

Detecting differentially expressed genes in longitudinal microarray experiments

Maurice Berk

Imperial College London, e-mail: maurice.berk@imperial.ac.uk

Giovanni Montana¹

Imperial College London, e-mail: giovanni.montana@imperial.ac.uk

Abstract: In this work we propose a statistical approach for the detection of differentially expressed genes. Specifically, we envisage a longitudinal study design where multiple biological replicates are available and each individual is observed at more than one time point. We present a semi-parametric random effects model based on smoothing splines and introduce a visual-distance metric for detecting time course profiles that differ across experimental conditions. Parameter estimation is carried out through the EM algorithm whereas model search is performed using a simplex optimization algorithm. We suggest the use of the corrected AIC for model selection and demonstrate that this criterion performs particularly well for small sample sizes. The performance of our approach is assessed through extensive Monte Carlo simulations and is compared to alternative statistical tests. Finally, we present a power study and illustrate the methods on real data sets.

Keywords: Differential analysis, longitudinal data, mixed-effects models, functional data analysis, smoothing splines, corrected AIC, simplex optimization

1. Introduction

Replication in microarray experiments is essential for making statistical inference and comes in two forms: technical and biological. Technical replicates arise when the same source of biological material is used in multiple hybridizations in order to account for experimental variation; for example, the same probe is spotted multiple times on a microarray or a dye-swap is performed. In this work we are mostly interested in situations where different biological units are sampled so that genetic variation across the population may be modeled. When different biological units are each repeatedly sampled over time, this gives rise to a longitudinal study design. To date, only a handful of methodologies have been proposed that are appropriate for analyzing data produced in longitudinal experiments. Two such examples are the multivariate empirical Bayes approach of Tai and Speed (2006), implemented in the *timecourse* package for BioConductor, and the semi-parametric mixed effect model of Storey *et al.* (2005) implemented in the *EDGE* software, also available as an R package. The latter approach assumes a model of form

$$y(t_{ij}) = \mu(t_{ij}) + v_i(t_{ij}) + \epsilon_{ij}$$

⁰The present paper is financially supported by Wellcome Trust

¹Address of correspondence: Imperial College London, Department of Mathematics, 180 Queen's Gate, London SW7 2AZ

where $y(t_{ij})$ is the j th gene expression measurement taken on individual i at time t_{ij} . The model postulates that the additive effect of two components explains the observed variation in the data: a population mean response $\mu(\cdot)$ and an individual-specific deviation from that response, $v_i(\cdot)$. As usual, ϵ_{ij} accounts for measurement error. Similar semi-parametric models for microarray time course data have been previously proposed in settings where no biological replicates are available, for instance by Bar-Joseph *et al.* (2003) and Luan and Li (2003). In those cases the focus was on clustering, and the unit was a gene, rather than an individual. Specific choices have to be made regarding the parameterization of mean response and individual effects. For instance, a mean response can be flexibly modeled using basis functions such as B-splines or cubic splines. In their proposal, Storey *et al.* (2005) used natural cubic splines for representing the mean curve and modelled the individual deviations as simple scalar shifts. Two main shortcomings of their approach are in order. Firstly, natural cubic splines require the selection of the number and location of knots. Although several standard solutions exist for dealing with these choices, including data-driven procedures, these issues are yet to be fully resolved. Secondly, although the choice of modeling individual deviations as scalar shifts is a parsimonious one and may be sufficient for modeling data sets with little biological variation, in our experience this approach may lead to problematic model fits in more noisy data. Motivated by these arguments, we set out to extend the model of Storey *et al.* (2005) in a number of directions. We propose a data-driven approach where we can trade off between model-fit and degrees of freedom, accounting for the small sample size, in order to use more flexible models where the individual deviations are also modeled as smooth functions of time, allowing for complex departures from the mean response. We do this by modeling both the mean and individual curves using smoothing splines. Moreover, we propose a test statistic that encapsulates the distance between the observed mean curves in a more natural or *visual* fashion.

2. A semi-parametric mixed effect model

We assume that the experiments have been carried out under K experimental conditions and the objective of the analysis is to detect gene expression profiles that differ across conditions. We treat the modeling of each gene independently as this greatly simplifies the procedure and is typical of other approaches in the area. Furthermore data in each group are fitted separately. If we define an incidence matrix $\mathbf{X}_i$ of dimension $n_i \times M$, where n_i is the number of observations on individual i and M is the number of design timepoints, such that in the j th row there is a single 1 in the column that corresponding to t_{ij} and 0s elsewhere, then we can express our model as

$$y_i(\mathbf{t}_i) = \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$$

This is a linear mixed-effects model (see, for instance, Harville (1977); Laird and Ware (1982)). The vector $\boldsymbol{\beta}$ includes values for the mean curve at the design timepoints. These are population parameters common to all individuals which are assumed to be fixed, yet unknown. In contrast, each $\mathbf{b}_i$ is a vector of shifts from that mean at the design timepoints for individual i . These shifts are assumed to be randomly sampled from the population. Specifically we assume that for all individuals the random effects and the errors $\boldsymbol{\epsilon}_i$ are independent multivariate normal variates with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}$ and $\mathbf{R}_i$, respectively. For simplicity we assume uncorrelated errors, hence $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$.

These distributional assumptions mean that $\mathbf{y}_i$ is itself marginally multivariate normally distributed with mean vector $\mathbf{X}_i\boldsymbol{\beta}$ and covariance $\mathbf{V}_i = \mathbf{X}_i\mathbf{D}\mathbf{X}_i^T + \sigma^2\mathbf{I}_{n_i}$. The standard approach to estimating the parameters $\boldsymbol{\beta}$ and $\mathbf{b}_i$, and the variance components $\mathbf{D}$ and σ^2 in the linear mixed-effects model is maximum likelihood estimation where the unobserved random-effects are treated as missing data in the EM algorithm. The log-likelihood $\log(\ell(\boldsymbol{\beta}, \mathbf{D}, \sigma^2|\mathbf{y}_i))$ is given by

$$-\frac{1}{2}n_i \log 2\pi - \frac{1}{2} \log |\mathbf{V}_i| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

As we model the fixed- and random-effects using smoothing splines, we penalize this likelihood with penalty terms for the roughness of these two curves, that is by adding terms $\lambda\boldsymbol{\beta}^T\mathbf{G}\boldsymbol{\beta}$ and $\lambda_v\mathbf{b}_i^T\mathbf{G}\mathbf{b}_i$, where λ and λ_v are smoothing parameters which control the size of the penalty and $\mathbf{G}$ is termed the roughness matrix. Note that we assume the same smoothing parameter λ_v for all individuals. In principle we can allow it to vary but this incurs a great computational cost. We incorporate the random-effects penalty by redefining the covariance of $\mathbf{y}_i$ as $\tilde{\mathbf{V}}_i = \mathbf{X}_i\mathbf{D}_v\mathbf{X}_i^T + \sigma^2\mathbf{I}_{n_i}$ where $\tilde{\mathbf{V}}_i$ and $\mathbf{D}_v = (\mathbf{D}^{-1} + \lambda_v\mathbf{G})^{-1}$ are the regularized covariance matrices of $\mathbf{y}_i$ and the random-effects, respectively. We then derive the ML estimators of $\boldsymbol{\beta}$, $\mathbf{b}_i$, $\mathbf{D}$ and σ^2 to be used within each EM iteration.

3. Model search and selection

Model selection involves the choice of the smoothing parameters λ and λ_v that optimize a given criterion such as the Akaike Information Criteria (AIC), the Bayesian Information Criteria (BIC) or generalized cross-validation. Exhaustive grid-based searches in the parameter space are generally adopted in practical settings, as in Wu and Zhang (2006). Alternatively, one can incorporate model selection in the estimation procedure by treating the smoothing parameters as variance components, as in Guo (2002). We prefer to disassociate the model selection from the estimation as this is a much more flexible approach and choose to optimize the *corrected* AIC or AICc which includes a correction for small sample sizes. Based on Monte Carlo simulations, we have found the AICc performs more favorably in the microarray context, as we would expect given the paucity of data. The search in the two-dimensional space is performed using the simplex method of Nelder and Mead (1965).

4. Detection of differential time series

In order to detect differences in the temporal profile of a gene between any two experimental conditions, we propose to use a distance measure, which we call *visual distance*, between the two fitted mean curves. Traditional distance measures are not appropriate to capture differences between temporal expression profiles as they would be normally detected by the biologist's eye. For instance, the L_2 distance typically used to measure distances between vectors is not applicable because it only measures the vertical distance between the two curves whereas a visual interpretation allows for both vertical and horizontal displacement. The suggested visual distance is calculated by first discretizing each curve into a large number of points. For each point (x, y) in one curve, we find

the closest point (x', y') in the second, calculating the euclidean distance in the standard way. We then sum the distances of each individual point to give an overall measure. Statistical inference is performed nonparametrically whereby a *non-parametric bootstrap* approach is proposed to estimate the null distribution of the test statistic. We resample both the within- and between-individual residuals in order to construct null samples. For a given gene we fit a null model that ignores the biological groupings and an alternative model which takes it into account. The within-individual residuals from the alternative model are given as $\hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij}^1$, where $\hat{y}_{ij}^1$ are the fitted values calculated under the alternative model. The between-individual residuals from the alternative model fit, b_i^1 , are those random-effects estimated using the EM algorithm. With these two sets of residuals, randomly sampled data can be generated as $y_{ij}^* = x_{ij}\hat{\beta}^0 + x_{ij}b_i^* + \epsilon_{ij}^*$. In the first term, $\hat{\beta}^0$, is the gene's mean curve fitted values from the null model fit. In the second term, b_i^* , is a randomly sampled individual deviation from the fitted between-individual residuals. Similarly, the third term, ϵ_{ij}^* , is a randomly sampled error from the within-individual residuals. B bootstrap data sets are then generated and used to estimate the null distribution. Finally, in order to evaluate the statistical power of our approach, we have carried out a large number of Monte Carlo simulations, varying a wide range of parameters including variance components, sampling rates and expression patterns, in addition to sample sizes. Our method for simulating differentially expressed genes uses interpolating splines in order to introduce features to steady-state curves, for instance spikes or jumps in expression levels. We also report on the analysis of two real longitudinal studies.

References

- Bar-Joseph Z., Gerber G.K., Gifford D.K., Jaakkola T.S. and Simon I. (2003) Continuous representations of time-series gene expression data, *Journal of Computational Biology*, 10, 3-4, 341–356.
- Guo W. (2002) Functional mixed effects models, *Biometrics*, 58, 121–128.
- Harville D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, 72, 358, 320–388.
- Laird N.M. and Ware J.H. (1982) Random-effects models for longitudinal data, *Biometrics*, 38, 963–974.
- Luan Y. and Li H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines, *Bioinformatics*, 19, 4, 474–482.
- Nelder J.A. and Mead R. (1965) A simplex method for function minimization, *Computer Journal*, 7, 308–313.
- Storey J.D., Xiao W., Leek J.T., Tompkins R.G. and Davis R.W. (2005) Significance analysis of time course microarray experiments., *Proc Natl Acad Sci U S A*, 102, 36, 12837–12842.
- Tai Y.C. and Speed T.P. (2006) A multivariate empirical bayes statistic for replicated microarray time course data, *Annals of Statistics*, 34, 5, 2387–2412.
- Wu H. and Zhang J.T. (2006) *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley.