Bayesian growth curves using normal mixtures with nonparametric weights

Luisa Scaccia and Peter J. Green*
University of Bristol, UK.

19 February 2002

Summary

Reference growth curves estimate the distribution of a measurement as it changes according to some covariate, often age. We present a new methodology to estimate growth curves based on mixture models and splines. We model the distribution of the measurement with a mixture of normal distributions with an unknown number of components, and model dependence on the covariate through the weights, using smooth functions based on B-splines. In this way the growth curves respect the continuity of the covariate and there is no need for arbitrary grouping of the observations. The method is illustrated with data on triceps skinfold in Gambian girls and women.

Some key words: Allocation, Bayesian hierarchical model, Centile curves, Finite mixture distributions, Growth curves, Heterogeneity, Markov chain Monte Carlo, Normal mixtures, Path sampling, Reversible jump algorithms, Semiparametric model, Splines, Split/merge moves.

*Corresponding author: Email: P.J.Green@bristol.ac.uk,
Address: School of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

1 Introduction

Centile reference charts are an important screening tool in medical practice. The general form of a centile chart is a series of smoothed curves, showing how selected centiles for a biometrical measurement, such as height, weight or middle-upper-arm-circumference, change when plotted against some independent, appropriate covariate, often age. Here we will refer to these curves as centile or growth curves. This second name comes from the fact that such charts are used widely in pediatrics, for measurements related to growth and development.

On the basis of the centile curves for a certain biometrical measurement, it is possible to identify patients who are unusual, in the sense that their value for that measurement lies in the tails of the reference distribution. Centiles are usually chosen from a symmetric subset of the 3^{rd}, 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, 97^{th}.

The simplest way to draw the centile curves is to calculate the empirical centiles, after grouping or smoothing with respect to the covariate. If empirical centiles are used, however, the more extreme are estimated relatively inaccurately, as the centile standard errors increase steeply towards the tails of the distribution. This problem can be overcome by fitting a theoretical distribution to the data and thereby obtaining the expected centiles. In doing so, two main criteria must be met:

1. Generally the covariate concerned takes continuous values and the distribution of the measurements can be assumed to vary smoothly with the covariate; centile curves should therefore be constructed in such a way to respect this continuity, and arbitrary discretisation or grouping should be avoided.

2. Even if biometrical measurements are often approximately normally distributed, it seems more appropriate not to make strong distributional assumptions, or to assume a particular parametric form for the dependence on the covariate.

In order to address the second requirement we fit a theoretical distribution to the data using a finite mixture model. Mixture models provide an appealing semi-parametric structure in which to model heterogeneity and unknown distributional shapes. We refer to the monographs by Titterington *et. al.* (1985) and Böhning (2000) for general background. In the present context we consider a mixture of normal distributions and model the dependence of the observations on the covariate through the weights. We allow the weights to be indexed by the covariate, so that they can vary from observation to observation.

In doing so we also meet the first requirement. We model the weights as a smooth function of the covariate using B-splines. In this way the centile curves respect the continuity of the covariate and there is no need for arbitrary grouping of the observations. We refer to Green and Silverman (1994) and Wahba (1990) for a comprehensive discussion of splines. An alternative approach to the nonparametric estimation of growth curves was taken by Cole and Green (1992); see also the other references therein. Bayesian approaches include those of Geisser (1970) and Fearn (1975).

The paper is structured as follows. In Section 2 we present the Bayesian hierarchical mixture model proposed for the density estimate and the calculation of the expected centiles. Splines and their application in modelling the weights of the mixture are introduced in Section 3. Computational implementation via Markov chain Monte Carlo methods is discussed in Section 4. In Section 5, performance of the methodology is assessed through application to a real data set, and we conclude in Section 6 with general discussion and some possibilities for future work.

2 Mixture model

2.1 Normal Mixture

Let $y = (y_i)_{i=1}^n$ be observations of a biometrical variable we want to construct growth curves for, and $t = (t_i)_{i=1}^n$ the corresponding observed values of a continuous covariate (such as time or age). The model we assume for y is

$$y_i \sim \sum_{j=1}^{k} w_j(t_i)\phi(\cdot;\mu_j,\sigma_j) \qquad \text{independently for } i = 1, 2, \dots, n, \tag{1}$$

conditional on weights, means and variances, where $\phi(\cdot;\mu,\sigma)$ is the density of the $N(\mu,\sigma^2)$ distribution with $\mu = (\mu_j)_{j=1}^k$ and $\sigma = (\sigma_j)_{j=1}^k$.

The weights satisfy $w_j(t) \geq 0$ with $\sum_{j=1}^k w_j(t) = 1$ for all t and they are allowed to vary continuously with t. Let w_j be the n-vector $w_j(t_i)_{i=1}^n$, for each $j = 1, 2, \dots, k$, with w the $k \times n$ matrix of all $w_j(t_i)$.

The number of components k is unknown and subject to inference, as are μ, σ and w. Note that making inference about $w_j(t)$ *as functions* allows us to make predictions about future observations

y for values of t lying between the observed t_i.

Conditional on weights, means and variances, the $100\alpha^{th}$ centile can be numerically evaluated from (1) as that value $C_\alpha(t_i)$ for which

$$\alpha = \sum_{j=1}^{k} w_j(t_i) \int_{-\infty}^{C_\alpha(t_i)} \phi(x;\mu_j,\sigma_j)\,\mathrm{d}x = \sum_{j=1}^{k} w_j(t_i)\Phi\left(\frac{C_\alpha(t_i)-\mu_j}{\sigma_j}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative density for a standard normal distribution.

Note that we have chosen to model the weights as varying with t, while keeping the means and variances fixed; it is possible to consider other formulations with varying means and/or variances, modelled in a similar way to our treatment of the weights in Section 3 below, but we have not explored these in any detail.

It is worth stressing that we are using the mixture representation primarily as a convenient semi-parametric density estimation device, and we are not greatly interested in the number of components of the mixture per se, or in a clustering of the observations.

2.2 Latent allocation variables

An alternative perspective leading to the same mixture model (1) involves the introduction of latent allocation variables $z = (z_i)_{i=1}^n$ and the assumption that each observation y_i arose from an unknown component z_i of the mixture. The allocation variables are given probability mass function

$$p(z_i = j) = w_j(t_i) \qquad \text{independently for } i = 1, 2, \ldots, n, \tag{3}$$

and conditional on them, the observations y are independently drawn from the densities

$$y_i|z,\mu,\sigma \sim \phi(\cdot;\mu_{z_i},\sigma_{z_i}). \tag{4}$$

Integrating out z_i in (4) using the distribution in (3) leads back to (1).

2.3 Priors on component parameters

From past experience, we would not expect inference about the density in (1) to be highly sensitive to prior specification. As in Richardson and Green (1997), our prior assumptions are that the μ_j and σ_j^{-2} are all drawn independently, with normal and gamma priors

$$\mu_j \sim N(\xi,\kappa^{-1}) \qquad \text{and} \qquad \sigma_j^{-2} \sim \Gamma(\eta,\zeta),$$

where the latter is parametrised so that the mean and the variance are η/ζ and η/ζ^2 respectively.

The prior on the weights w_j will be discussed in section 3.2.

The number of components k will also be considered unknown and subject to inference. For this purpose, we assume for the number of components k a uniform prior on the values $\{1, 2, \ldots, k_{max}\}$, where k_{max} is a pre-specified integer. As in other mixture model contexts (or indeed in almost all model choice problems), it seems difficult to argue objectively for any specific prior for k. Our choice here is for similar reasons to those in Richardson and Green (1997), namely that with this choice it is easy to adjust results to get posteriors corresponding to other priors, by importance sampling (see, for example, Hammersley and Handscomb, 1964).

In order to allow for weakly informative priors for the model parameters, we introduce a hyperprior structure and hyperparameter choices which correspond to making only minimal assumptions on the data. Following Richardson and Green (1997) we take the $N(\xi, \kappa^{-1})$ prior for μ to be rather flat over the range of the data, by letting ξ equal to the mid-point of this range, and κ equal to a small multiple of $1/R^2$, where R is the length of the range.

For σ^2 we instead introduce an additional hierarchical level by allowing ζ to follow a Gamma distribution with parameters f and h, with $\eta > 1 > f$ and h a small multiple of $1/R^2$. This means that the support for σ^2 is not fixed *a priori* but determined by the value sampled for ζ.

3 Modelling dependence using splines

A particular feature of our mixture model is that the weights in (1) are evaluated at t_i, so that they are allowed to vary from observation to observation, according to the value recorded for the covariate. In this way, we introduce dependence on the covariate through the modelling of the weights. In particular we want to reflect the fact that observations corresponding to values of the covariate t not too far from each other, are somewhat similar. This is especially true when thinking of growth curves, for which the distribution of the biometrical measurement can be assumed to vary smoothly with the covariate. In our model this is achieved through requiring the weights (and thus also the allocation probabilities) to be continuous functions of t.

The model we propose makes use of a linear combination of cubic B-splines, after a suitable transformation of the weights.

3.1 Mathematical formulation

The weights $w_j(t)$ are constrained to be non-negative and sum to one for all t. For convenience, we impose the necessary constraints by transformation and, therefore, express the weights as

$$w_j(t) = \frac{\exp(g_j(t))}{\sum_{j'=1}^{k} \exp(g_{j'}(t))}, \tag{5}$$

where $g_j(t)$ is some continuous function of the covariate t. In order to allow the biometrical measurement to change smoothly and slowly with the covariate, we choose to model $g_j(t)$ and, thereby the weights $w_j(t)$, using natural cubic splines. For more details of the mathematical and statistical properties of cubic splines, needed in this and the following section, see Green and Silverman (1994, chapters 2 and 3).

Consider a set of distinct real numbers $t_1, \ldots, t_n$ (such as, for example, possible values of our covariate) on some interval $[a, b]$, satisfying $a < t_1 < t_2 < \ldots < t_n < b$. A function g defined on $[a, b]$ is a *cubic spline* on the *knots* t_i if:

i) g is a cubic polynomial on each of the intervals $(a, t_1), (t_1, t_2), (t_2, t_3), \ldots, (t_n, b)$;

ii) the polynomial pieces fit together at the points t_i in such a way that g itself and its first and second derivatives are continuous at each t_i, and hence on the whole of $[a, b]$.

The continuity of the second derivative is enough to give "visual smoothness" of the resulting function. Cubic splines can be specified in many equivalent ways. One of them is to give the four polynomial coefficients of each cubic piece, for example in the form

$$g(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i \qquad \text{for } t_i \le t \le t_{i+1}.$$

A cubic spline on an interval $[a, b]$ is said to be a *natural cubic spline* if its second and third derivatives are zero at a and b. These conditions, called the *natural boundary conditions*, imply that $d_0 = c_0 = d_n = c_n = 0$, so that g is linear on the two extreme intervals $[a, t_1]$ and $[t_n, b]$.

The importance of these functions derives from their variational characterisation: among all functions g on $[a, b]$ that are twice continuously differentiable and interpolate a given set of data points (t_i, y_i), where the t_i are distinct, that minimising the integrated-squared-second-derivative roughness penalty $\int (g_j''(t))^2 \mathrm{d}t$ is the unique interpolating natural cubic spline. This result motivates both the use of such splines in modelling smooth dependence, and the use of the prior specification in the following section.

Cubic splines are very convenient functions to deal with computationally; the existence of banded matrices in representations of interpolating and smoothing splines guarantees that computation times are $O(n)$ for n data points. However, the computational requirements of our Bayesian methodology are more demanding, and it is convenient to impose a finite-dimensional structure on the problem, by restricting the choice of $g(t)$ to the span of a prescribed set of basis functions, $\beta_1, \ldots, \beta_q$, and, thus, considering only functions $g(t)$ that can be expressed in the form

$$g(t) = \sum_{l=1}^{q} \gamma_l \beta_l(t)$$

for some numbers $\gamma_1, \ldots, \gamma_q$.

A possible choice for the basis functions is the set of natural cubic *B-splines* on a fixed grid of knots $s_1 < s_2 < \cdots < s_q$, usually taken to be equally spaced to cover the range of points t_i. The B-splines form a set of natural cubic splines that are non-negative and have only limited support: for $3 \leq l \leq q-2$ the function β_l is zero outside (s_{l-2}, s_{l+2}), whilst β_1, β_2, β_{q-1} and β_q are similar but linear outside (s_1, s_q). Restricting $g(t)$ to lie in the span of a set of, say, 10 B-splines typically has minimal impact on the quality of fit to the data.

Now, in our model, we have a function $g(t)$ for each component of the mixture. Modelling them as a linear combination of B-splines on the same knots $s_1 < s_2 < \cdots < s_q$, we can write them as

$$g_j(t) = \sum_{l=1}^{q} \gamma_{lj} \beta_l(t). \tag{6}$$

Let γ_j be the q-vector $(\gamma_{lj})_{l=1}^{q}$, for each $j = 1, 2, \ldots, k$, with γ the $q \times k$ matrix of all γ_{lj}.

3.2 Priors on weights

The prior on the weights is specified via that on g_j or γ_j, beginning with the natural integral-squared-second-derivative penalty

$$\lambda \int (g_j''(t))^2 \, \mathrm{d}t = \lambda \gamma_j^T K \gamma_j \tag{7}$$

for $j = 1, 2, \ldots, k$, where λ is a parameter always positive and K is the $q \times q$ matrix with $K_{lm} = \int \beta_l''(t) \beta_m''(t) \, \mathrm{d}t$. The prior $p(g_j) \propto \exp\{-(1/2)\lambda \gamma_j^T K \gamma_j\}$ would be "partially improper", since the matrix K has rank $q-2$; see Wahba (1978). Essentially the prior is invariant to the addition of a linear trend in t. Combined with the fact that while w_j are identifiable from the data, the g_j are not, this can cause problems with impropriety in the posterior. To circumvent these, the prior is converted to be proper by substituting the matrix K with a full rank matrix, and using

$$p(g_j) \propto \exp\{-(1/2)\gamma_j^T(\lambda K + \delta I)\gamma_j\} \tag{8}$$

(where I is the identity matrix and δ is a positive parameter).

The natural integral-squared-second-derivative $\int (g_j''(t))^2 \, \mathrm{d}t$ is a measure of the roughness of the curve $g_j(t)$. There are many ways of measuring how rough a curve is, but this is particularly appealing for different reasons. First of all, a natural requirement for any measure of roughness is that if two functions differ only by a constant or a linear function then their roughness should be identical. This logically leads to the idea of a roughness measure based on the second derivative of the curve under consideration. Secondly, the integral-squared-second-derivative has considerable computational advantages. Thirdly, there is the connection with the variational characterisation of natural cubic splines, mentioned in the previous section.

The role of λ is that of a smoothing parameter. As λ increases towards ∞, there is a stronger shrinkage of each γ_j towards a linear trend. As a result the curves $g_j(t)$ become smoother and so do the weights. Overall, the centile growth curves will also be very smooth and display little curvature. In the opposite limiting case, as $\lambda \to 0$ the centile growth curves will track the empirical ones very closely at the expense of being rather variable. This variability would affect particularly the more extreme centile curves and the accuracy of their estimates, as the centile standard errors increase steeply towards the tail of the distribution.

In this perspective, the model chosen for the weights introduces, through the natural integral-squared-second-derivative $\int (g_j''(t))^2 \, \mathrm{d}t$ and the smoothing parameter λ, a roughness penalising element and a trade-off between smoothness and goodness of fit of the centile curves.

The parameter δ can also be regarded as a smoothing parameter. As its value increases, not only there is a stronger shrinkage of the γ_j towards zero, but this effect is also reinforced by the fact that the variance for the γ_j is reduced and their distribution is more concentrated around the zero-mean.

Obviously the choice of the parameters λ and δ is of some importance. This matter will be discussed later in section 5.3.

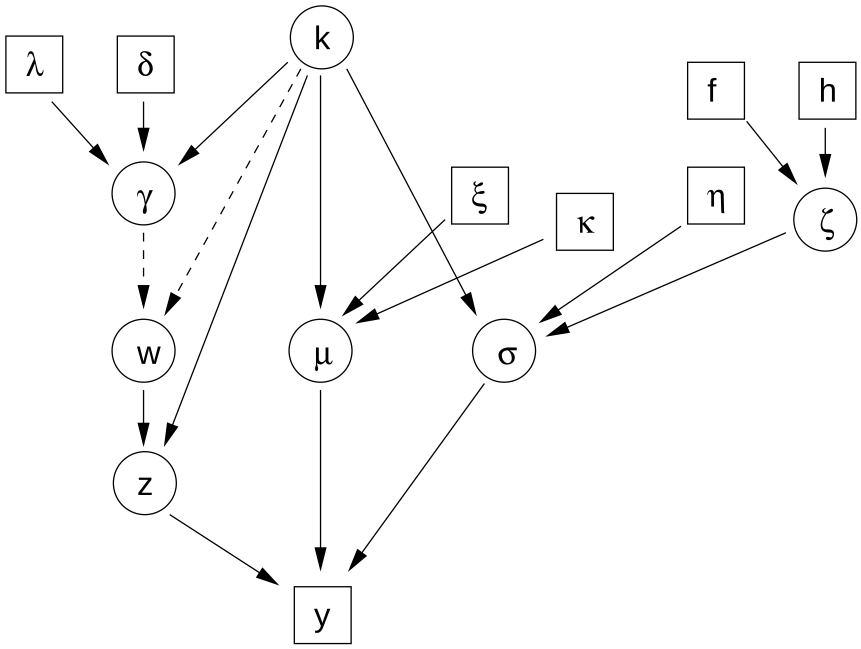

Figure 1: Directed acyclic graph for the complete hierarchical model

3.3 Complete hierarchical model

The joint distribution of all variables conditional on fixed hyperparameters may be written

$$p(k,\gamma,\mu,\zeta,\sigma,z,y|\lambda,\delta,\xi,\kappa,\eta,f,h) = p(k)p(\gamma|k,\lambda,\delta)p(\mu|k,\xi,\kappa)p(\zeta|f,h)p(\sigma|k,\eta,\zeta)p(z|\gamma,k)p(y|z,\mu,\sigma).$$

We have

$$p(z|\gamma,k) = \prod_{i=1}^{n} w_{z_i}(t_i),$$

with the relationship between w and γ given by (5), and

$$p(y|z,\mu,\sigma) = \prod_{i=1}^{n} \phi(y_i;\mu_{z_i},\sigma_{z_i}).$$

The prior distribution $p(\gamma|k,\lambda,\delta)$ is given in section 3.2, while $p(\mu|k,\xi,\kappa)$, $p(\zeta|f,h)$ and $p(\sigma|k,\eta,\zeta)$ are given in section 2.3. The complete hierarchical model is displayed in Figure 1 as a directed acyclic graph (DAG). We follow the usual convention that square boxes represent fixed or observed quantities and circles represent the unknowns. Relationships that are deterministic, as opposed to stochastic, are indicated by broken lines.

4 Computational implementation

The complexity of the mixture model presented requires Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution. Details of these computational methods can be found for example in Tierney (1994) and Besag *et. al.* (1995).

Our sampler uses five different fixed-dimension moves, each updating one of the variables of the model, plus a variable dimension move for updating k. The way in which the first five moves are performed is quite standard and thus we will go through them rather quickly. The last move, for updating k, is performed using a reversible jump method (Green, 1995).

4.1 Random walk Metropolis move for the weights

We can update the weights by means of a simultaneous random walk Metropolis method applied to γ. Thus we draw

$$\gamma'_{lj} \sim N(\gamma_{lj}, \tau^2)$$

independently, compute the corresponding weights $w'_j(t_i)$, and accept this proposal with the usual probability equal to $\min\{1, Q\}$ where

$$Q = \frac{p(k, \gamma', \mu, \zeta, \sigma, z, y)}{p(k, \gamma, \mu, \zeta, \sigma, z, y)} = \frac{p(\gamma'|k)}{p(\gamma|k)} \frac{p(z|\gamma', k)}{p(z|\gamma, k)},$$

which simplifies into

$$Q = \exp\left(-(1/2)\sum_{j=1}^{k}[(\gamma'_j)^T(\lambda K + \delta I)\gamma'_j - \gamma_j^T(\lambda K + \delta I)\gamma_j]\right)\prod_{i=1}^{n}\frac{w'_{z_i}(t_i)}{w_{z_i}(t_i)}.$$

Faster convergence in the algorithm was obtained introducing a small bias in the proposal. Thus, instead of proposing to update γ_{lj} to the new value

$$\gamma'_{lj} = \gamma_{lj} + r,$$

where r is a random number from a $N(0, \tau^2)$, we propose, as a new value for γ_{lj},

$$\gamma'_{lj} = \gamma_{lj} + r + \varrho \sum_{i:z(i)=j} \beta_l(t_i),$$

where ϱ is a small number (we used $\varrho = 0.001$ on the basis of some limited pilot runs). The acceptance ratio for this proposal then becomes

$$Q' = \frac{p(\gamma'|k)}{p(\gamma|k)} \frac{p(z|\gamma', k)}{p(z|\gamma, k)} \frac{p(\gamma'|\gamma)}{p(\gamma|\gamma')} = Q \exp\left(-\frac{2\varrho}{\tau^2}\sum_{j=1}^{k}\sum_{l=1}^{q}(\gamma'_j - \gamma_j)\sum_{i:z(i)=j}\beta_l(t_i)\right),$$

where $p(\gamma'|\gamma)$ is the probability of proposing γ' when the current value is γ and $p(\gamma|\gamma')$ is the probability of proposing γ when the current value is γ'.

The biased proposal has the overall effect of proposing new weights which tend slightly to favour the current allocation of the observations.

4.2 Gibbs move for the allocations

For the allocations we have

$$p(z|\gamma,\mu,\sigma,k,y) \propto p(\gamma,\mu,\sigma,z,k,y) \propto p(z|\gamma,k)p(y|z,\mu,\sigma)$$

(that is, proportional as a function of z), so the allocation variable z_i has conditional probability

$$p(z_i = j|\gamma,\mu,\sigma,k,y) = \frac{w_j(t_i)\phi(y_i;\mu_j,\sigma_j)}{\sum_{j'} w_{j'}(t_i)\phi(y_i;\mu_{j'},\sigma_{j'})}. \tag{9}$$

We can sample directly from this distribution and update the allocation variables independently by means of Gibbs sampling.

4.3 Moves for the parameters and the hyperparameter

Updating μ

Before considering the updating of the μ_j, we comment briefly on the issue of labelling the components. The whole model is, in fact, invariant to permutation of the labels $j = 1, 2, \ldots, k$. For identifiability, Richardson and Green (1997) adopt a unique labelling in which the μ_j are in increasing numerical order. As a consequence the joint prior distribution of the μ_j is $k!$ times the product of the individual normal densities, restricted to the set $\mu_1 < \mu_2 < \ldots < \mu_k$.

The μ_j can be updated by means of Gibbs sampler, drawing them independently from the distribution

$$\mu_j|\cdots \sim N\left(\frac{\sigma_j^{-2}\sum_{i:z_i=j} y_i + \kappa\xi}{\sigma_j^{-2}n_j + \kappa}, (\sigma_j^{-2}n_j + \kappa)^{-1}\right),$$

where $n_j = \#\{i : z_i = j\}$ is the number of observations currently allocated to the j component of the mixture. Here and later, '$\cdots$' denotes 'all other variables'. In order to preserve the ordering constraints on the μ_j, the move is accepted provided the ordering is unchanged and rejected otherwise.

Except for very small values of k, this updating move has the drawback of producing a very small acceptance ratio, due to the fact that the ordering of the μ_j seldom remains unchanged. For this reason it is preferable to update μ_j using a trick similar to the one that Green and Richardson (2000) adopt to update their component risk parameters. We propose simultaneous independent zero-mean normal increments to each μ_j; the modified values of μ_j are then placed in increasing order to give μ' say. The complete set of updates is accepted with probability, formed from prior ratio and likelihood ratio, which reduces to $\min\{1, S\}$ where

$$\mathrm{S} = \exp\left\{\sum_{j=1}^{k}\left[-\frac{\kappa}{2}\left((\mu_j'^2 - \mu_j^2) - 2\xi(\mu_j' - \mu_j)\right) - \sum_{i:z_i=j}\frac{1}{2\sigma_{z_i}^2}\left((\mu_{z_i}'^2 - \mu_{z_i}^2) - 2y_i(\mu_{z_i}' - \mu_{z_i})\right)\right]\right\}.$$

An alternative to imposing identifiability constraints on the parameters, is to order the parameters, according to some unique labelling, *a posteriori*, after the whole sample of parameters drawn from the posterior is available. In the present case, where the main concern is the inference on the posterior density of the data and its centiles, rather than on the single parameters of the model, this labelling is not even required. In this case Gibbs sampler can be used for updating μ_j without any need for their order to stay unchanged.

For reason of completeness and because of some interest, however, we preferred to make inference also on the single parameters of the model and we decided to use the first approach (i.e. to impose an ordering on the μ_j *a priori*) after having obtained much the same results from both of them.

Updating σ

The full conditionals for σ_j^2 are

$$\sigma_j^{-2}|\cdots \sim \Gamma\left(\eta + \frac{1}{2}n_j, \zeta + \frac{1}{2}\sum_{i:z_i=j}(y_i - \mu_j)^2\right).$$

We update σ_j^2 independently using a Gibbs move, sampling from their full conditionals.

Updating ζ

The only hyperparameter we are not treating as fixed is ζ. Conditional on all the other parameters and the data, ζ has a Gamma distribution

$$\zeta|\cdots \sim \Gamma(f + k\eta, h + \sum_{j=1}^{k}\sigma_j^{-2}).$$

We update ζ by a Gibbs move, sampling from its full conditional.

4.4 Variable dimension move for updating k

Updating the value of k implies a change of dimensionality for the components μ and σ, the allocation variables z and the weights w (through the change of dimensionality for g and γ). We follow the approach used by Richardson and Green (1997) consisting of a random choice between splitting an existing component into two, and merging two existing components into one. The probabilities of these alternatives are b_k and $d_k = 1 - b_k$, respectively, when there are currently k components. Of course, $d_1 = 0$ and $b_{k_{max}} = 0$, and otherwise we choose $b_k = d_k = 0.5$, for $k = 2, 3, \ldots, k_{max} - 1$.

For the combine proposal we randomly choose a pair of components (j_1, j_2) that are adjacent in terms of the current value of their means, which means $\mu_{j_1} < \mu_{j_2}$, with no other μ_j in the interval $[\mu_{j_1}, \mu_{j_2}]$. These two components are merged into a new one, labelled $\mu_{j^\star}$, reducing k by 1. We then reallocate all those observations y_i with $z_i = j_1$ or j_2 to the new component $j^\star$ and create values for $\mu_{j^\star}, \sigma_{j^\star}$ in such a way that:

$$\begin{aligned}
\mu_{j^\star} &= (\mu_{j_1} + \mu_{j_2})/2 \\
\mu_{j^\star}^2 + \sigma_{j^\star}^2 &= [(\mu_{j_1}^2 + \sigma_{j_1}^2) + (\mu_{j_1}^2 + \sigma_{j_2}^2)]/2.
\end{aligned}$$

To create the new values $w_{j^\star}(t_i)$ we first have to create $\gamma_{lj^\star}$, for $l = 1, 2, \ldots, q$. We do this by setting

$$\gamma_{lj^\star} = \log[\exp(\gamma_{lj_1}) + \exp(\gamma_{lj_2})], \qquad \text{for } l = 1, 2, \ldots, q$$

and calculate $g_j^\star(t_i)$ and $w_j^\star(t_i)$ using (6) and (5), respectively. Note that in this way all the weights change slightly.

The split proposal starts by choosing a component $j^\star$ at random. This component is split into two new ones labelled j_1 and j_2, augmenting k by 1. Then we have to reallocate all those observations y_i with $z_i = j^\star$ between the two new components, and create values for $(w_{j_1}, w_{j_2}, \mu_{j_1}, \mu_{j_2}, \sigma_{j_1}, \sigma_{j_2})$. Let us start by splitting $\mu_{j^\star}$ and $\sigma_{j^\star}$. We generate a two-dimensional random vector v to specify the new parameters. We use Beta distributions

$$v_1 \sim Be(1,1) \qquad \text{and} \qquad v_2 \sim Be(1,1)$$

for this and set

$$\mu_{j_1} = \mu_{j^\star} - v_1\sigma_{j^\star} \quad , \quad \mu_{j_2} = \mu_{j^\star} + v_1\sigma_{j^\star},$$
$$\sigma^2_{j_1} = 2v_2(1-v_1^2)\sigma^2_{j^\star} \quad \text{and} \quad \sigma^2_{j_2} = 2(1-v_2)(1-v_1^2)\sigma^2_{j^\star}.$$

We then need to check that the constraints on the means are satisfied. If not the move is rejected forthwith, as the mis-ordered vector μ has zero density under the ordered prior. If the constraints are satisfied we move on and split the weights.

In doing so, we generate a q-dimensional vector u from

$$u_l \sim Be(0.5, 0.5), \qquad \text{independently for } l = 1, 2, \ldots, q,$$

and we set

$$\gamma_{lj_1} = \gamma_{lj^\star} + \log(u_l) \quad , \quad \gamma_{lj_2} = \gamma_{lj^\star} + \log(1-u_l) \qquad \text{for } l = 1, 2, \ldots, q.$$

We then calculate $g_{j_1}(t_i)$, $g_{j_2}(t_i)$ and $w_{j_1}(t_i)$, $w_{j_2}(t_i)$ using (6) and (5), respectively. We denote the proposed new weights by w'.

Finally we reallocate those y_i with $z_i = j^\star$ between j_1 and j_2 in a way analogous to the standard Gibbs allocation move; see equation (9). We denote the proposed new allocation vector by z'.

According to the reversible jump framework, the acceptance probability for the split move is $\min(1, A)$, where

$$\begin{aligned}
A \;=\; & (\text{likelihood ratio}) \times \frac{p(k+1)}{p(k)} \times (k+1) \times \prod_{i=1}^{n} \frac{w'_{z'_i}(t_i)}{w_{z_i}(t_i)} \\
\times\; & \sqrt{\frac{\kappa}{2\pi}} \exp\left\{-\frac{\kappa}{2}\left[(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 - (\mu_{j^\star}-\xi)^2\right]\right\} \\
\times\; & \frac{\zeta^\eta}{\Gamma(\eta)} \exp\left[-\zeta\left(\frac{1}{\sigma^2_{j_1}} + \frac{1}{\sigma^2_{j_2}} - \frac{1}{\sigma^2_{j^\star}}\right)\right]\left(\frac{\sigma_{j^\star}}{\sigma_{j_1}\sigma_{j_2}}\right)^{2(\alpha-1)} \\
\times\; & \frac{|\lambda K + \delta I|^{1/2}}{(2\pi)^{q/2}} \exp\left\{-\frac{1}{2}\left[\gamma_{j_1}^T(\lambda K + \delta I)\gamma_{j_1} + \gamma_{j_2}^T(\lambda K + \delta I)\gamma_{j_2} - \gamma_{j^\star}^T(\lambda K + \delta I)\gamma_{j^\star}\right]\right\} \\
\times\; & \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \left[be_{1,1}(v_1)be_{1,1}(v_2)\prod_{l=1}^{q} be_{0.5,0.5}(u_l)\right]^{-1} \\
\times\; & \frac{8(1-v_1^2)\sigma^3_{j^\star}}{\prod_{l=1}^{q} u_l(1-u_l)},
\end{aligned} \tag{10}$$

where $\Gamma(\cdot)$ is the Gamma function, P_{alloc} is the probability of this particular allocation, $be_{p,r}$ denotes the Beta(p,r) density, and (likelihood ratio) is the ratio of the product of the $f(y_i|z_i, \mu_{z_i}, \sigma_{z_i})$ terms for the new parameter set to that for the old one. The quantity $|\lambda K+\delta I|$ represents the determinant of the inverse of the covariance matrix for each of the g_j. The need to make the prior distribution on the g_j proper and to introduce the positive parameter δ is now evident.

The first four lines of (10) are the product of the likelihood ratio and the prior ratio for the parameters of the model. The fifth line is the proposal ratio. The last line is the Jacobian of the transformation from the vector $(\mu_{j^\star}, \sigma_{j^\star}, \gamma_{1j^\star}, \ldots, \gamma_{qj^\star}, v_1, v_2, u_1, \ldots, u_q)$ to the vector $(\mu_{j_1}, \sigma_{j_1}, \gamma_{1j_1}, \ldots, \gamma_{qj_1}, \mu_{j_2}, \sigma_{j_2}, \gamma_{1j_2}, \ldots, \gamma_{qj_2})$.

The acceptance probability for the combine move is $\min(1, A^{-1})$, with some obvious substitutions in the expression for A.

4.5 Within-model simulation and path sampling

The alternative general approach to sample-based joint inference about a model indicator k and model parameters is to conduct separate simulation within each model, and piece the results together afterwards. Inference about k is then based on the estimate of the posterior model probabilities

$$p(k|Y) \propto p(k)p(Y|k).$$

These require estimates of the marginal likelihoods $p(Y|k)$ separately for each k, using individual MCMC runs. Let $\psi_k = (\mu, \sigma, w)$ denote the unknowns μ, σ and w in the model with k components. Then, possible estimates of marginal likelihoods, using importance sampling, are for example (see Newton and Raftery (1994))

$$\widehat{p}_1(Y|k) = M / \sum_{m=1}^{M} \left\{ p(Y|k, \psi_k^{(m)}) \right\}^{-1},$$

based on an MCMC sample $\psi_k^{(1)}, \psi_k^{(2)}, \ldots, \psi_k^{(M)}$ from the posterior $p(\psi_k|Y,k)$,

$$\widehat{p}_2(Y|k) = M^{-1} \sum_{m=1}^{M} p(Y|k, \psi_k^{(m)}),$$

based on a sample from the prior $p(\psi_k|k)$, or

$$\widehat{p}_3(Y|k) = \frac{\sum_{m=1}^{M} p(Y|k, \psi_k^{(m)}) / \left\{ \epsilon \widehat{p}_3(Y|k) + (1-\epsilon) p(Y|k, \psi_k^{(m)}) \right\}}{\sum_{m=1}^{M} \left\{ \epsilon \widehat{p}_3(Y|k) + (1-\epsilon) p(Y|k, \psi_k^{(m)}) \right\}^{-1}},$$

based on a sample from a $(\epsilon, 1-\epsilon)$ mixture of the prior $p(\psi_k|k)$ and the posterior $p(\psi_k|Y,k)$ and supposed to perform better than the previous two. It is well known, however, that when the distance between the prior and the posterior densities is big (as it is in our case), the variability of this last estimator can become so large that the estimate is virtually unusable. A gain of efficiency can be obtained using the idea of the bridge sampling and choosing a sensible density which serves as a "bridge" between the prior and the posterior densities. The idea of creating a bridge can obviously be pushed further if the prior and the posterior densities are so far separated that the estimator based on the bridge sampling is too variable to use in practice. In such cases it is useful to construct an infinite series of intermediate densities, i.e. a whole "path", from which we can make draws. We refer to Gelman and Meng (1998) for a detailed discussion of bridge sampling and path sampling.

Briefly, to estimate the marginal likelihood, we construct a geometric path between the prior and the posterior parameter densities using a scalar parameter $\theta \in [0,1]$,

$$q(\psi_k|\theta) = \{p(\psi_k|k)\}^{1-\theta}\{p(\psi_k|k)p(Y|\psi_k,k)\}^{\theta} = p(\psi_k|k)\{p(Y|\psi_k,k)\}^{\theta}. \tag{11}$$

We use a notation in which $q(\cdot)$ represents an un-normalised density, $p(\cdot)$ is the corresponding normalised density and $c(\cdot)$ is the normalising constant. From (11) it is evident that estimating $p(Y|k)$ is equivalent to estimating $c(1)$, i.e. the normalising constant when $\theta = 1$. Following Gelman and Meng (1998) it can be proved that

$$\log[p(Y|k)] = \log[c(1)] = \int_0^1 E_\theta[U(\psi_k,\theta)]\,\mathrm{d}\theta, \tag{12}$$

where E_θ denotes the expectation with respect to the sampling distribution $p(\psi_k|\theta)$ (which is the unknown normalised version of $q(\psi_k|\theta)$) and where

$$U(\psi_k,\theta) = \frac{d}{d\theta}\log q(\psi_k|\theta).$$

For estimating $\log[p(Y|k)]$ we then numerically evaluated the integral in (12) over a grid of values for θ. We chose an exponential spacing for the grid to account for the sharp variation of $q(\psi_k|\theta)$ for θ close to 0. Given the indexing $0 = \theta_{(1)} < \ldots < \theta_{(h)} < \ldots < \theta_{(H)} = 1$ and a (possibly dependent) sample of draws $(\psi_k^{(m)}, \theta^{(m)})$ from $p(\psi_k, \theta)$, applying the trapezoidal rule, we estimate $\log[p(Y|k)]$ by

$$\log[\widehat{p}(Y|k)] = \log\widehat{c}(1) = \frac{1}{2}\sum_{h=1}^{H-1}(\theta_{(h+1)} - \theta_{(h)})(\bar{U}_{(h+1)} + \bar{U}_{(h)}), \tag{13}$$

where $\bar{U}_{(h)}$ is the average of the values of $U(\psi_k^{(m)}, \theta^{(m)})$ for all simulation draws m for which $\theta^{(m)} = \theta_{(h)}$.

We now come to the matter of obtaining an MCMC sample of draws $(\psi_k^{(m)}, \theta^{(m)})$ from the joint distribution $p(\psi_k, \theta)$, using a different MCMC sample from that used to compute the within-model posterior distribution of the parameters. We can easily draw ψ_k from $p(\psi_k|\theta)$, but the problem of updating θ is more complicated. Assuming a discrete uniform distribution for θ on the values specified in the grid, we have

$$p(\theta|\psi_k) \propto p(\theta)p(\psi_k|\theta) = p(\theta)\frac{q(\psi_k|\theta)}{c(\theta)} \propto \frac{q(\psi_k|\theta)}{c(\theta)},$$

and $c(\theta)$ is unknown.

A first solution would be to use nested loops: for each value θ in the grid, run an iterative simulation algorithm until approximate convergence to obtain $p(\psi_k|\theta)$. The problem with this solution is that when the grid for θ is very fine, as in the present case, many runs of the algorithm are required and convergence needs to be assessed for each of them.

We thus preferred to draw from the joint density of (ψ_k, θ), combining the simulations of ψ_k and θ in a single loop of iterative simulation, alternately updating ψ_k and θ. In this way, convergence can be checked looking at the sampled distribution for θ which should be uniform on the values specified in the grid. Also, we avoid the need to do burn-in runs for each θ in the grid, separately. To overcome the problem of drawing θ from $p(\theta|\psi_k)$ we used a rough estimate $c^\star(\theta)$ instead of $c(\theta)$ and simulated θ with target

$$p(\theta)\frac{q(\psi_k|\theta)}{c^\star(\theta)} = \left(p(\theta)\frac{c(\theta)}{c^\star(\theta)}\right) p(\psi_k|\theta),$$

so that the output distribution for θ is only slightly altered (provided $c^\star(\theta)$ is not too far from $c(\theta)$) while the output distribution for $(\psi_k|\theta)$ is unaltered. We proceeded starting from a first rough estimate $c^\star(\theta)$. This is obtained keeping $\theta = \theta_{(h')}$ fixed for 20 sweeps of the algorithm, updating only ψ_k, estimating $c^\star(\theta_{(h')})$ from these 20 sweeps using (13) where the sum is over $h = 1, \ldots, h'-1$, and then moving to $\theta = \theta_{(h'+1)}$. Notice that there is no need to estimate $c^\star(0)$ since $c(0) = 1$. After $c^\star(\theta)$ was calculated for each θ a further 100 000 sweeps were used to draw values of θ and ψ_k from their joint distribution and to re-estimate $c^\star(\theta)$ after an exponentially increasing number of sweeps ($m = 67, 91, 168, \ldots, 73\ 783, 100\ 000$). We then used the last estimate of $c^\star(\theta)$ to run the final 100 000 iterations to be used to estimate $p(Y|k)$ from (13). The whole procedure is quite cumbersome. It is however required to get values of $c^\star(\theta)$ as close as possible to those of $c(\theta)$. This is necessary to obtain an output distribution of θ not too far from the uniform and to ensure, in this way, that the averages $\bar{U}_{(h)}$ are calculated over a considerable number of values for each h.

5 Results

The method for determining growth curves illustrated so far is applied to data on triceps skinfold in Gambian females. These data come from an anthropometry survey of 892 girls and women up to age 50 in three Gambian villages, seen during the dry season of 1989; 620 (70 %) of the subjects were aged under 20. Five different anthropometric measurements were taken. Here we only discuss the triceps skinfold for a subset of the original data including 584 subjects aged from about 3 to 26 years. In this covariate range, the observations on triceps skinfold vary in the interval $[3.2; 21.0]$, determining a length R for the range of the data equal to 17.8. This dataset has already been analysed in Cole and Green (1992) and Green and Silverman (1994, §6.7).

The results reported correspond to runs of 100 000 sweeps after a burn-in period of 50 000 sweeps. The following settings were used for previously unspecified constants: $\kappa = 10.0/R^2$, $\eta = 2.0$, $f = 0.2$, $h = 10.0/R^2$, $\lambda = 0.7$, $\delta = 0.09$. Some experimentation indicates that results are quite robust to reasonable perturbations of these values.

Models with a number of components up to $k_{max} = 5$ were considered. The number of knots was fixed at 10, with the knots equally spaced between 5 and 23.

5.1 Posterior inference on k

The reversible jump approach allows joint (or across-model) inference about the number of components k and the other parameters of the model. The posterior distribution $p(k|y)$ can be, therefore, directly obtained from the MCMC sample. Unfortunately, in the present case, a few runs of the algorithm showed an extremely small acceptance rate of the split/combine move, approximately equal to 0.06 %. This is partly due to the large number of parameters involved in the model, that makes it difficult to move from a state of dimension $12k$ to a new one with dimension $12(k-1)$ or $12(k+1)$. However, the main reason for such bad performance is the large size of the data set.

This leads to very peaked posterior distributions for the parameters, causing a very low acceptance rate of the split/combine proposals.

An improvement in the acceptance rate was found considering one quarter of the original data set: the data were sorted according to the value of t_i and a reduced data set was created taking one datum at random from every consecutive four, in order to preserve the original structure of the data. In this way the acceptance rate increases 10 times and a further increase was obtained repeating the split/combine move more than once in each sweep and considering the move accepted overall, if accepted at least once.

Reversible jump was therefore used on the reduced data set to simulate a "partial" posterior distribution for k. The mode of this distribution was considered as an estimate of the number of components for the mixture. The MCMC algorithm was then run again to obtain the density estimate for the whole data set, skipping this time the variable dimension move and fixing k to its estimated value. This is not 'using the data twice', but is a valid approach to approximating the posterior of all parameters conditional on $k = k^\star$, where $k^\star$ maximises $p(k|y)$; the probability that $k^\star$ differs between the reduced and full data sets is negligible.

In addition to the across-model approach on the partial data set, we also tried the within-model approach of Section 4.5. We mainly did so in order to check the estimate of k we obtained from the reduced data set and also to explore an alternative solution when the amount of data makes the use of reversible jump on the whole data set infeasible, at least with the moves we have considered.

Even proceeding in this way, the estimate still shows some variability. In order to reduce this variability, the final estimate for each model was therefore obtained as a trimmed mean of four estimates (i.e. the mean of the middle two of the four) resulting from four different runs of the algorithm. Figure 2 shows the posterior distribution $p(k|Y)$ obtained in this way for the whole data set (continuous line) and for the partial data set (dashed line) compared with the one resulting from the reversible jump sample on the partial data set (dotted line). In spite of their variability, the different estimates agree on the mode of the posterior distribution for k and favour an explanation of the data using 4 components.

A further visual comparison to evaluate the goodness of fit of the different models can be based on the plot of the cumulative density of the data against the covariate. If the data are well fitted by the model, the points must be uniformly spread (in the vertical direction) over the plot. The

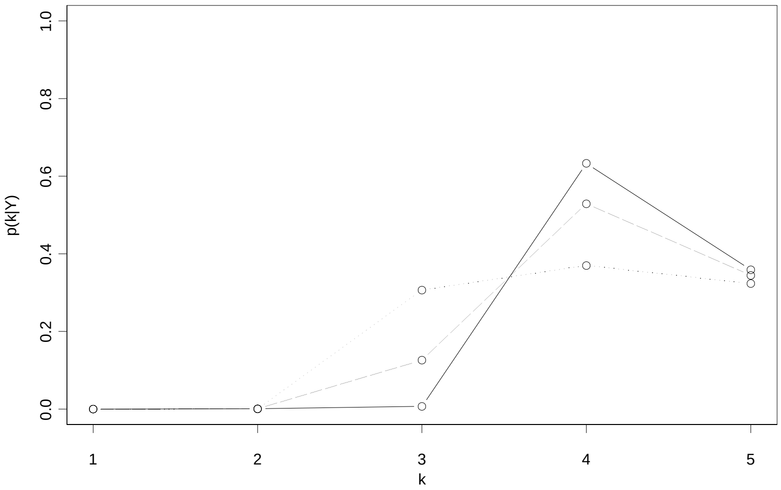

Figure 2: Posterior distribution $p(k|Y)$ estimated using path sampling on the whole data set (continuous line) and on the partial data set (dashed line), and using reversible jump on the partial data set (dotted line).

cumulative density, conditional on weights, means and variances can be written as

$$F(y_i) = \sum_{j=1}^{k} w_j(t_i) \int_{-\infty}^{y_i} \phi(x; \mu_j, \sigma_j)\, \mathrm{d}x = \sum_{j=1}^{k} w_j(t_i) \Phi \left(\frac{y_i - \mu_j}{\sigma_j} \right).$$

The cumulative density for all t_i was computed at each sweep and then averaged over the number of sweeps. Figure 3 shows these cumulative densities plotted against the age for the five different models. Clearly the models with 1 and 2 components show their limits in fitting the data, while the models with 3 to 5 components seem to be all reasonable. In the bottom right corner, the fit corresponding to the across-model inference is also shown.

5.2 Posterior inference for centile curves

We briefly illustrate how to evaluate the centile curves from the MCMC output. First of all, centile curves were not evaluated for every t_i but only over a grid of 47 equally spaced values for t, between 4.0 and 26.0. We indicate these values as $t^\star = (t_i^\star)_{i=1}^{n^\star}$, where $n^\star = 47$. This was done to reduce the computing time required to evaluate the centiles numerically at each sweep of the algorithm.

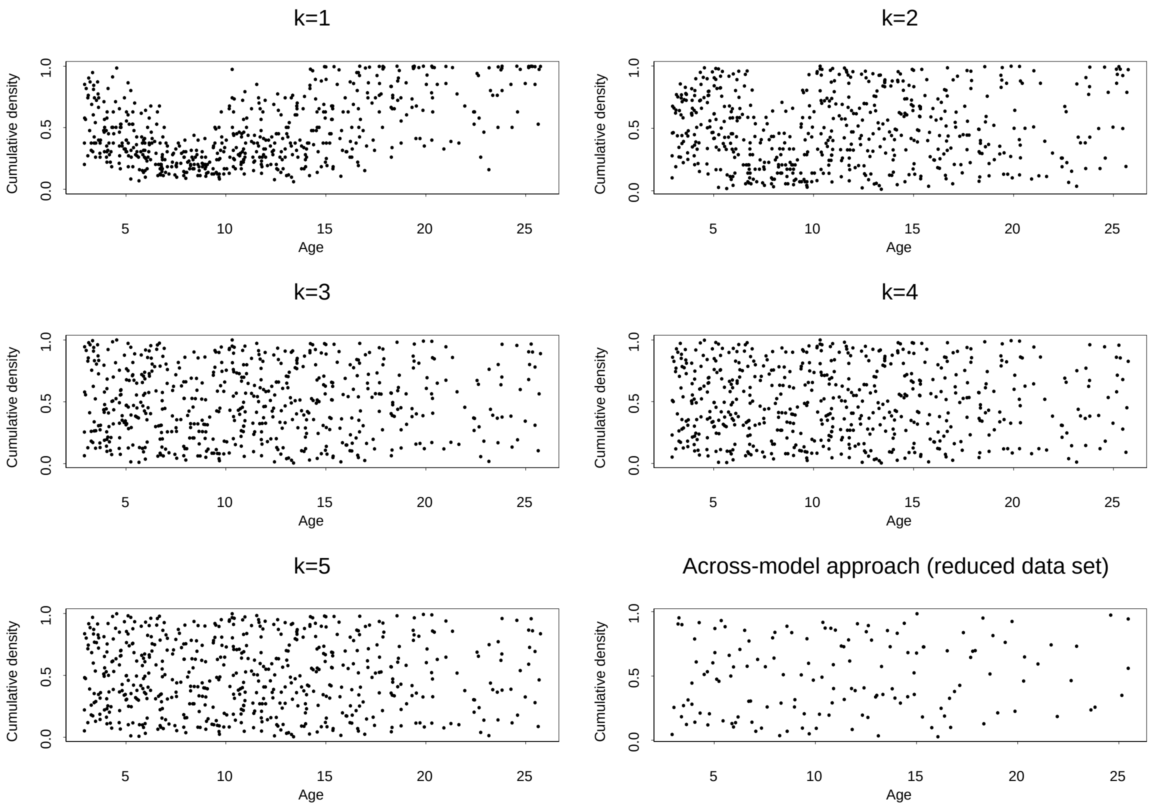

Figure 3: Cumulative density for y_i plotted against t_i.

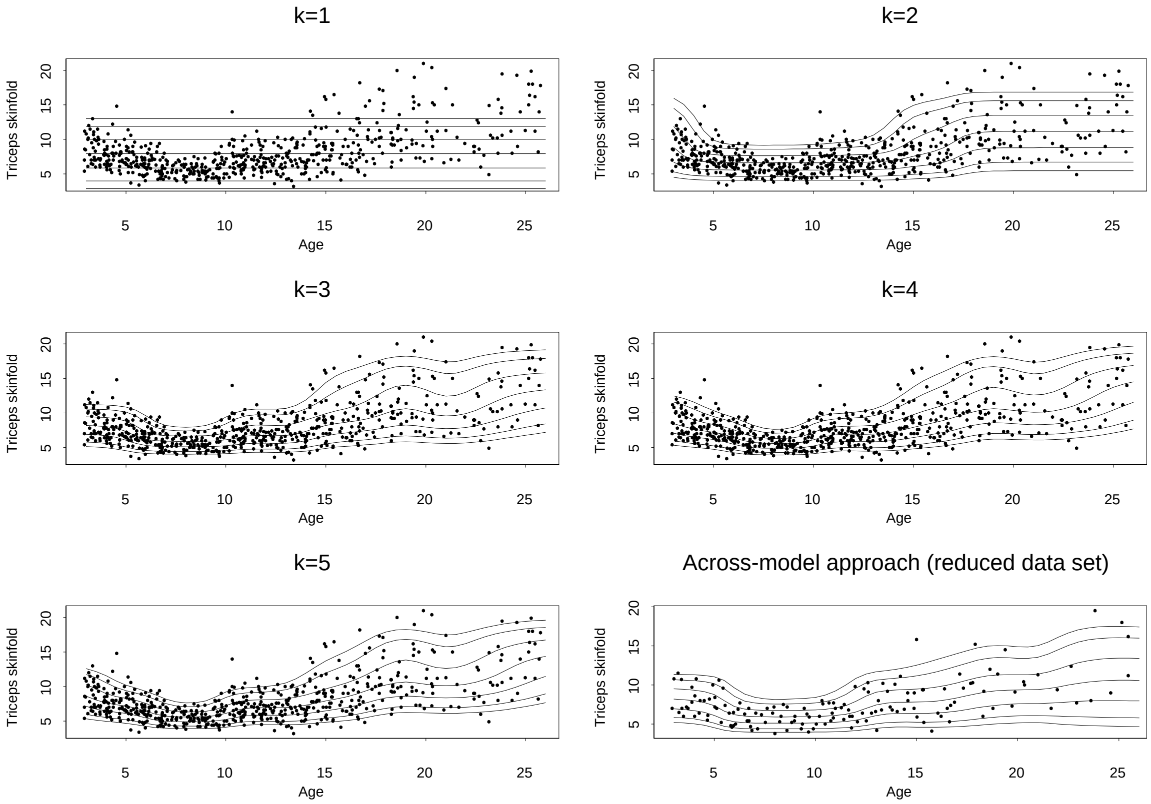

Figure 4: Centiles curves obtained with the five different models for triceps skinfold among Gambian females: 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles.

Then we set a grid of values $a = (a_h)_{h=1}^{200}$. For each sweep $m = 1, 2, \ldots, 100\ 000$ and for each value $t_i^\star$, and recalling (2), we evaluate, for a given value α, the centile $C_\alpha^{(m)}(t_i^\star)$ as that value a_h for which

$$\sum_{j=1}^{k} w_j^{(m)}(t_i^\star)\Phi\left(\frac{a_h - \mu_j^{(m)}}{\sigma_j^{(m)}}\right) \leq \alpha < \sum_{j=1}^{k} w_j^{(m)}(t_i^\star)\Phi\left(\frac{a_{h+1} - \mu_j^{(m)}}{\sigma_j^{(m)}}\right).$$

The centiles $C_\alpha^{(m)}(t_i^\star)$ are then averaged over the different sweeps to obtain finally $C_\alpha(t_i^\star)$. Figure 4 shows the centile curves obtained for the five different models. The results for the model with 1 component are given only for reasons of completeness and as a visual aid, to see how the fitting of the data changes as the number of components increases. It expresses no dependence of the data on the covariate and simply models the data as a normal distribution with some estimated mean and variance. Obviously it has no pretension of explaining the data. The 2 components mixture seems to be not completely adequate to model the density of the data, either, while mixtures with 3 to 5 components show similar results in terms of growth curves. In the bottom right corner of Figure 4, the centile curves estimated for the partial data set, using the across-model approach are also given. In this case inference results from mixing over k, while inference on the other centile curves is conditional upon k.

Data are characterised by high triceps values in early childhood, followed by a fall and then a second continuous rise until the end of the age interval. The centile curves obtained using the three models with 3 to 5 components closely follow this pattern. They show the same "notch" in the dependence at around 9 years, found in Cole and Green (1992). They also show a gradual increase in the spread of the data after age 10.

The across-model approach gives much the same results, with some obvious differences due to the fact that here only one quarter of the data are involved in the model fitting.

The centile curves are reasonably well smoothed up to age 20. After that they appear somewhat ragged. This might well be due to the fact that data for women aged more than 20 are very sparse. Smoother curves could be obtained by increasing the value of the smoothing parameters λ or δ.

5.3 The choice of the smoothing parameters

The choice of the smoothing parameters λ and δ obviously involve an arbitrary decision. A possible solution to overcome this problem would be to assign a hyperprior to λ and δ. This would allow

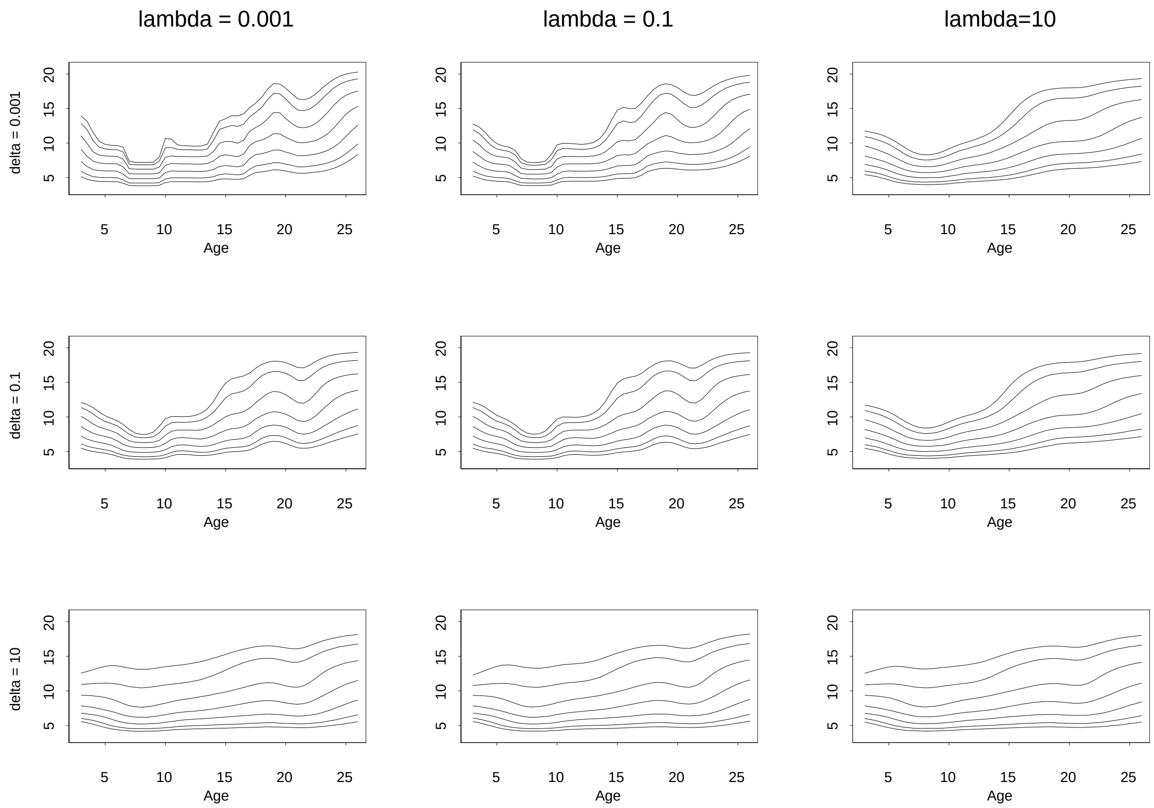

Figure 5: Centile curves obtained for different values of the smoothing parameters λ and δ, conditioning on $k = 4$.

extra flexibility to the model and cope with uncertainty about these parameters. In this way, in fact, the smoothing parameter values would be chosen by the data.

In the present context we preferred to regard the free choice of the smoothing parameters as an advantage of the model rather than a problem to be solved. By varying the smoothing parameters, in fact, features of the data that arise on different "scales" can be explored. A final value for λ and δ can then be obtained by a subjective choice. Figure 5 shows the centile curves obtained for increasing values of λ (by row) and increasing values of δ (by column), conditioning on $k = 4$. The curves become smoother as the parameters increase and δ, as expected, has a stronger effect than λ, since δ directly affects the variance of the γ_j. For $\lambda = \delta = 10$ the curves show very little curvature and a farther increase in one or both parameters would determine straight curves, parallel to the

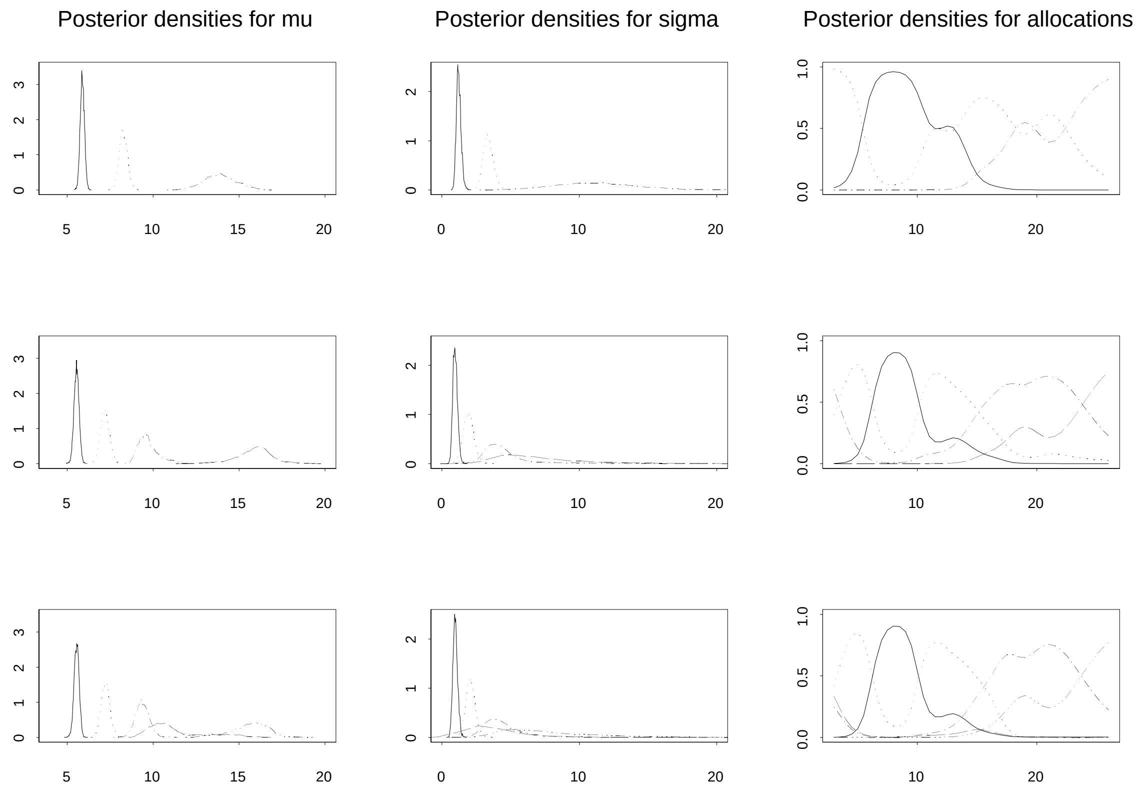

Figure 6: Marginal posterior densities for μ and σ^2 and z. Models with 3 to 5 components.

covariate axis.

Our final choice of the smoothing parameters, with λ larger than δ ($\lambda = 0.7$, $\delta = 0.09$) aimed to get smoothed curves without placing any strong restriction on the prior distribution of γ_j.

5.4 Posterior inference on μ, σ^2 and $w_j(t_i)$

Even if the focus of our analysis is on the inference about the growth curves, it is also interesting to spend a few words on the posterior distributions of the single components of the mixture and the posterior distributions of the allocation variables. Figure 6 shows the marginal posterior distribution of μ (left panel), σ^2 (central panel) and z (right panel) conditioning on increasing values of k, from $k = 3$ to $k = 5$. Here, a line of a given type is used to represent the posterior densities of μ and σ^2, for a given component, and the probability, conditional on the value of the covariate, of an observation to be allocated to that component. For all the models it can be noticed that the posterior distribution of μ_j becomes more and more spread as we move from μ_1 to μ_k. This is due to the fact that the mean value of σ_j^2 increases with j, and also to the fact that fewer observations are generally allocated to the last components. If we consider the modal allocation, for the model with five components, 526 observations are allocated to the first three components and only 58 observations are allocated to the last two.

It is also interesting to notice how the effect of going from $k = 3$ to $k = 4$ components is that all the components are shifted, while the principal effect of going from $k = 4$ to $k = 5$ components is that the third component is split into two.

Looking at the posterior distribution of σ_j^2, a general increase in the variance of the distribution is again evident, together with an increase in the shift towards right, as we move from σ_1^2 to σ_k^2. These increases are caused by a smaller number of observations allocated to components with large μ, but also by the fact that these observations are the most disperse ones. The only exception in the location of the posterior densities for σ^2 can be noticed for $k = 5$ where the posterior mode for σ_4^2 is less than the posterior mode for σ_3^2.

The posterior densities of z_i, i.e. the weight curves $w_j(t_i)$, can be analysed using both Figure 6 and Figure 7. Figure 7 describes the modal allocation of each observation and gives a more immediate idea than Figure 6.

In the model with $k = 3$, the observations at ages less than five are very likely associated to the second component. Then the weight of this second component decreases in favour of the weight of the first one to which all the observations in the notch are allocated with probability very close to 1. The observations at ages more than 15 are associated again with the second component or with the third one, when their value is high.

In the models with $k = 4$ and $k = 5$, the first component is still associated with the notch around the age 9. Observations on the edges of this notch are instead very likely to be allocated to the second component. The fourth component in the model with $k = 4$ has the same role of the fifth component in the model with $k = 5$: it is associated with observations aged more than 15 with high triceps values. As noted before, going from $k = 4$ to $k = 5$ has the effect of splitting the third component. For $k = 4$, the third component is both associated with observations at young ages and with observations at ages around 15 and older, characterised by triceps around the

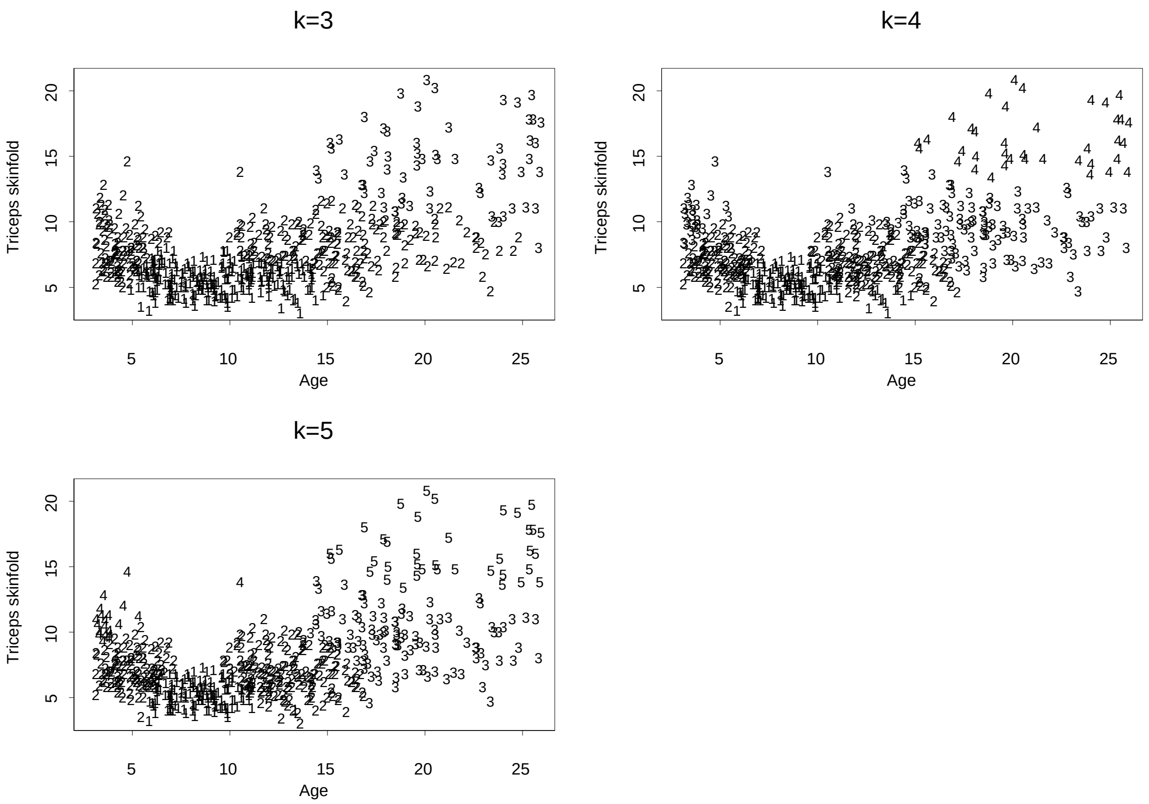

Figure 7: Modal allocation of the observations for models with 3 to 5 components

value 10. In the model with $k = 5$, observations at young ages and observations at ages around 15 and older are instead separated and allocated with high probability to the fourth and the third component respectively. It is evident that very few observations have a reasonable probability of being allocated to the fourth component. On the other side, these are much less dispersed than those likely to be allocated to the third one. This explains the fact that the posterior mode for σ_4^2 is less than the posterior mode for σ_3^2 when $k = 5$, and also the fact that σ_4^2 is more dispersed.

6 Discussion

We believe that the model introduced and discussed in this paper provide an interesting new methodology for the estimate of growth curves. The main difficulty lies, in fact, in deciding whether a bump or dip observed on a centile curve at a particular age is a real feature of the data, or whether it is simply sampling error. The Bayesian modelling approach, together with the flexible semiparametric nature of mixture models, the adaptability of the splines and the roughness penalising element they introduce provide an elegant solution to this problem. Data are smoothed preserving their heterogeneity, no age cut-offs need to be specified and the only arbitrariness in the whole procedure is the choice of the smoothing parameters λ and δ.

The model could be extended in several directions. One interesting extension is the use of B-splines to model not only the weights of the mixture, but also the component parameters μ_j and σ_j. This would give the model more flexibility. Let us think of an extreme example. If we consider a hypothetical data set for which the underlying theoretical model is a simple linear regression on the covariate, it is clear that the mixture model presented in the paper would require a large number of components to fit the data. Modelling the component parameters as a smoothed function of the covariate, would instead allow to fit the data using only one component.

Although the validation of the MCMC code did not receive much space in the paper, numerous checks were conducted on the correctness of our sampler. In particular we checked that without any data, our estimate of the joint posterior distribution tallies with the chosen prior. We are satisfied that the values chosen for the smoothing parameters λ and δ allow substantial smoothing in the centile curves, so that higher values of λ and δ, for which mixing could be slower, are not necessary.

Finally, we draw attention on the difficult matter of estimating the marginal likelihood for complex, non-regular problems, when the large amount of data induce very peaked posterior distributions for the parameters and makes the use of reversible jump infeasible.

Acknowledgements

We acknowledge partial support of this work by the UK Engineering and Physical Sciences Research Council, and are grateful for helpful comments by the referees and editor, which have led to improvement of the presentation.

References

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion), *Statistical Science*, **10**, 3–66.

Böhning, D. (2000) *Computer-assisted analysis of mixtures and applications.* Monograph on Statistics and Probability 81, Chapman and Hall/CRC, London.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–1319.

Fearn, T. (1975) A Bayesian approach to growth curves. *Biometrika*, **62**, 89–100.

Geisser, S. (1970) Bayesian analysis of growth curves. *Sankhyā*, **32**, 53–64.

Gelman, A. and Meng, X. L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163–185.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Richardson, S. (2000) Hidden Markov models and disease mapping.

Green, P. J. and Silverman B. W. (1994) *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

Hammersley, J. M. and Handscomb, D. C. (1964) *Monte Carlo Methods*, Chapman and Hall, London.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, series B*, **56**, 3–48.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, series B*, **59**, 731–792.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–1762.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, series B*, **40**, 364–372.