

Bayesian methods for time course microarray analysis: from genes' detection to clustering¹

Claudia Angelini²

Istituto per le Applicazioni del Calcolo, e-mail: c.angelini@iac.cnr.it

Daniela De Canditiis

Istituto per le Applicazioni del Calcolo, e-mail: d.decanditiis@iac.cnr.it

Marianna Pensky

Department of Mathematics, University of Central Florida, USA, e-mail:
mpensky@pegasus.cc.ucf.edu

Abstract: Time-course microarray experiments are an increasingly popular approach for understanding the dynamical behavior of a wide range of biological systems. In this paper we discuss some recently developed functional Bayesian methods specifically designed for time-course microarray data. The methods allow one to identify differentially expressed genes, to rank them, to estimate their expression profiles and to cluster the genes associated with the treatment according to their behavior across time. The methods successfully deal with various technical difficulties that arise in this type of experiments such as a large number of genes, a small number of observations, non-uniform sampling intervals, missing or multiple data and temporal dependence between observations for each gene. The procedures are illustrated using both simulated and real data.

Keywords: Bayesian Functional Approaches, Clustering, Estimation, Testing

1. Introduction

Time course microarray experiments are an increasingly popular approach for understanding the dynamical behavior of a wide range of biological systems. From one side, experiments can be carried out both to identify the genes that are associated to a given condition or respond to a given treatment, to determine similar responses and behaviors, and to study from the gene network their regulation mechanisms. On the other hand, analyzing high-dimensional time course-microarray data requires very specific statistical methods which are usually not included in standard software packages, as a consequence the potential of these experiments has not yet been fully exploited.

In this paper we first discuss some functional Bayesian methods recently developed in *Angelini et al (2007)* and *Angelini et al (2009a)* for detecting differentially expressed genes in time course microarray experiments. These methods allow one to identify differentially expressed genes in both the “one-sample” and the “two-sample” experimental designs, to rank them and to estimate their expression profiles. Then, we describe the novel functional approach for clustering the genes associated to the treatment or condition of interest according to their temporal expression profile. The latter approach automatically determines the number of existing groups and the ‘optimal’ partition of the genes in

¹Supported by CNR-Bioinformatics Project and CNR-DG.RSTL.004.002

²Address of correspondence: Dr. Claudia Angelini, IAC-CNR, via Pietro Castellino 111, 80131 Napoli

these groups, moreover it can provide a bunch of ‘quasi-optimal’ solutions that one can further investigate. All the methods successfully deal with various technical difficulties that arise in microarray time-course experiments such as a large number of genes, a small number of observations, non-uniform sampling intervals, missing or multiple data and temporal dependence between observations for each gene. The procedures are illustrated using both simulated and real data, moreover they are compared with competitive recent approaches.

2. Detection of differentially expressed genes in time course microarray experiments

In a ‘one sample’ experimental design the data consists of the records, for N genes, of the differences in gene expression levels between the sample of interest and a reference (i.e., treated and control) in course of time. Each record is modeled as a noisy measurement of a function $s_i(t)$ at a time instant $t^{(j)} \in [0, T]$ which represents the differential gene expression profile: $z_i^{j,k} = s_i(t^{(j)}) + \zeta_i^{j,k}$, $i = 1, \dots, N$, $j = 1, \dots, n$, $k = 1, \dots, k_i^{(j)}$. Here the number n of time points is relatively small, with very few replications ($k_i^{(j)}$) available at each time instants, while the number N of genes is very large. $M_i = \sum_{j=1}^n k_i^{(j)}$ observations are available for each gene. The objective is to identify the genes showing different functional expression between treated and control, and then to evaluate the effect of the treatment. For each gene i , we expand its expression profile $s_i(t)$ into a series over some standard orthonormal basis on $[0, T]$ with coefficients $c_i^{(l)}$, $l = 0, \dots, L_i$:

$$s_i(t) = \sum_{l=0}^{L_i} c_i^{(l)} \phi_l(t). \quad (1)$$

Following *Angelini et al. (2007)*, genes are treated as conditionally independent and their expressions are modeled as $\mathbf{z}_i = \mathbf{D}_i \mathbf{c}_i + \boldsymbol{\zeta}_i$. Here, $\mathbf{D}_i$ is the 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 $k_i^{(j)}$ times; $\mathbf{z}_i = (z_i^{1,1} \dots z_i^{1,k_i^{(1)}} \dots z_i^{n,1} \dots z_i^{n,k_i^{(n)}})^T$, $\mathbf{c}_i = (c_i^{(0)} \dots c_i^{(L_i)})^T$ and $\boldsymbol{\zeta}_i = (\zeta_i^{1,1} \dots \zeta_i^{1,k_i^{(1)}} \dots \zeta_i^{n,1} \dots \zeta_i^{n,k_i^{(n)}})^T$ are, respectively, the column vectors of all measurements for gene i , the coefficients of $s_i(t)$ in the chosen basis and the random errors. The following hierarchical model is imposed on the data:

$$\mathbf{z}_i \mid L_i, \mathbf{c}_i, \sigma^2 \sim \mathcal{N}(\mathbf{D}_i \mathbf{c}_i, \sigma^2 \mathbf{I}_{M_i}) \quad L_i \sim \text{Truncated Poisson}(\lambda, L_{\max}) \quad (2)$$

$$\mathbf{c}_i \mid L_i, \sigma^2 \sim \pi_0 \delta(0, \dots, 0) + (1 - \pi_0) \mathcal{N}(0, \sigma^2 \tau_i^2 \mathbf{Q}_i^{-1}) \quad (3)$$

All parameters in the model are treated either as random variables or as nuisance parameters, recovered from the data. Noise variance σ^2 is assumed to be random, $\sigma^2 \sim \rho(\sigma^2)$, and the following priors allow to account for possibly non-Gaussian errors:

Model 1: $\rho(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$, the point mass at σ_0^2 .

Model 2: $\rho(\sigma^2) = IG(\gamma, b)$, the Inverse Gamma distribution.

Model 3: $\rho(\sigma^2) = c_\mu \sigma^{M_i-1} e^{-\sigma^2 \mu/2}$.

The automatic detection of differentially expressed genes is carried out on the basis of Bayes Factors (BF), which are used for taking into account multiplicity of errors. Subsequently, the curves $s_i(t)$ are estimated by the posterior means. The algorithm is self-contained and the hyperparameters are estimated from the data. Explicit formulae and other details can be found in *Angelini et al. (2007)*; the procedure is implemented in *Angelini et al (2008)*.

The above approach has been extended to the case of the two-sample experimental design in *Angelini et al (2009a)*. Although mathematically the two-sample set-up appears homogeneous, in reality it is not. In fact, it may involve comparison between the cells under different biological conditions (e.g., for the same species residing in different parts of the habitat) or estimating an effect of a treatment (e.g., effect of estrogen treatment on a breast cell). Hence, we will distinguish between “interchangeable” and “non-interchangeable” models. For a gene i in the sample $\aleph$, we assume that evolution in time of its expression level is governed by a function $s_{\aleph i}(t)$ and each of the measurements $z_{\aleph i}^{j,k}$ involves some measurement error. The quantity of interest is the difference between expression levels $s_i(t) = s_{1i}(t) - s_{2i}(t)$.

The natural extension of (1) and (2-3) is: $\mathbf{z}_{\aleph i} \mid L_i, \mathbf{c}_{\aleph i}, \sigma^2 \sim \mathcal{N}(\mathbf{D}_{\aleph i} \mathbf{c}_{\aleph i}, \sigma^2 \mathbf{I}_{M_{\aleph i}})$ where the vectors of coefficients $\mathbf{c}_{\aleph i}$ are modeled differently in the interchangeable and non-interchangeable cases.

Interchangeable set-up Assume that a-priori vectors of the coefficients $\mathbf{c}_{\aleph i}$, $\aleph = 1, 2$ are either equal to each other or are independent and have identical distributions:

$$\mathbf{c}_{1i}, \mathbf{c}_{2i} \mid L_i, \sigma^2 \sim \pi_0 \mathcal{N}(\mathbf{c}_i \mid \mathbf{0}, \sigma^2 \tau_i^2 \mathbf{Q}_i^{-1}) \delta(\mathbf{c}_{1i} = \mathbf{c}_{2i}) + (1 - \pi_0) \prod_{\aleph=1}^2 \mathcal{N}(\mathbf{c}_{\aleph i} \mid \mathbf{0}, \sigma^2 \lambda_i^2 \mathbf{Q}_i^{-1}).$$

Non-interchangeable set-up Assume that the expression level in sample 2 is the sum of the expression level in sample 1 and the effect of the treatment $\mathbf{d}_i$, i.e. $\mathbf{c}_{2i} = \mathbf{c}_{1i} + \mathbf{d}_i$: $\mathbf{c}_{1i} \mid L_i, \sigma^2 \sim \mathcal{N}(\mathbf{c}_{1i} \mid \mathbf{0}, \sigma^2 \tau_i^2 \mathbf{Q}_i^{-1})$ and $\mathbf{d}_i \mid L_i, \sigma^2 \sim \pi_0 \delta(\mathbf{d}_i = \mathbf{0}) + (1 - \pi_0) \mathcal{N}(\mathbf{d}_i \mid \mathbf{0}, \sigma^2 \lambda_i^2 \mathbf{Q}_i^{-1})$.

In both set ups, parameter σ^2 is treated as a random variable with the same three choices for $\rho(\sigma^2)$. The inference is carried out similarly to the “one sample case”. For brevity, we do not report the formulae, see *Angelini et al. (2009a)* for details. Note that the model allows that the two samples can be observed on different grids $t_{\aleph}^{(j)}$ taken in the same interval $[0, T]$.

A great advantage of both Bayesian models described above is that all evaluations are carried out in analytic form, with very efficient computations.

3. Clustering time course profiles

Once the genes associated to the condition and treatment of interest have been selected, in order to investigate gene-regulation’s mechanisms, it is useful to perform a clustering according to their expression profile across time. Starting from (1) we have recently proposed a novel functional clustering procedure that is detailed described in *Angelini et al (2009b)*. We assume that the coefficients $\mathbf{c}_i$ follows the two-level hierarchical model:

Level 1: For each gene i , given the degree of the polynomial L_i , the vector of coefficients $\mathbf{c}_i$ and τ_i^2 we assume $\mathbf{z}_i \mid L_i, \mathbf{c}_i, \tau_i^2 \sim \mathcal{N}(\mathbf{D}_i \mathbf{c}_i, \sigma^2 \tau_i^2 \mathbf{I}_{M_i})$.

Level 2: Given the degree of the polynomial L_i , the mean $\boldsymbol{\mu}_i$ and the precision τ_i^2 we assume $\mathbf{c}_i \mid L_i, \boldsymbol{\mu}_i, \tau_i^2 \sim \mathcal{N}(\boldsymbol{\mu}_i, \tau_i^2 \mathbf{I}_{L_i})$.

In the above notation, τ_i^2 represents the cluster's variability, while σ^2 is a parameter that scales the variances of instrumental errors. Integrating out vectors of parameters $\mathbf{c}_i$ yields the following marginal distributions of the data vectors

$$\mathbf{z}_i \mid L_i, \boldsymbol{\mu}_i, \tau_i^2 \sim \mathcal{N}(\mathbf{D}_i \boldsymbol{\mu}_i, \tau_i^2 \mathbf{A}_i). \quad (4)$$

where $\mathbf{A}_i = \mathbf{D}_i \mathbf{D}_i^t + \sigma^2 \mathbf{I}_{M_i}$ is a known gene-specific design-related matrix.

Assuming that two genes, i and j , belong to the same cluster if and only if the corresponding vectors of coefficients $\mathbf{c}_i$ and $\mathbf{c}_j$ are of the same dimension ($L + 1$) and are sampled from the same $(L + 1)$ -variate normal distribution with mean $\boldsymbol{\mu}$ and scale parameter τ^2 , we can characterize each cluster by the triplet of parameters $(L, \boldsymbol{\mu}, \tau^2)$. Then, introducing a latent vector γ of length N (number of profile to be clustered) such that $\gamma_i = \gamma_j$ if and only if genes i and j belong to the same cluster. Then, expression (4) can be rewritten as $\mathbf{z}_i \mid \gamma_i, L_{\gamma_i}, \boldsymbol{\mu}_{\gamma_i}, \tau_{\gamma_i}^2 \sim \mathcal{N}(\mathbf{D}_i \boldsymbol{\mu}_{\gamma_i}, \tau_{\gamma_i}^2 \mathbf{A}_i)$.

A Dirichlet process prior is elicited on the distribution of the triplet $(L, \boldsymbol{\mu}, \tau^2)$ of parameters that define the cluster: $(L, \boldsymbol{\mu}, \tau^2) \sim DP(\alpha, G_0)$ where α is the scale parameter and G_0 the base distribution. Note that $DP(\alpha, G_0)$ is a discrete process and it has the great advantage to allow any number of clusters to be present in the dataset. We choose the following conjugate base distribution $G_0(L, \boldsymbol{\mu}, \tau^2) = g_\lambda(L) NIG(\boldsymbol{\mu}, \tau^2 \mid a, b, 0, \rho^2 \mathbf{Q})$, $NIG(\boldsymbol{\mu}, \tau^2 \mid a, b, 0, \rho^2 \mathbf{Q}) = N(\boldsymbol{\mu} \mid 0, \tau^2 \rho^2 \mathbf{Q}) IG(\tau^2 \mid a, b)$, and $g_\lambda(L)$ as in (2).

Finally, we make inference on the posterior $p(\gamma \mid \mathbf{z})$ using the improved MCMC Split-Merge procedure proposed by *Dahl (2003)*. After burn-in, we chose the MAP criterion for selecting the ‘optimal’ allocation. From one side, the great advantage of this Bayesian model is that, at the price of higher computational cost, it automatically identifies the number of clusters hidden in the data-sets and the ‘optimal’ partition. On the other hand, pooling together results of different chains or looking at the marginal posterior probability of each visited model, it can also provide a limited number of ‘quasi optimal’ solutions that the experimenter can easily investigate.

References

- Angelini C., Cutillo L., De Canditiis D., Mutarelli M. and Pensky M. (2008) Bats: A bayesian user friendly software for analyzing time series microarray experiments, *BMC Bioinformatics*, 9.
- Angelini C., De Canditiis D., Mutarelli M. and Pensky M. (2007) A bayesian approach to estimation and testing in time-course microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, 6.
- Angelini C., De Canditiis D. and Pensky M. (2009a) Bayesian models for the two-sample time-course microarray experiments, *Computational Statistics & Data Analysis*, 53, 1547–1565.
- Angelini C., De Canditiis D. and Pensky M. (2009b) Clustering time-course microarray data using functional bayesian infinite mixture model, *Technical report IAC-CNR*.
- Dahl D. (2003) An improved merge-split sampler for conjugate dirichlet process mixture models, *Technical Report 1086, Department of Statistics, University of Wisconsin*.