expression profiles

Finally, we estimate the coefficients $\mathbf{c}_i$ for the significant genes using the posterior expectation over (12)

$$\hat{\mathbf{c}}_i = \frac{(1-\pi_0)/\pi_0}{BF_i + (1-\pi_0)/\pi_0} \left(\mathbf{D}_i^T \mathbf{D}_i + \mathbf{Q}_i/\hat{\tau}_i^2\right)^{-1} \mathbf{D}_i^T \mathbf{z}_i,$$

and, from (3), the estimator of $s_i(t)$ is

$$\hat{s}_i(t) = \begin{cases} \sum_{l=0}^{\hat{L}_i} \hat{\mathbf{c}}_i^{(l)} \phi_l(t) & \text{if} \quad H_0 \quad \text{is rejected,} \\ 0 & \text{if} \quad H_0 \quad \text{is accepted.} \end{cases} \tag{16}$$

2.7 Estimation of global parameters

To complete the theory, we need to address the problem of estimating the global parameters of the model: the common parameter π_0 and the case-specific parameters σ_0^2 (for case 1), γ and b (for case 2) or μ (for case 3).

Several options are available in literature to accomplish this task, see for example Efron *et al.* (2001), Ishwaran and Rao (2003), Storey and Tibshirani (2003), Pounds and Cheng (2004), Pawitan *et al.* (2005). In this section, we describe the methods we used in simulation and in real data analysis. However, we note that a different set of methods can be applied without changing the main algorithm.

First we observe that, according to our model, some genes may have missing observations, since a few values may be filtered out during preprocessing. Estimation of the global parameters is then based only on the N_c genes for which the complete set of M observations is available. Therefore, we have M arrays with N_c observations at each time point $t^{(j)}$, $j = 1, \ldots, M$. Note that here we are using a one-index enumeration of time points, so in case multiple observations are made at the same time, then $t^{(j)} = t^{(j+1)} = t^{(j+2)} \ldots$.

For each array of observations at a time point $t^{(j)}$, the standard deviation $\sigma^{(j)}$ is estimated by the sample standard deviation $\hat{\sigma}^{(j)}$ using the sparsity assumption (the majority of array components are zeros). If normality of the data can be justified, the sample variance can be replaced by a more robust estimator like MAD. Note that if a sufficiently large number of arrays is available, then one can also use a gene specific estimator of σ or an unbiased pooled estimator. The estimator $\hat{\sigma}^2$ is obtained by averaging of $(\hat{\sigma}^{(j)})^2$, $j = 1, \ldots, M$.

Given $\hat{\sigma}$, we estimate the global parameter π_0 by averaging over the arrays the proportion of data points which fall below the universal threshold

$\hat{\sigma}\sqrt{2\log N_c}$ (following Donoho (1992)). Note that this method tends to overestimate π_0 when the error is normally distributed. However, this is no longer true when the error distribution has heavier tails, a very common condition in real data. Alternatively, π_0 can be estimated using the empirical Bayes approach of Johnstone and Silverman (2004), whose set up is in concordance with ours. In this case, we obtain an estimator of π_0 for each array and average the results.

The set of estimators $\hat{\sigma}^{(j)}$, $j = 1, \ldots, M$, is subsequently treated as the sample of values of σ and is used for estimating parameters of $\rho(\sigma^2)$. If the prior model chosen is the *case 1*, σ_0^2 is estimated by $\hat{\sigma}^2$. In *case 2*, hyper-parameters γ and b can be estimated by using the MLE (note that if $(\hat{\sigma}^{(j)})^2 \sim IG(\gamma, b)$, then $(\hat{\sigma}^{(j)})^{-2} \sim$ Gamma(γ, b)). An alternative procedure is to fix one of the two parameters, γ or b, and then estimate the other one by matching the mean of $IG(\gamma, b)$ with $\hat{\sigma}^2$. Similarly, in *case 3*, μ can be estimated by $\hat{\mu} = (M_i - 1)/\hat{\sigma}^2$, so that the mean of the prior distribution $\rho(\sigma^2)$ is centered at $\hat{\sigma}^2$.

2.8 Algorithm

In this Section we describe the algorithm for automatic identification and estimation of the gene expression profiles in a time-course microarray experiment. We again point out that the input data in (4) should be pre-processed and normalized to remove systematic sources of variation.

The algorithm can be performed by carrying out the following steps:

1. A preliminary step is to fix prior parameters λ, $L_{\max}$ and ν.

2. Estimate global parameters: σ^2 and π_0, and additional case-specific hyper-parameters σ_0^2 (for case 1), γ and b (for case 2) or μ (for case 3).

3. For each gene i, estimate the gene specific parameter τ_i by maximizing the marginal pdf of the data (10). Subsequently, plug in $\hat{\pi}_0$, $\hat{\sigma}^2$, $\hat{\gamma}$, $\hat{b}$ or $\hat{\mu}$ instead of π_0, σ_0^2, γ, b or μ when required.

4. For each gene i, estimate the most appropriate degree L_i as the mean or the mode of the posterior pdf (11).

5. For each gene i, conditionally on $\hat{L}_i$, compute Bayes Factor BF_i using formula (13).

6. Perform Bayesian multiple testing procedure for controlling the multiplicity error and rank the genes according to the ordered Bayes

factors. For this purpose, order Bayes factors so that $BF_{(1)} \leq BF_{(2)} \leq \ldots \leq BF_{(N)}$, and re-index the corresponding hypotheses. After that, start from the most plausible null hypothesis $BF_{(N)}$ and continue accepting the null hypotheses as long as (15) is true. After that, reject all the remaining hypotheses. All genes corresponding to the rejected hypotheses are declared significant.

The choices of the prior $\pi(r)$ in (15) are discussed in Section 2.5.

7. Estimate the gene expression profiles by $\hat{s}_i(t)$ defined in (16).

3 Evaluations and comparisons

In this section we evaluate the methodology proposed above using simulated and real data sets. We also carry out its comparison with the recent competitive methods by Storey *et al.* (2005) and Tai and Speed (2006).

3.1 Simulations

To investigate the performance of the proposed method, we carried out a simulation study. We generated data according to the model (4) with $N = 8000$, $n = 11$, $k_i^j = 2$ for all $j = 1, \ldots, 11$ except $k_i^{2,5,7} = 3$ to mimic the structure of the real data set described in the next section. We also used the same time points as in the real data set. We randomly chose 600 "significant" curves, simulating their profiles according to (3). The other 7400 curves were set to identical zero. This scenario corresponds to the situation when 7.5 % of the total number of genes are "differentially expressed". For each significant curve, we first sampled the degree of the polynomial L_i^{true} from a discrete uniform distribution in $[1, L_{\max}]$. We avoided polynomials of degree zero since a nonzero constant signal is questionable from a biological point of view. In simulations we selected $L_{\max} = 6$.

After that, we randomly sampled $\mathbf{c}_i$ from $\mathcal{N}(0, \sigma^2\tau_i^2\mathbf{Q}_i^{-1})$ where we chose $\sigma = 0.27$, calculated from the real data set. Matrix $\mathbf{Q}_i$ is set to $\mathbf{Q}_i = \mathrm{diag}(1^{2\nu_i}, 2^{2\nu_i}, ..., {L_i}^{2\nu_i})$ where $\nu_i \sim U([0,1])$ and τ_i^2 was sampled uniformly in order to produce the signal-to-noise ratio (SNR) in the interval $[2, 6]$, mimicking both weak and strong signals. The choice of the sampling interval $[0, 1]$ for the parameter ν_i was motivated by the belief that biological responses to the treatment may depend on the particular gene and that the profiles are continuous or at most differentiable. Note that in the estimation algorithm we chose a common prior parameter ν. However, our method could be modified

to allow gene dependent ν_i, thus paying a price of heavier calculations without much gain in precision.

We performed simulations with three kinds of i.i.d. noise: normal $N(0, \sigma^2)$ and Student T with 5 or 3 degrees of freedom, respectively noted $T(5)$ and $T(3)$. Student noise was rescaled to have the same variance σ^2 of the normal case. In addition, very large values (with a threshold of 5) were filtered out and substituted with "missing values", mimicking real data preprocessing where unrealiable values are eliminated. The number of missing values per gene did not exceed 5, and at most 8% of the profiles were affected.

For each noise distribution, the simulations were repeated with 30 randomly generated sets of profiles s_i. In order to make the results comparable and independent of the particular choice of functions s_i, the same sets of functions were used with all three noise scenarios. For each set of profiles, the experiment was replicated with 10 different noise realizations. The data were processed using the methods proposed in *cases-1-2-3*, and the results were averaged.

Simulations were carried out with various choices of the parameters λ, $L_{\max}$ and ν and the results were robust with respect to these choices. For this reason, Tables 1 and 2 present the results only with $L_{\max} = 6$, $\lambda = 9$ (corresponding to an expected degree of about 3) and $\nu = 0$ or $\nu = 1$, respectively. Tables 1 and 2 report the average number of rejected hypotheses (genes declared differentially expressed) (reje), the average number of the correctly rejected hypotheses (corr), the False Discovery Rate (FDR), computed as the average proportion of the falsely rejected hypotheses over the total number of rejected hypotheses, and the False Negative Rate (FNR), computed as the average proportion of the significant curves which were not detected over the total number of curves declared nonsignificant. Results in the two cases are comparable, although a slightly higher number of curves is usually detected with $\nu = 1$. Simulations show good performance of the procedure in the case of the normal and the T(5) noise, and acceptable performance in the case of heavy-tailed T(3) noise. In addition, the results are robust not only in the number of detected genes but also in their ranking: the ranks of the top 200-300 significant genes assigned by the different methods vary at most by a few positions. Moreover, in the normal and T(5) cases the highly ranked false positives only appear after the first 450 and 400 positions respectively, while in the T(3) case there are almost no false positives among the first 300 genes declared significant.

Simulations reported in the paper used the MAP procedure to estimate the gene-wise degree L_i, although there was no remarkable difference when the same degree was estimated using the posterior mean. Estimation of the hyper-parameters was carried out as described in Section 2.7. In particular, σ was estimated by the sample standard deviation averaged over the arrays.

Table 1: Simulated results for the case $L_{\max} = 6$, $\lambda = 9$, $\nu = 0$.

	N noise				$T(5)$ noise				$T(3)$ noise			
	reje	corr	FDR	FNR	reje	corr	FDR	FNR	reje	corr	FDR	FNR
case-1	488.8	488.8	.00001	.0148	505.0	491.9	.0260	.0144	575.1	502.1	.1270	.0132
case-2-I	483.3	483.3	.00005	.0155	497.4	488.0	.0188	.0149	572.1	504.5	.1182	.0129
case-2-II	485.0	485.0	.00002	.0153	500.7	488.8	.0237	.0148	578.4	501.8	.1325	.0132
case-3	474.6	474.5	.00003	.0167	502.7	481.8	.0416	.0158	621.7	501.3	.1937	.0134

Table 2: Simulated results for the case $L_{\max} = 6$, $\lambda = 9$, $\nu = 1$.

	N noise				$T(5)$ noise				$T(3)$ noise			
	reje	corr	FDR	FNR	reje	corr	FDR	FNR	reje	corr	FDR	FNR
case-1	502.3	502.3	.00004	.0130	514.2	504.7	.0185	.0127	574.3	515.8	.1018	.0113
case-2-I	499.1	499.0	.00011	.0135	511.2	503.7	.0155	.0129	582.4	519.4	.1082	.0109
case-2-II	499.9	499.9	.00004	.0133	513.1	503.1	.0195	.0129	587.6	516.5	.1210	.0113
case-3	491.4	491.4	.00008	.0145	517.6	497.6	.0385	.0137	634.5	516.9	.1854	.0113

In case-2, various strategies to select γ and b were tested, but the tables only report case-2-I (where we fixed $\gamma = 15$ and selected b to make the mean of the prior IG distribution coincide with the estimated $\hat{\sigma}^2$) and case-2-II (where simultaneous estimation of γ and b was performed by the MLE). Note that although the two cases provide different estimates of γ and b, there was very little difference in the detection of significant genes. However, case-2-I may be preferred by an experienced user who wants to use a tuning parameter to slightly adjust the number of selected genes.

The quality of the selection was also evaluated in terms of functional errors in estimating the response curves. In Table 3 we show several types of functional errors associated with our decision, namely

$$errA = ||s_i - \hat{s}_i||_2^2 \Big/ ||s_i||_2^2$$

is the relative L_2 error averaged over the functions s_i correctly declared significant and estimated by $\hat{s}_i$;

$$errB = ||\hat{s}_i||_2^2$$

is the absolute L_2 error averaged over false positive functions for which $s_i = 0$

Table 3: Average relative estimation error (errA); Average false positive error (errB); Average false negative error (errC). Results are obtained with $L_{\max} = 6$, $\lambda = 9$, $\nu = 0$.

	N noise			$T(5)$ noise			$T(3)$ noise		
	errA	errB	errC	errA	errB	errC	errA	errB	errC
case-1	.07905	.00086	.46856	.08828	.63378	.45311	.10035	.83435	.43564
case-2-I	.07792	.00235	.49161	.07898	.29420	.47501	.08428	.32265	.44137
case-2-II	.07799	.00087	.48094	.08162	.45795	.46422	.09078	.51321	.43956
case-3	.07590	.00141	.51823	.08126	.26061	.49212	.08821	.32450	.45070

but s_i is incorrectly declared significant and estimated with $\hat{s}_i$; finally,

$$errC = ||s_i||_2^2$$

is the absolute L_2 error averaged over false negative functions which are not detected as significant in spite of $s_i \neq 0$. Note that all three methods for all three noise scenarios have comparable values of errors $errA$ and $errC$. The only exception is $errB$ which is significantly larger when the noise has heavier tails since in this case it is much easier to confuse noise with signal. Note also that results of case-2-II are comparable with those of case-1 and are quite different from those of case-2-I. The reason for this is that the MLE of parameters γ and b are higher than in case-2-I. Hence, the T-distribution in case-2-II approximates closely the normal distribution of case-1 and are quite distant from the T-distribution in case-2-I.

ErrC is comparable for all the methods and represents a sort of detection limit (in L_2-norm), under which the effect of the treatment cannot be detected by the proposed methods. Figure 1 shows the histograms of errA, errB and errC computed on data simulated with T(3) noise and processed by our method (case 1, $\lambda = 9$, $L_{\max} = 6$ and $\nu = 0$). We have to point out that the histograms are drawn for the absolutely worst noise scenario when the actual noise has T(3) distribution and the data is processed as if the noise was Gaussian. In spite of this, the histograms show that the pdfs of the errors are unimodal, centered around small values and have thin tails.

3.2 Real data application

We applied the proposed methods to the time-course microarray study described in Cicatiello *et al.* (2004) (the data are available on the

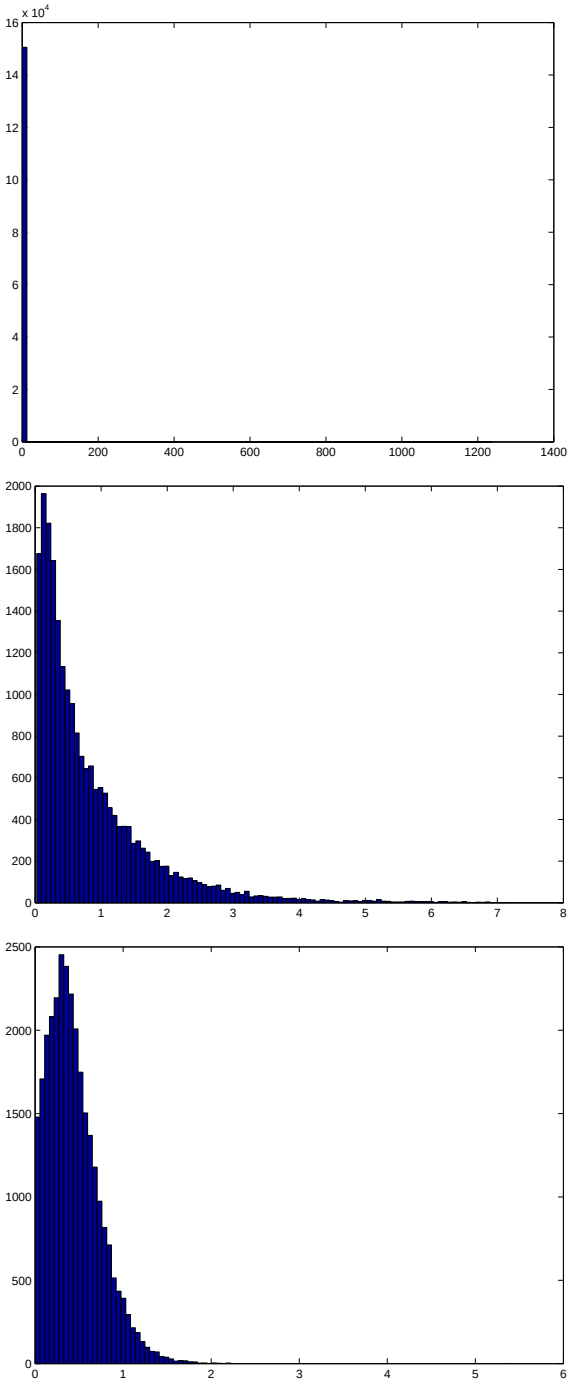

Figure 1: Histograms of the estimation errors, errA, errB and errC computed with the data simulated with T(3) noise and processed according to case 1, $\lambda = 9$, $L_{\max} = 6$, $\nu = 0$.

Table 4: Average relative estimation error (errA); Average false positive error (errB); Average false negative error (errC). Results are obtained with $L_{\max} = 6$, $\lambda = 9$, $\nu = 1$.

	N noise			$T(5)$ noise			$T(3)$ noise		
	errA	errB	errC	errA	errB	errC	errA	errB	errC
case-1	.07536	.00550	.42738	.08300	.71074	.41384	.09344	.90208	.39634
case-2-I	.07524	.01141	.45080	.07695	.25375	.43511	.07620	.26662	.40459
case-2-II	.07497	.00560	.43933	.07798	.42431	.42401	.08429	.43825	.40088
case-3	.07426	.00894	.47344	.07922	.21030	.45074	.07918	.25972	.41210

GEO repository - `http://www.ncbi.nlm.nih.gov/geo/`, accession number GSE1864). The objective of the experiment was to identify genes involved in the estrogen response in a human breast cancer cell line. Estrogen has a known role in promoting cell proliferation and thus in cancer development in hormone-responsive tissues such as breast and ovary. In the original experiment, ZR-75.1 cells were stimulated with a mitogenic dose of 17β-estradiol, after 5 days of starvation on an hormone-free medium, and samples were taken after $t = 1, 2, 4, 6, 8, 12, 16, 20, 24, 28, 32$ hours, with a total of 11 time points covering the completion of a full mitotic cycle in hormone-stimulated cells. For each time point at least two replicates were available (three replicates at $t = 2, 8, 16$).

From the 8400 genes we first removed the genes that did not pass the image analysis quality control flag. Then, we filtered out individual spots with low intensity values in any single channel, red or green ($\min(\log_2(R), \log_2(G)) < 5$), or in both channels ($\log_2(R * G) < 11$), and replaced the corresponding values with a missing. Here R (red) and G (green) are the values measured on the cDNA microarray on the red and green channel, respectively. We also removed a few spots that showed opposite signs and big difference between the expression values of replicates at the same time point. After that, a gene was removed from further investigation if more than 20% of values were missing. As a result, the total number of analyzed genes was 8161 (among them about 350 contained at least one missing value). We then normalized data using the standard lowess normalization procedure with span parameter 0.3 in order to remove various nuisance sources of systematic variation in the measured fluorescence intensities (e.g. different labeling efficiencies and scanning properties of the two fluorochromes or suboptimal choice of the scanning parameters). We refer the reader to Yang *et al.* (2002) and Cui *et al.*

(2002) for a review of normalization procedures for microarrays. We carried out our procedure with various choices of parameters at the preprocessing stage but it made very little difference in the approximation and testing results.

We analyzed this data set using the proposed methods and various choices of the parameters ν and λ. Table 5 shows the number of genes declared affected by the treatment for $L_{\max} = 6$, $\nu = 0$ and λ's ranging between 6 and 12, corresponding to an expected prior degree of polynomials from 2.5 to 3.5. Table 5 shows that the results are quite robust with respect to the number of detected significant genes, with smaller λ providing larger lists ($\lambda > 12$ does not provide any noticeable changes in the list). The technique is also very robust with respect to the list of genes declared significant: 574 genes were common to all 28 lists (combination of different methods and different parameter values) while 958 genes have been selected in at least one of the 28 lists. Moreover, we note a very substantial overlap also in any sublist of genes (for example, for any i, there is an overlap of about 85% in the top i ranked genes of all lists). Note also that our list of 574 common genes includes 270 genes out of the 344 genes identified as significant in Cicatiello *et al.* (2004). Among the remaining 74 genes, 16 were filtered out in our analysis, due to a more stringent selection in the preprocessing stage, and 58 genes were selected by our method with some combinations of priors and parameters but not with all of them. Indeed, the list of 958 contains 309 genes already detected in Cicatiello *et al.* (2004). By examining the raw expression profiles, we found those 58 genes having between weak and very weak responses, compared to the noise. On the other hand, our list contains 304 genes not detected in Cicatiello *et al.* (2004). While looking at the newly selected response curves, we noticed that the raw data show much more variability between replicates than the gene profiles selected in Cicatiello *et al.* (2004). Since the Cicatiello *et al.* (2004) analysis was carried out manually, point by point, the data for those genes was probably considered unreliable and genes were discarded. However, the functional approach which lies at the core of our method allows one to estimate the gene profiles with enough precision even with missing or less reliable individual data points. Figure 2 shows an example of a gene expression profile selected as significant by both our method and Cicatiello *et al.* (2004) and an example of a gene selected by our method but not by Cicatiello *et al.* (2004).

Moreover, interestingly, 17 of the 304 newly selected genes were replicate spots of genes already selected in the Cicatiello *et al.* (2004). Most of them are known to be involved in biological processes related to estrogen response, such as cell cycle and cell proliferation (AREG, NOLC1, cyclin D1), DNA replication (MCM7, RFC5), mRNA processing (SFRS1) and lipid metabolism

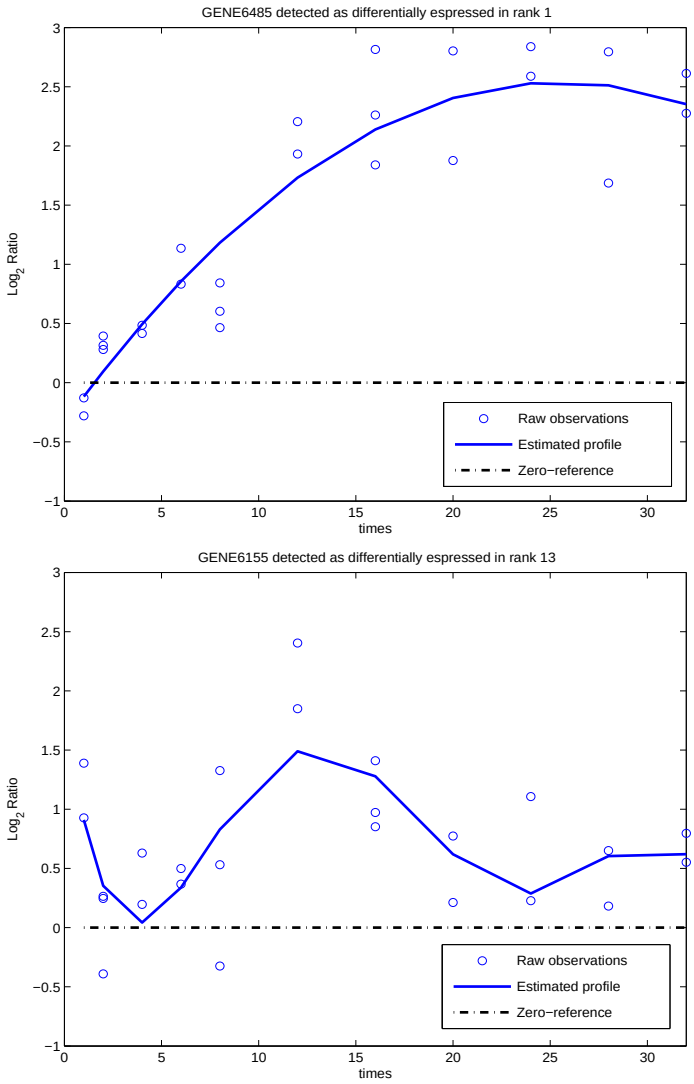

Figure 2: Gene6485 (TFF1, a well-known target of the estrogen receptor) has been selected with rank 1 by our method and included in the list of 574 genes selected by all the 28 combinations. This gene has been detected in Cicatiello *et al.* (2004) and by Tai and Speed (2006) and Storey *et al.* (2005) approaches as well. Gene6155 (MKI67, a gene involved in cell-cycle control but with a less clear association with estrogen action in literature) has been selected with rank 13 by our method and included in the list of 574 genes selected by all the 28 combinations. This gene has not been detected by Cicatiello et. al (2004) or EDGE (with q-value=0.1), while it has been detected by Tai and Speed (2006) with rank 2. Results are obtained for case 1, $\lambda = 9$, $L_{\max} = 6$ and $\nu = 0$.

Table 5: Total number of genes in Cicatiello *et al.* (2004) dataset detected as significant by our method (with $\nu = 0$ and $Ł_{\max} = 6$)

	$\lambda = 6$	$\lambda = 7$	$\lambda = 8$	$\lambda = 9$	$\lambda = 10$	$\lambda = 11$	$\lambda = 12$
case-1	867	808	753	712	692	688	691
case-2-I	893	823	765	711	679	657	650
case-2-II	869	810	755	714	694	690	693
case-3	855	786	726	676	640	617	609

(APOD and LDHA).

The analysis was also repeated on the same data set using $\nu = 1$ and the same range of λ's leading to an expected result of detecting a larger number of significant genes (since simulation results show that $\nu = 1$ is less conservative). The difference of about $100 - 200$ more genes is not negligible, however, the genes selected with $\nu = 0$ were all present in the list of $\nu = 1$. Finally, we also repeated the analysis with different choices at the preprocessing stage (the size of the span parameter of the lowess method, the cut off on the intensities, etc.) and we observed that the results are quite robust to these choices with deviations of only $20 - 40$ genes.

3.3 Comparisons with other methods

In order to further evaluate the performance of our method, we compared it with two recent competitive methods: Storey *et al.* (2005) and Tai and Speed (2006). The first method was implemented by the EDGE software (Leek *et al.* 2006) while the second by the R-package *timecourse*. Since all three methods apply to different experimental designs, account for different biological information and are valid under different assumptions, we felt that it would be more fair to compare our method with the others using a real data set that does not conform to the assumptions in the present paper. For comparisons, we chose the above mentioned Cicatiello *et al.* (2004) dataset since it does not comply with any artificial assumptions. In addition, Cicatiello *et al.* (2004) provides a "biology-guided" selection of significant genes that can be used as a "benchmark" in our comparisons.

We should mention that EDGE was originally designed for the "two-sample" problem following the methodological paper of Storey *et al.* (2005) and afterwards equipped with a special tool to handle the "one-sample" problem. Tai and Speed (2006) approach applies both on the "one-sample"

and the "two-sample" problem for classical longitudinal data where replicates are biologically meaningful.

Since the EDGE software does not automatically account for missing values but only suggests a preliminary procedure (K-nearest-neighbors) for filling them in, we repeated the analysis both using this procedure and filtering out genes with missing values. Additionally, EDGE allows the user to choose the degree of the splines or the polynomials common to all genes. Hence, similarly to Table 5, we carried out the analysis with different choices of λ and we found results were robust with respect to those choices (data not shown). To estimate the distribution of the statistics under the null hypothesis EDGE uses a bootstrap approach, thus requiring a high computational effort and appropriate memory resources. We used 1000 permutations in our comparisons and we discovered that the gene selections were robust with different random seeds (only a few different genes). In order to control the multiplicity error, EDGE uses the q-values, which we chose to be $q = 0.05$ and $q = 0.1$ in our analysis.

The method of Tai and Speed (2006) neither allows missing values nor suggests a specific procedure for treating them. Moreover, it requires that each time point has the same number of replicates (different number of replicates are allowed between different genes). In order to apply the method, we first filtered out all the genes with missing observations and then discarded the third observations which was available at time points $t = 2, 8, 16$. To be fair, we should mention that the method of Tai and Speed (2006) is designed for data where replicates are biologically meaningful. Hence, since Cicatiello *et al.* (2004) dataset contains only technical (indistinguishable) replicates, the method of Tai and Speed (2006) could not take advantage of the replicate identification. On the other hand, the information about the time measurements is not used by their method. Since the method only provides rank-ordered list of genes (without any automatic cut off point), we perform the comparisons taking the top 500 and 1000 genes in their list. Table 6 shows the number of detected genes with different procedures and the overlap with the genes detected as significant in Cicatiello *et al.* (2004).

Table 6 shows that the proposed approach has a noticeably wider overlap with the "biology guided" selection of significant genes of Cicatiello *et al.* (2004). Moreover, most genes selected by EDGE, *timecourse* and Cicatiello *et al.* (2004) were also selected by our method. Indeed, out of the 186 genes selected by EDGE and declared significant in Cicatiello *et al.* (2004), 165 were contained in the list of 574 genes common to all the lists. On the other hand, only 186 out of 767 genes selected by EDGE were present on the Cicatiello *et al.* (2004) list, and among the 500 top-ranked genes by *timecourse* only

Table 6: Total number of genes declared affected by the treatment and overlap with Cicatiello *et al.* (2004)

	Selected genes	Overlap
All of the 28 methods in Table 5	574	270
At least one of the 28 methods in Table 5	958	309
Case 1, $\lambda = 9$ in Table 5 (default choice)	712	295
EDGE with default choices and q=0.05	767	186
EDGE with default choices and q=0.1	1178	219
Timecourse	500	174
Timecourse	1000	215

174 were detected as significant. Note that out of 174 genes selected by the R-package *timecourse*, 166 were present in the list of 574 genes common to all methods. Finally, 138 genes were common to all selections (Cicatiello *et al.* (2004), all versions of our method, EDGE and *timecourse*).

The comparisons show that in the case of analysis of the Cicatiello *et al.* (2004) data, the algorithm proposed above provides results which are much closer to a "biologist's choice" and delivers a lower percentage of false positive and negative answers than the competitive algorithms.

Finally, we carried out a limited simulation study to compare performances of EDGE, *timecourse* and our method. To this purpose, with the same experimental structure and parameters' choice as in Section 3.1, we generated data sets with 8000 genes among which 600 or 1500 were significant and with the same kinds of noise (normal, $T(5)$ and $T(3)$ rescaled to have $\sigma = 0.27$). In order to apply EDGE, missing data were filled in using the K-nearest-neighbors algorithm. For the *timecourse* R-package, the records containing missing values were removed and only the first two replicates were used (note that at most 19 record were removed from the analysis). Moreover, since *timecourse* does not provide an automatic cut off point and only provides a ranked list, for the sake of comparison we used the same number of significant genes as in our method. Also, similarly to the case of analysis of Cicatiello *et al.* (2004) data set we used EDGE with $q = 0.05$ and $q = 0.1$ and our technique was applied with $\lambda = 9$ and normal noise model. The following tables report the number of rejected genes (reje) and the number of correctly rejected genes (corr), averaged over 5 simulated data sets. The results show that our method provides more accurate results than its competitors.

Table 7: Total number of genes declared significant and the number of correctly rejected genes (600 significant out of 8000 genes)

	Normal		$T(5)$		$T(3)$	
	reje	corr	reje	corr	reje	corr
Our method (case 1, $\lambda = 9$)	491.6	491.6	497.4	485.0	580.8	506.4
EDGE with $q = 0.05$	409.8	377.2	416.8	379.0	452.8	403.8
EDGE with $q = 0.1$	354.2	343.6	359.2	345.4	387.6	372.6
Timecourse	491.6	453.0	495.4	414.2	580.8	453.0

Table 8: Total number of genes declared significant and the number of correctly rejected genes (1500 significant out of 8000 genes)

	Normal		$T(5)$		$T(3)$	
	reje	corr	reje	corr	reje	corr
Our method (case 1, $\lambda = 9$)	1110	1110	1096	1093	1152	1129
EDGE with $q = 0.05$	1199	1085	1183	1077	1232	1120
EDGE with $q = 0.1$	1050	1002	1025	992	1093	1045
Timecourse	1110.4	1099	1096	1025	1132	1076

4 Discussion

In this paper we present a fully Bayesian approach for detecting differentially expressed genes in a time-course experiment. The proposed method contributes to the new and increasingly popular research area of the analysis of time-course microarray data (see Figure 1 of Ernst *et al.* (2005) for examples of experiments in which the proposed procedure may be helpful). To the best of our knowledge, our approach is the first functional fully Bayesian procedure available in literature for such kind of problem. Our method can also be complemented with a somewhat similar Bayesian approach for the cluster analysis of gene profiles proposed by Heard *et al.* (2006). Indeed, any clustering procedure may benefit of a preliminary step where differentially expressed genes are selected.

The Bayesian formulation allows one to explicitly use the prior information that biologists may provide and successfully deals with various technical difficulties that arise in microarray time-course experiments such as a small number of observations, non-uniform sampling intervals, missing or multiple

data and temporal dependence between observations for each gene. Moreover, the method can accommodate a wide variety of errors, thus avoiding the two undesirable extreme cases: treating the error distribution as completely unknown, which leads to rather expensive computational procedures (as e.g. in Storey *et al.* (2005)), or assuming that the errors are normally distributed, which is unrealistic. As a result, all evaluations are based on explicit expressions, thus leading to fast and simple computational procedures that are attractive to a practitioner. In addition, since the computational cost of the method is relatively small, fine tuning of the prior parameters can be easily done.

The proposed procedure was evaluated using simulated and real microarray data provided by Cicatiello *et al.* (2004), describing the estrogen response in a human breast cancer cell line. It was also compared to the recent competitive methods by Storey *et al.* (2005) and Tai and Speed (2006) using the same data set and the "biology-guided" selection of significant genes by Cicatiello *et al.* (2004). The comparison shows that for the Cicatiello *et al.* (2004) data, the proposed algorithm provides results much closer to a "biologist's choice" and delivers a lower percentage of false positive and negative answers than the competitive algorithms. Moreover, application of our technique allows one to estimate the profiles for those genes with sufficient precision even in the presence of missing or less reliable individual data points.

In addition, in the course of our study, we detected a number of genes that resulted strongly affected by the treatment but were discarded in Cicatiello *et al.* (2004) because data on those genes were considered unreliable.

The pre-processed data and the Matlab routines used for carrying out simulations and analysis of real data are available upon request from the first two authors. A software package BATS that implements the methodology described in the paper is under preparation.

5 Appendix

5.1 Derivation of $p(L_i|\mathbf{z}_i)$

Combination of the model and the priors leads to the following joint pdf

$$\begin{aligned} p(\mathbf{z}_i, \mathbf{c}_i, L_i, \sigma^2) &= g_\lambda(L_i)\rho(\sigma^2)(2\pi)^{-M_i/2}\sigma^{-M_i}\exp\left\{-\frac{(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i-\mathbf{D}_i\mathbf{c}_i)}{2\sigma^2}\right\} \\ &\quad \left[\pi_0\delta(0,\ldots,0) + (1-\pi_0)\frac{\sqrt{|\mathbf{Q}_i|}}{(2\pi\sigma^2\tau_i^2)^{(L_i+1)/2}}\exp\left(-\frac{\mathbf{c}_i^T\mathbf{Q}_i\mathbf{c}_i}{2\sigma^2\tau_i^2}\right)\right]. \end{aligned} \tag{17}$$

Using equality

$$(\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i) + \tau_i^{-2}\mathbf{c}_i^T\mathbf{Q}_i\mathbf{c}_i = \\ \left[\mathbf{c}_i - (\mathbf{D}_i^T\mathbf{D}_i + \tau_i^{-2}\mathbf{Q}_i)^{-1}\mathbf{D}_i^T\mathbf{z}_i\right]^T \left[\mathbf{c}_i - (\mathbf{D}_i^T\mathbf{D}_i + \tau_i^{-2}\mathbf{Q}_i)^{-1}\mathbf{D}_i^T\mathbf{z}_i\right] + H_i(\mathbf{z}_i),$$

where $H_i(\mathbf{z}_i)$ is defined in (8), we integrate out $\mathbf{c}_i$:

$$\begin{aligned} p(\mathbf{z}_i, L_i, \sigma^2) \quad &= g_\lambda(L_i)\rho(\sigma^2)(2\pi)^{-M_i/2}\sigma^{-M_i}\left[\ \pi_0 \exp\left(-\tfrac{\mathbf{z}_i{}^T\mathbf{z}_i}{2\sigma^2}\right) + \right. \\ &\left. \tfrac{(1-\pi_0)[(L_i+1)!]^\nu}{(\tau_i^2)^{(L_i+1)/2}\sqrt{|\mathbf{D}_i^T\mathbf{D}_i + \frac{\mathbf{Q}_i}{\tau_i^2}|}} \exp\left(-\tfrac{H_i(\mathbf{z}_i)}{2\sigma^2}\right)\ \right]. \end{aligned} \tag{18}$$

Integrating out σ^2 and using definition of F given in (5) we obtain

$$p(\mathbf{z}_i, L_i) = (2\pi)^{-M_i/2}\, g_\lambda(L_i)\ \left[\pi_0 F(M_i, \mathbf{z}_i^T\mathbf{z}_i) + (1-\pi_0)\ V(\mathbf{z}_i, L_i, M_i)\right]. \tag{19}$$

Summing over all possible degrees $L_i = 0, \ldots, L_{\max}$, we derive the marginal pdf of the data (10). The posterior pdf (11) of the degree L_i can be obtained by dividing (19) by (10).

5.2 Estimation of $\mathbf{c}_i$

Plugging an estimate $\hat{L}_i$ in (17) and dividing it by $g_\lambda(\hat{L}_i)$ we obtain $p(\mathbf{z}_i, \mathbf{c}_i, \sigma^2|\hat{L}_i)$. Integrating out σ^2, we derive

$$p(\mathbf{z}_i, \mathbf{c}_i|\hat{L}_i) = (2\pi)^{-M_i/2}\left[\pi_0 F(M_i, (\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i))\delta(0, \ldots, 0) + \right. \\ \left. \tfrac{(1-\pi_0)[(\hat{L}_i+1)!]^\nu}{(2\pi\tau_i^2)^{(\hat{L}_i+1)/2}} F(M_i + \hat{L}_i + 1, (\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i)^T(\mathbf{z}_i - \mathbf{D}_i\mathbf{c}_i) + \tau_i^{-2}\mathbf{c}_i^T\mathbf{Q}_i\mathbf{c}_i)\ \right]. \tag{20}$$

Similarly, plugging $\hat{L}_i$ in (19) and dividing by $g_\lambda(\hat{L}_i)$ we obtain

$$p(\mathbf{z}_i|\hat{L}_i) = (2\pi)^{-M_i/2}\left[\pi_0 F(M_i, \mathbf{z}_i^T\mathbf{z}_i) + (1-\pi_0)\ V(\mathbf{z}_i, \hat{L}_i, M_i)\right]. \tag{21}$$

Finally we obtain the posterior pdf (12) of $\mathbf{c}_i$ dividing (20) by (21). The estimator $\hat{\mathbf{c}}_i$ of coefficients $\mathbf{c}_i$ is the mean of the pdf (12).

References

[1] Abramovich, F. and Angelini, C. (2006). Bayesian maximum a posteriori multiple testing procedure. *Sankhya*, **68**, 436-460.

[2] Bar–Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, vol. 20, no 16, 2493–2503.

[3] Bar–Joseph, Z. , Gerber, G., Jaakkila, T., Gifford, D., and Simon, I. (2003a). Comparing the continuous representation of time series expression profiles to identify differentially expressed genes. *Proc. Nat. Acad Sci. USA,* **100**, 10146–10151.

[4] Bar–Joseph, Z. , Gerber, G., Jaakkila, T., Gifford, D., and Simon, I. (2003b). Continuous representation of time series gene expression data. *J. Comput. Biol.,* **3-4**, 341–356.

[5] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer series in Statistics.

[6] Burnham, K.P., and Anderson, D.R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach.* Second ed. Springer-Verlag, New York.

[7] Chipman, H., George, E.I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. With discussion. In *IMS Lecture Notes Monogr. Ser. Model selection*, **38**, 65–134.

[8] Cicatiello, L., Scarfoglio, C., Altucci, L., Cancemi, M., Natoli, G., Facchiano, A., Iazzetti G., Calogero, R., Biglia, N., De Bortoli, M., Sfiligol, C., Sismondi, P., Bresciani, F. and Weisz, A. (2004). A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive trascriptome. *Journal of Molecular Endocrinology,* **32**, 719–775.

[9] Conesa, A., Nueda, M.J., Ferrer, A., and Talon, M. (2006). MaSigPro: a method to identify significantly dofferential expression profiles in time-course microarray-experiments.*Bioinformatics,* **22,** 1096–1102.

[10] Cui, X., Kerr, M.K., and Churchill, G. A. (2003). Transformation for cDNA Microarray Data. *Statistical Applications in Genetics and Molecular Biology,* **2** (1).

[11] Donoho, D.L. (1992). De-noising by soft thresholding. *IEEE transaction on Information Theory*, **41**, 613–627.

[12] de Hoon, M.J.L., Imoto, S., and Miyano, S. (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. *Bioinformatics,* **18,** 1477–1485.

[13] DiCamillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S.K., Trajanosky, Z., and Cobelli, C. (2005). A quantization method based on threshold optimization for microarray short time series. *BMC Bioinformatics,* **6**.

[14] Dudoit, S., Yang, Y.H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica,* **12**, 111–140.

[15] Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *J. Amer. Statist. Assoc.*, **96**, 1151–1160.

[16] Ernst, J., Nau, G.J., and Bar–Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, **21**, 1159–1168.

[17] Heard, N.A., Holmes, C.C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the Immune response of Anopheline Mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.*, **101**, 18–29.

[18] Johnstone, I.M., and Silverman, B.W. (2004). Finding needles hay in haystacks: Risk bounds for empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594–1649.

[19] Kerr, M.K., Martin M., and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**, 819–837.

[20] Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G.A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica,* **12**, 203–218.

[21] Ishwaran, H., and Rao, J.S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.*, **98**, 438–455.

[22] Leek, J.T., Monsen, E., Dabney, A.R., and Storey, J.D. (2006). EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, **22**, 507–508.

[23] Lonnstedt, I., and Speed, T. (2002). Replicated microarray data. *Statistica Sinica,* **12**, 31–46.

[24] McLachlan, G., Do, K.A., and Ambroise, C. (2004). *Analyzing microarray gene expression data.* Wiley series in Probability and Statistics.

[25] Opgen-Rhein, R., and Strimmer, K. (2006). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, **4**, 53–65.

[26] Park, T., Yi, S.G., Lee, S., Lee, S.Y., Yoo, D.H., Ahn, J.I., and Lee, Y.S. (2003). Statistical tests for identifying differentially expressed genes in time course microarray experiments. *Bioinformatics,* **19,** 694–703.

[27] Pawitan, Y., Krishna Murthy, K.R., Michiels, S., and Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21,** 3865–3872.

[28] Pounds, S., and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics,* **20**, 1737–1745.

[29] Smyth, G.K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Eds. Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., Springer, New York, pp. 397–420.

[30] Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., and Davis, R.W. (2005). Significance analysis of time course microarray experiments (with supplementary material). *Proc. Nat. Acad. Soc.*, **12**, 12837-12842.

[31] Storey, J.D., and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differantial gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software.* Eds. Parmigiani, G.,Garrett, E.S., Irizarry, R. A., and Zeger, S.L. Statistics for Biology and Health. Springer, pp. 272–290.

[32] Tai, Y. C., and Speed, T.P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Stat.*, bf 34, 2387–2412.

[33] Tusher, V., Tibshirani, R., and Chu, C. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Soc.*, **98**, 5116–5121.

[34] Yang, Y.H., Dudoit, S., Luu, P., Lin, M.D., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a

robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4).

[35] Wit, E., and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*, Wiley, Chichester, West Sussex, England.

[36] Wu, H., Kerr, M. K., Cui, X., and Churchill, G.A. (2003). MAANOVA: A software package for Analysis of spotted cDNA Microarray experiments. In *The Analysis of Gene Expression Data: Methods and Software.* Eds. Parmigiani, G.,Garrett, E.S., Irizarry, R. A., and Zeger, S.L. Statistics for Biology and Health. Springer, pp. 313–341.