Joint segmentation of multivariate Gaussian Processes using mixed linear models

by

Franck Picard, Emilie Lebarbier, Eva Budinská and Stéphane Robin

Research Report No. 5
March 2007

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

Joint segmentation of multivariate Gaussian Processes using mixed linear models

Franck Picard * Émilie Lebarbier † Eva Budinská ‡
Stéphane Robin §

Abstract

We consider the joint segmentation of multiple series. We use a mixed linear model to account for both covariates and correlations between signals. We propose an estimation algorithm based on EM which involves dynamic programming for the segmentation step. We show the computational efficiency of this procedure. An application to microarray CGH profiles from multiple patients is presented.

Keyword : Dynamic programming; EM algorithm; Mixed linear model; Segmentation.

1 Introduction

Many application fields in statistics provide signals which are in the form of non-stationary time series. To simplify the interpretation of such signals, segmentation models are often used to identify intervals in which the signal is homogeneous. To do this the data are modelled by a random process whose parameters are subject to abrupt changes at unknown coordinates. This is the so-called *off-line* multiple changed-point problem. When considering

*UMR CNRS 8071 - INRA 1152 - Univ. d'Evry, 523 place des Terrasses, 91000 Evry, France (e-mail: picard@genopole.cnrs.fr)

†UMR INA P-G/ENGREF/INRA-MIA 518, 16 rue Claude Bernard, 75231 Paris cedex 05, France (e-mail: lebarbie@inapg.fr)

‡Center of Biostatistics and Analysis, Faculty of Science and Faculty of Medicine, Masaryk University, Kamenice 126/3, 625 00, Brno, Czech Republic (e-mail: budinska@cba.muni.cz)

§UMR INA P-G/ENGREF/INRA-MIA 518, 16 rue Claude Bernard, 75231 Paris cedex 05, France (e-mail: robin@inapg.fr)

univariate processes, the objective is to identify the number and the position of the change-points, as well as the value of the parameter between two changes. Many strategies exist in this framework, and intensive research has been conducted to develop efficient segmentation algorithms.

Despite a wide range of applications, little has been done in the case of multivariate processes, when the purpose is to detect and characterize structure in two or more related series [4]. In the following we focus on the *joint* segmentation problem, for which each series is segmented jointly with other series, compared with the *simultaneous* segmentation problem for which changes are common among series. We use the linear model approach to model the change-points, as already used by [1, 2], and we introduce random effects to structure the covariance matrix of the process. This model allows us to introduce a correlation structure among series at every instant, and to incorporate additional covariates in the model which are not subjet to changes. Consequently we consider segmentation models with partial structural changes.

One main issue when using segmentation methods is to define an efficient estimation procedure for change-point positioning. This optimization problem can be viewed as a partitioning problem whose purpose is to segment N data points into K segments, K being fixed. When using the maximum likelihood estimation method, dynamic programming (DP) has shown excellent performance [9], as it reduces the algorithmic complexity from $\mathcal{O}(N^K)$ to $\mathcal{O}(KN^2)$. However the use of the traditional DP is not possible when considering partial structural changes [1], and the problem becomes even more intricate with the introduction of random effects.

In this article, we propose to solve this issue using the ECM (Expectation/Conditional Maximization) algorithm to maximize the likelihood [7]. ECM is an instance of the traditional EM algorithm [3], which replaces a complicated M-step of EM with several computationally simpler CM-steps. This algorithm can be used in this context as linear mixed models can be put in the more general framework of models with incomplete data. Among the CM-steps, one is dedicated to the estimation of the breakpoints, and we show that DP can be used at this step. However, despite a drastic decrease in the complexity, DP may not be sufficient to segment multivariate processes whose data points may reach the hundreds of thousands points, as mentioned by [2]. We develop a two-stage dynamic programming procedure to solve this problem. Our method is applied to the detection of recurrent changes among the genomes of patients with colorectal cancer.

2 Model and notations

We consider M time series with n_m observations each, and we note $N = \sum_m n_m$ the total number of observations. We denote by t the position of the signal and by $n_{\max}$ the maximum number of points in a series, $t \in \{1, \ldots, n_{\max}\}$. We observe Y_{mt} the signal of series m at position t. We suppose that part of the mean of the process $\{Y_{mt}\}_t$ is subject to $K_m - 1$ abrupt changes at breakpoints $\{t_k^m\}$ for series m, (with convention $t_0^m = 0$ and $t_{K_m}^m = n_{\max}$) and is constant between two breakpoints within the interval $I_k^m =]t_{k-1}^m, t_k^m]$. In the following we denote by $K = \sum_m^M K_m$ the total number of segments accross series which is fixed in this section.

We consider the following linear mixed effect model

$$\forall t \in I_k^m, \;\; Y_{mt} = \mu_{mk} + \mathbf{x}_{mt}\boldsymbol{\theta} + U_t + E_{mt},$$

with $\mathbf{x}_{mt}$ a $[1 \times p]$ vector of covariates, $\boldsymbol{\theta}$ the corresponding parameter which is not subject to changes. When $p = 0$ we obtain a pure structural change model. U_t is the random effect at position t, which models the correlations among series. E_{mt} stands for the noise.

To use the matricial formulation of linear models, we introduce the $[N \times K]$-incidence matrix of breakpoints denoted by $\mathbf{T} = \text{Bloc}\,[\mathbf{T}_m]$ with $\mathbf{T}_m = \text{Bloc}\left[\mathbb{1}_{n_{K_m}^m}\right]$ of size $[n_m \times K_m]$, and with $n_k^m = t_k^m - t_{k-1}^m$ being the length of segment k for series m. We also introduce notation $\boldsymbol{\mu} = [\mu_{mk}]$ which corresponds to the fixed effects subject to changes (of size $[K \times 1]$). Using the matricial formulation of linear models, we have

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{U} + \mathbf{E},$$

where $\mathbf{Y}$ ($[N \times 1]$) stands for the observed data, and where $\mathbf{T}, \mathbf{X}, \mathbf{Z}$ are incidence matrices of breakpoints, constant parameter and random effects with respective size $[N \times K]$, $[N \times p]$ and $[N \times n_{\max}]$. Note that compared with classical linear models, incidence matrix $\mathbf{T}$ is unknown and should be estimated. As for the random part of the model, $\mathbf{U}$ ($[n_{\max} \times 1]$) stands for the random effects and $\mathbf{E}$ ($[N \times 1]$) for the noise. $\mathbf{U}$ is centered Gaussian with covariance matrix $\mathbf{G}$; $\mathbf{E}$ is centered Gaussian with diagonal covariance matrix $\mathbf{R}$; $\mathbf{U}$ and $\mathbf{E}$ are independent. Consequently, the covariance matrix of $\mathbf{Y}$ is $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

3 Parameter estimation using the ECM algorithm

We propose to estimate the parameters of the model by maximum likelihood. In the following, we denote by $\phi = (\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{G}, \mathbf{R}, \mathbf{T})$ the set of parameters to be estimated. The use of the EM algorithm [3] is now well established in the context of parameter estimation for mixed linear models [11], as these models can be put in the general framework of models with incomplete data. In this case, random effects $\mathbf{U}$ constitute the unobserved data, and the use of EM lies in the decomposition of the complete-data log-likelihood such that: $\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi) = \log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\mu}, \mathbf{R}) + \log \mathcal{L}(\mathbf{U}; \mathbf{G})$. We denote by $\mathbb{E}_\phi\{\cdot\}$ the expectation operator using ϕ as the parameter value and $\mathbb{V}_\phi\{\cdot\}$ the corresponding variance. The conditional expectation $Q(\phi; \phi^{(h)})$ of $\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi)$ given $\mathbf{Y}$ is also a sum of two terms $Q_0(\phi; \phi^{(h)})$ and $Q_1(\phi; \phi^{(h)})$:

$$\begin{aligned}
-2Q_0(\phi; \phi^{(h)}) &= -2\mathbb{E}_{\phi^{(h)}}\left\{\log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\mu}, \mathbf{R})|\mathbf{Y}\right\} \\
&= N\log(2\pi) + \log|\mathbf{R}|^{-1} + \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}\|^2_{\mathbf{R}^{-1}} \\
&+ \operatorname{Tr}\left(\mathbf{R}^{-1}\mathbf{Z}\mathbf{W}\mathbf{Z}'\right), \\
-2Q_1(\phi; \phi^{(h)}) &= -2\mathbb{E}_{\phi^{(h)}}\left\{\log \mathcal{L}(\mathbf{U}; \mathbf{G})|\mathbf{Y}\right\} \\
&= M\log(2\pi) + \log|\mathbf{G}|^{-1} + \widehat{\mathbf{U}}'\mathbf{G}^{-1}\widehat{\mathbf{U}} + \operatorname{Tr}\left(\mathbf{G}^{-1}\mathbf{W}\right),
\end{aligned}$$

where $\widehat{\mathbf{U}} = \mathbb{E}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$ stands for the best linear unbiased predictor (BLUP) of the random effects $\mathbf{U}$, $Tr(A)$ for the trace of matrix A, $|A|$ for its determinant and where $\mathbf{W} = \mathbb{V}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$.

3.1 E-step

This step consists in the calculation of $Q(\phi; \phi^{(h)})$ which only requires the calculation of $\widehat{\mathbf{U}}$ and $\mathbf{W}$. The BLUP is such that $\widehat{\mathbf{U}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{T}\boldsymbol{\mu})$, and we use Henderson's trick which avoids the inversion of $\mathbf{V}$. So we get at iteration $(h+1)$

$$\begin{aligned}
\widehat{\mathbf{U}}^{(h+1)} &= \mathbf{W}^{(h)}\mathbf{Z}'\mathbf{R}^{(h)-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h)} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)}\right), \\
\mathbf{W}^{(h+1)} &= \left(\mathbf{Z}'\mathbf{R}^{(h)-1}\mathbf{Z} + \mathbf{G}^{(h)-1}\right)^{-1}.
\end{aligned}$$

3.2 CM-steps

The principle of the ECM algorithm is to breakdown the maximization of $Q(\phi; \phi^{(h)})$ with respect to ϕ (global M-step) into simpler CM-steps which focus on one parameter, the others being fixed. The convergence properties of ECM are provided in [7].

Estimation of $\boldsymbol{\theta}$. The update of $\boldsymbol{\theta}$ is done with the classical least-squares estimator

$$\mathbf{X}'\mathbf{R}^{(h)-1}\mathbf{X}\boldsymbol{\theta}^{(h+1)} = \mathbf{X}'\mathbf{R}^{(h)-1}(\mathbf{Y} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}).$$

Estimation of Variance components. We get the estimates $\mathbf{G}^{(h+1)}$ and $\mathbf{R}^{(h+1)}$ as $\arg\max_{\mathbf{G}} Q_1(\phi; \boldsymbol{\theta}^{(h+1)}, \mathbf{T}^{(h)}, \boldsymbol{\mu}^{(h)}, \mathbf{G}^{(h)}, \mathbf{R}^{(h)})$ and $\arg\max_{\mathbf{R}} Q_0(\phi; \boldsymbol{\theta}^{(h+1)}, \mathbf{T}^{(h)}, \boldsymbol{\mu}^{(h)}, \mathbf{G}^{(h+1)}, \mathbf{R}^{(h)})$, respectively. Note that when $\mathbf{G}$ is diagonal, analytic formulas can be derived for the estimates.

Estimation of segmentation parameters. This step is done such that

$$\left\{\mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)}\right\} = \arg\max_{\mathbf{T},\boldsymbol{\mu}} Q_0\left(\phi; (\boldsymbol{\theta}^{(h+1)}, \mathbf{G}^{(h+1)}, \mathbf{R}^{(h+1)})\right), \quad (1)$$

and the computation of this particular CM-step is done in the next Section.

3.3 Estimating breakpoints

The optimization problem (1) is equivalent to the minimization of the residual sum of squares:

$$\begin{aligned} SSR_K(\boldsymbol{\mu}, \mathbf{T}) &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h+1)} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}\|^2_{\mathbf{R}^{(h+1)-1}} \\ &= \sum_{m=1}^{M}\sum_{k=1}^{K_m} SSR_k^m(\boldsymbol{\mu}_m, \mathbf{T}_m) \end{aligned}$$

under the constraint $\sum_m K_m = K$. This sum is additive according to the number of segments, which allows us to use the dynamic programming algorithm in this step. The purpose is to partition the interval $[1, N]$ into K intervals structured according to series such that $[1, N] = \bigcup_{m=1}^{M} J^m = \bigcup_{m=1}^{M}\bigcup_{k=1}^{K_m} I_k^m$, with $J^m =]t_1^m, t_{K_m}^m]$. To account for this structure, and to get an efficient algorithm when N is large, we propose a two-stage dynamic programming.

Stage-1. We denote by $SSR_k^m(J^m)$ the residual sum of squares when partitionning interval J^m of series m into k segments. This segmentation step is based on the calculus of $SSR_1^m(]i, j])$ and on the recursive minimization

$$\begin{aligned} &\forall k \in [1 : K_m], \\ &SSR_k^m(]t_1^m, j]) = \min_h \left\{SSR_{k-1}^m(]t_1^m, h]) + SSR_1^m(]h, j])\right\}. \end{aligned}$$

Stage-2. We denote by $SSR_K(J^1, \ldots, J^m)$ the total sum of square for a model with K segments spread over m series. The second step consists in the repartition of segments among time series. This step is based on the calculus of $SSR_k^m(J^m)$ which has been done in Step-1, and on the recursive minimization:

$$\begin{aligned} &\forall m \in [1:M], \\ SSR_K(J^1, \ldots, J^m) &= \min_{k'+k''=K} \left\{ SSR_{k'}(J^1, \ldots, J^{m-1}) + SSR_{k''}^m(J^m) \right\}. \end{aligned}$$

Complexity. The first stage corresponds to the segmentation of individual series into a given number of segments whose complexity is $\mathcal{O}(\sum_m n_m^2 K_m)$. The complexity of the second stage is $\mathcal{O}(K^2 \times M)$ which makes the overall complexity of order $\mathcal{O}(\sum_m n_m^2 K_m + K^2 M)$. Using dynamic programming on the whole dataset would result in a complexity of $\mathcal{O}(N^2 K)$.

If all series had the same length $n_m = n$ (so $N = Mn$) and were segmented into $K_m = k$ segments each (so $K = Mk$) and assuming that $k = \lambda n$ (with $\lambda \ll 1$), the two-stage dynamic programming algorithm has a complexity of $\mathcal{O}(\lambda M n^2 [n + \lambda M^2])$ whereas the overall one has complexity $\mathcal{O}(\lambda M^3 n^3)$.

3.4 Strategy for the complete estimation algorithm

The regular ECM algorithm described above consists in the calculation of $Q(\phi; \phi^{(h)})$ (E-step) and in the maximization of this quantity (CM-steps). This last step could be achieved via the circular estimation of all the elements of ϕ until convergence. However, this should require numerous dynamic programming steps, which are the most computationally demanding. To reduce the numbers of segmentation steps, we iterate the E-step and the CM-steps for every element of ϕ except $\mathbf{T}$ and $\boldsymbol{\mu}$ until convergence, then we update $\mathbf{T}$ and $\boldsymbol{\mu}$. We applied these two algorithms on the same data to verify that they provide the same estimates in a reduced computational time.

4 Application to the analysis of multiple CGH profiles

CGH data. In this section we present an application of this method to the analysis of CGH (Comparative Genomic Hybridization) microarray data. This technology aims at detecting and mapping chromosomal aberrations along the genome. A continuous fluorescence signal is obtained by hybridizing the genomic DNA of a patient (target DNA) on a glass slides where

mapped DNA fragments (probes) are spotted (see [10] for more details). Once measured this signal is ordered according to the position of the probes, and shows some discontinuities when the number of target DNAs is different from a reference number. Segmentation methods are currently used to analyse these data but they treat each CGH profile separately (see [5] for a review on these methods). As this technology becomes a comprehensive screening tool, its used has recently been generalized to the study of shared chromosomal aberrations among patients with homogeneous clinical diagnosis.

We consider CGH profiles of chromosome 20 from a cohort of 121 colorectal cancer patients described in [8]. The cohort is divided into 4 clinical groups with respective size 11, 37, 35 and 38. The aim is to characterize the clinical status of the patients according to hotspots of genomic instability.

Model. We denote by $Y_{g\ell t}$ the observed signal at position t for patient ℓ in group g ($g = 1, \ldots, 4$). We use the following model:

$$\forall t \in I_k^{g\ell} \;\; Y_{g\ell t} = \mu_{g\ell k} + U_{gt} + E_{g\ell t} \tag{2}$$

No covariate is considered here. We consider homoskedastic errors with variance σ_0^2 and independent random effects with heteroskedastic variances σ_g^2. This variance heterogeneity means that correlation among profiles may be different from one clinical group to another.

Results. The number of segments is estimated using a penalized log-likelihood criterion from [6]. We obtain a total number of $\widehat{K} = 240$ segments spread in patients of groups 2, 3 and 4 (only 3 patients from group 1 present breakpoints). Figure 1 presents the results for groups 1 and 3, group 2 and 4 being very similar to group 3. Most patients from group 3 present genomic instabilities between positions 35 and 40. Only few patients from group 1 present instabilities on this interval.

If we do not account for correlations among series, *i.e.* if we remove the random effect, we obtain $\widehat{K}_0 = 230$ segments. Some breakpoints detected without the random effect at positions 36 and 85-86 vanish with the mixed model (Figure 1, top panel). This is illustrated by the two profiles displayed in Figure 2. The predictions of the random effect at these particular position are large (bottom panel). The predicted random effects in the different groups are present the same trend. Furthermore, the estimated variances are also very similar: $\widehat{\sigma}_g = (0.056, 0.052, 0.060, 0.058)$. This suggest that the random effect reveal some intrinsic characteristic of the position or of

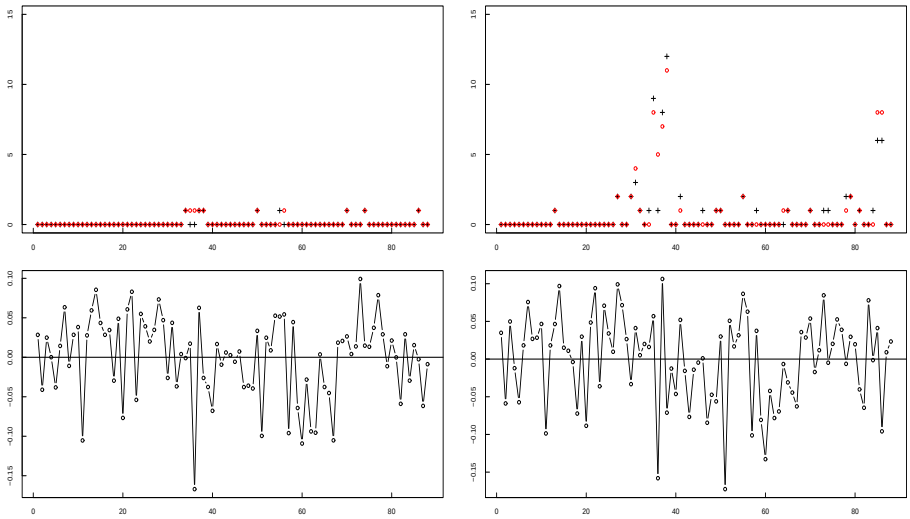

Figure 1: Segmentation results for groups 1 (left) and 3 (right). Top: number of patients having a breakpoint at each position with (+) or without (∘) random effect. Bottom: predicted random effect at each position.

the spot. Typically, the high value of $\widehat{U}$ at positions 36 and 86 is probably due to either a bad spot quality or a some annotation error regarding its position. Such spots do not reveal any biological information. They are not considered by the segmentation part of the model $\mathbf{T}\boldsymbol{\mu}$, but by the random part $\mathbf{ZU}$, which seems biologically relevant.

References

[1] J. Bai and P. Perron, *Computation and analysis of multiple structural change models*, J. Appl. Econ. **18** (2003), 1–22.

[2] H. Caussinus and O. Mestre, *Detection and correction of artificial shifts in climate series*, JRSS-C **53** (2004), no. 3, 405–425.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society Series B **39** (1977), 1–38.

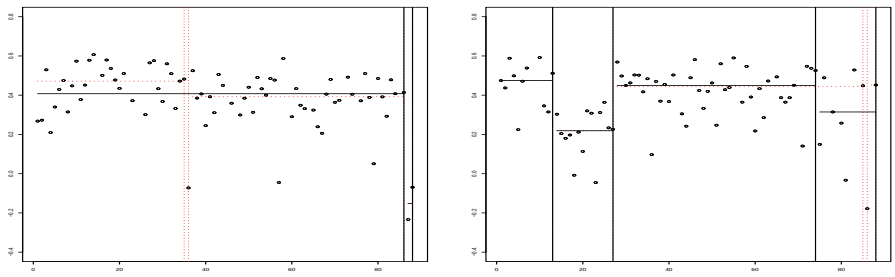

Figure 2: Segmentation results of one profile of group 1 (letf) and one of group 3 (right). Dotted line: segmentation with the random effect. Solid line: segmentation with the random effect.

[4] N. Dobigeon, J.-Y. Tourneret, and J. Scargle, *Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model*, IEEE Trans. Signal Processing **55** (2007), no. 1.

[5] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P. J. Park, *Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data*, Bioinformatics **0** (2005), no. 0, 1–8.

[6] E. Lebarbier, *Detecting multiple change-points in the mean of Gaussian process by model selection*, Signal Processing **85** (2005), 717–736.

[7] X.-L. Meng and D.B. Rubin, *Maximum likelihood estimation via the ecm algorithm: a general framework*, Biometrika **80** (1993), no. 2, 267–278.

[8] K. Nakao, K.R. Mehta, J. Fridlyand, D. H. Moore, A.J.Jain, A. Lafuente, J.W. Wiencke, J.P. Terdiman, and F.M. Waldman, *High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization*, Carcinogenesis **25** (2004), no. 8, 1345–1357.

[9] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, *A statistical approach for array CGH data analysis*, BMC Bioinformatics **6** (2005), no. 27, 1.

[10] A. M. Snijders, N. Nowak, R. Segraves, S. Blakwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law,

K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A.N. Jain, D. Pinkel, and D. G. Albertson, *Assembly of microarrays for genome-wide measurement of* DNA *copy number*, Nature Genetics **29** (2001), 263–264.

[11] D.A. van Dyck, *Fitting mixed-effects models using efficient em-type algorithms*, Jour. Comp. and Graph. Statistics **9** (2000), 78–98.