Bayesian Mixture of Parametric and Nonparametric Density Estimation: A Misspecification Problem*

Hedibert F. Lopes**
Ronaldo Dias***

Abstract

In this paper we study the effect of model misspecifications for probability density function estimation. We use a mixture of a parametric and nonparametric density estimation. The former can be modeled by any suitable parametric probability density function, including mixture of parametric models. The latter is given by the known B-spline estimation. The procedure also deals with the situation when a highly structured data are collected so that it is difficult to propose a parametric model with a large number of mixture components. Then a nonparametric part would help to postulate an appropriate model. In addition, in order to reduce the computational cost of getting a nonparametric density for high dimensional data a parametric mixture of densities could be used as the starting point for modeling such dataset. Our procedure is computed by using EM-type algorithm for a non-Bayesian approach and MCMC algorithm under a Bayesian point of view. Simulations and real data analysis show that our proposed procedure have performed quite well even for non structured datasets.

Keywords: Nonparametric Density Estimation, B-Splines, Mixtures Models, MCMC, EM-Algortihm.

JEL Codes: C11, C14, C15.

*Submitted in November 2011. Revised in July 2012. We would like to thank the referee for the comments and suggestions that made this work better and clearer. Part of this work was done while Prof. Ronaldo Dias was a Visiting Professor at University of Chicago, Business School. This research was partially supported by CNPq grant no. 302182/2010-1 and 475504/2008-9. Color plots upon request `dias@ime.unicamp.br`

**GSB, University of Chicago. E-mail: `hedibert.lopes@chicagogsb.edu`
***Universidade Estadual de Campinas. E-mail: dias@ime.unicamp.br

1. Introduction

Let $X_1, \ldots, X_n$ be i.i.d. random variables with an unknown continuous probability density function g on a compact interval $\mathcal{X}$. Based on the sample, one could be interested in estimating the density g. If it is possible to conjecture a parametric family for the distribution, life is much easier and parametric methods can be used. However, most of the time we cannot fit a parametric model and a nonparametric estimation is necessary, such as spline fitting Kooperberg and Stone (1991), Eilers and Marx (1996), Dias (1998) or Kernel fitting Parzen (1962).

The idea of combining a parametric with nonparametric model under the non-Bayesian point of view, was first introduced by Olkin and Spiegelman (1987). By taking a convex combination

$$g(x|\Psi) = \delta f(x|\Psi) + (1-\delta)h(x),$$

they expected that a parametric model does not hold when δ is close to zero, and consequently a nonparametric model should be more appropriate to explain the data.

Lenk (2003) proposes a Bayesian approach to a semiparametric density estimation by adding a Gaussian process to the kernel of the exponential family and expanding it through Karhunen-Loève expansion. Cai and Meyer (2011) developed, under a Bayesian point of view, a procedure based on a mixture of B-splines with fixed number of interior knots to estimate a baseline of proportional hazards model. In this work, we propose a completely different Bayesian semiparametric model. The idea is to represent the prior information as the parametric part and to capture a possible departure from this prior through the nonparametric component. In this way, when data have too much structure it is unlikely that an experimenter would believe in a large number of mixtures of parametric models to describe this data set. Then a nonparametric approach as proposed could show a lack of fit of the first suggested parametric model. In other words, a parametric model guided by a nonparametric model. On the other hand, a full nonparametric approach can be very costly in high dimensional data. Thus, having a parametric model as starting guess would help to lower the dimension. In that matter, a misspecification model could be treated very well by our proposed approach.

2. Methodology

2.1 Mixture of B-spline

The methodology developed for mixtures of parametric densities is well understood and readers can follow McLachlan and Peel (2000) for details of this subject. On the contrary, mixtures of B-splines are not so widespread and we think it deserves some comments.

Due to their simple structure and good approximation properties, polynomials are widely used in practice for approximating functions. For this propose, one

usually divides the interval $[a, b]$ in the function support into sufficiently small subintervals of the form $[\xi_0, \xi_1], \ldots, [\xi_q, \xi_{q+1}]$ and then uses a low degree polynomial p_i for approximation over each interval $[\xi_i, \xi_{i+1}]$, $i = 0, \ldots, q$. This procedure produces a piecewise polynomial approximating function $s(\cdot)$;

$$s(x) = p_i(x) \text{ on } [\xi_i, \xi_{i+1}],\ i = 0, \ldots, q.$$

In the general case, the polynomial pieces $p_i(x)$ are constructed independently of each other and therefore do not constitute a continuous function $s(x)$ on $[a, b]$. This is not desirable if the interest is on approximating a smooth function. Naturally, it is necessary to require the polynomial pieces $p_i(x)$ to join smoothly at knots $\xi_1, \ldots, \xi_q$, and to have all derivatives up to a certain order, coincide at knots. As a result, we get a smooth piecewise polynomial function, called a *spline function*.

Definition 2.1 *The function $s(x)$ is called a spline function (or simply "spline") of degree r with knots at $\{\xi_i\}_{i=1}^q$ if $-\infty =: \xi_0 < \xi_1 < \ldots < \xi_q < \xi_{q+1} := \infty$, where $-\infty =: \xi_0$ and $\xi_{q+1} := \infty$ are set by definition,*

- *for each $i = 0, \ldots, q$, $s(x)$ coincides on $[\xi_i, \xi_{i+1}]$ with a polynomial of degree not greater than r;*
- *$s(x), s'(x), \ldots, s^{r-1}(x)$ are continuous functions on $(-\infty, \infty)$.*

The set $\mathcal{S}_r(\xi_1, \ldots, \xi_q)$ of spline functions is called *spline space*. Moreover, the spline space is a linear space with dimension $q+r+1$ Schumaker (1972) and Schumaker (1993).

Definition 2.2 *For a given point $x \in (a, b)$ the function*

$$(t-x)_+^r = \begin{cases} (t-x)^r & \text{if } t > x \\ 0 & \text{if } t \leq x \end{cases}$$

is called the truncated power function of degree r with knot x.

Hence, we can express any spline function as a linear combination of $q + r + 1$ basis functions. For this, consider a set of interior knots $\{\xi_1, \ldots, \xi_q\}$ and the basis functions $\{1, t, t^2, \ldots, t^r, (t-\xi_1)_+^r, \ldots, (t-\xi_q)_+^r\}$. Thus, a spline function is given by,

$$s(t) = \sum_{k=0}^{r} \theta_k t^k + \sum_{k=r+1}^{q+r} \theta_k (t - \xi_{k-r})_+^r$$

It would be interesting if we could have basis functions that make it easy to compute the spline functions. It can be shown that B-splines form a basis of spline

spaces Schumaker (1972), Schumaker (1993) and Dierckx (1993). Also, B-splines have an important computational property, they are splines which have smallest possible support. In other words, B-splines are zero on a large set. Furthermore, a stable evaluation of B-splines with the aid of a recurrence relation is possible.

Definition 2.3 *Let $\Omega_\infty = \{\xi_k\}_{\{k\in\mathbb{Z}\}}$ be a nondecreasing sequence of knots. The k-th B-spline of order q for the knot sequence Ω_∞ is defined by*

$$B_k^q(t) = -(\xi_{k+q} - \xi_k)[\xi_k, \ldots, \xi_{k+q}](t - \xi_k)_+^{q-1} \quad \textit{for all} \quad t \in \mathbb{R},$$

where, $[\xi_k, \ldots, \xi_{k+q}](t-\xi_k)_+^{q-1}$ is $(q-1)$-th divided difference of the function $(x - \xi_k)_+^q$ evaluated at points $\xi_k, \ldots, \xi_{k+q}$.

From the Definition 2.3 we notice that $B_k^q(t) = 0$ for all $t \notin [\xi_k, \xi_{k+q}]$. It follows that only q B-splines have any particular interval $[\xi_k, \xi_{k+1}]$ in their support. That is, of all the B-splines of order q for the knot sequence Ω_∞, only the q B-splines $B_{k-q+1}^q, B_{k-q+2}^q, \ldots, B_k^q$ might be nonzero on the interval $[\xi_k, \xi_{k+1}]$. (See de Boor (1978) for details.) Moreover, $B_k^q(t) > 0$ for all $t \in (\xi_k, \xi_{k+q})$ and $\sum_{k\in\mathbb{Z}} B_k^q(t) = 1$, that is, the B-spline sequence B_k^q consists of nonnegative functions which sum up to 1 and provides a partition of unity. Thus, a spline function can be written as linear combination of B-splines,

$$s(t) = \sum_{k\in\mathbb{Z}} \beta_k B_k^q(t).$$

The value of the function $s(.)$ at point t is simply the value of the function $\sum_{k\in\mathbb{Z}} \beta_k B_k^q(t)$ which makes good sense since the latter sum has at most q nonzero terms. Particularly, there is a well know representation in terms of the natural splines, for the sequence of K interior knots $\{\xi_1 \ldots, \xi_K\}$ with degree $r = q - 1$

$$B_k(t) = \lambda_{0k} + \lambda_{1j}\, t + \sum_{l=1}^{K} \lambda_{(l+1)\,j}\, (t - \xi_l)_+^r$$

then

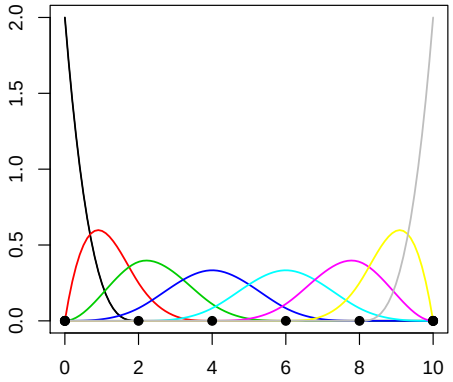
$$\beta_k\, B_k(t) = \beta_k\, \lambda_{0k} + \beta_k\, \lambda_{1j}\, t + \sum_{l=1}^{K} \beta_k \lambda_{(l+1)\,j}\, (t - \xi_l)_+^r$$

One may notice that, we have $K + 2$ coefficients to estimate which is also the dimension of the spline space. An attempt to reduce the dimension space to keep the computational cost low, one would necessarily have to lower the degree of the spline and the number of the interior knots. However, a change in the degree of the polynomial splines directly affects the function estimate smoothness. For instance, if one chooses $r = 3$, a spline function estimate would have 2 continuous derivatives. For $r = 1$ a spline function estimate would be just a constant function. Decreasing the number of interior knots reduces the function estimate variance but it increases

the bias. Nevertheless, under the computational viewpoint, the B-splines basis used, in general, are the cubic spline basis which have larger correlations among them than linear spline basis, hence the ill-conditioning problem is more serious. Hence, to choose the number of B-splines is a very important issue.

Figure 1 shows an example of B-splines basis and their compact support property. This property makes the computation of B-splines easier and numerically stable.

Figure 1
B-spline basis: $n_k = 6$ knots (4 interior knots) and eight basis



In some applications the normalized B-splines,

$$\frac{q}{\xi_{k+q} - \xi_k} B_k^q(t)$$

is used instead. Since, we have

$$\int_{\xi_k}^{\xi_{k+q}} B_k^q(t)\, dt \;=\; \frac{\xi_{k+q} - \xi_k}{q}.$$

It is well known that any probability density function can be well approximated by a finite mixture of B-splines basis functions. The procedures such as, Kooperberg and Stone (1991), Dias (1998) are automatic, flexible and very adaptive ones. Moreover, these procedures are shown to be computational efficiently in one dimensional case. However, it is still one of the most challenging problem how to select the dimension of the approximant space. In the context of mixture

of densities this problem is the same as to find the optimal number of the mixture components. In the case of basis functions density estimation several authors suggested algorithms in order to provide a good choice of the dimension of the approximant space as a function of the sample size, see for example Gu (1993), Antoniadis (1994), De Vore et al. (2003), Bodin et al. (2000) and Kohn et al. (2000). However, all of these procedures including adaptive ones Kooperberg and Stone (1991), Luo and Wahba (1997) and Dias (1998) deal with a non-random choice of K.

Assuming that there is a finite but unknown K such that f is an element of a space that can be generated by K mixture components of B-splines basis, Dias (2000), Dias and Garcia (2004) and Dias and Garcia (2007) suggested to use a proxy of the Kullback-Leibler distance in order to determine $\hat{K}$, an estimate of the true dimension. The Kullback-Leibler distance is attractive not only because of the linearization of the logistics transformation, but also it is asymptotically equivalent to maximize the likelihood function and to minimize the Hellinger distance. Note that K acts the smoothing parameter, large values of K cannot avoid the *Dirac's disaster* and small values of K provides a flat density estimate, (see Dias and Garcia (2007)). Consequently, an optimal, in some sense, density estimate strongly depends on a good estimate of the number of B-splines components. Moreover, cubic B-splines are related to Bernstein polynomials of order 3. However, B-splines are more suitable to describe data highly structured due to their computational stability, local properties and their degree of approximation does not depend on the order of polynomial as it is when Bernstein polynomials are being used. Figure 2 show an example of the relationship between Bernstein polynomials and B-splines.

Properties of B-splines, Bernstein polynomials and their use in density estimation can be found in de Boor (1978), Dias (1998) Dias (2000), Dias and Garcia (2004), Dias and Garcia (2007), Kooperberg and Stone (1991), Koo and Kooperberg (2000), O'Sullivan (1988) Petrone (1999) and Petrone and Wasserman (2002).

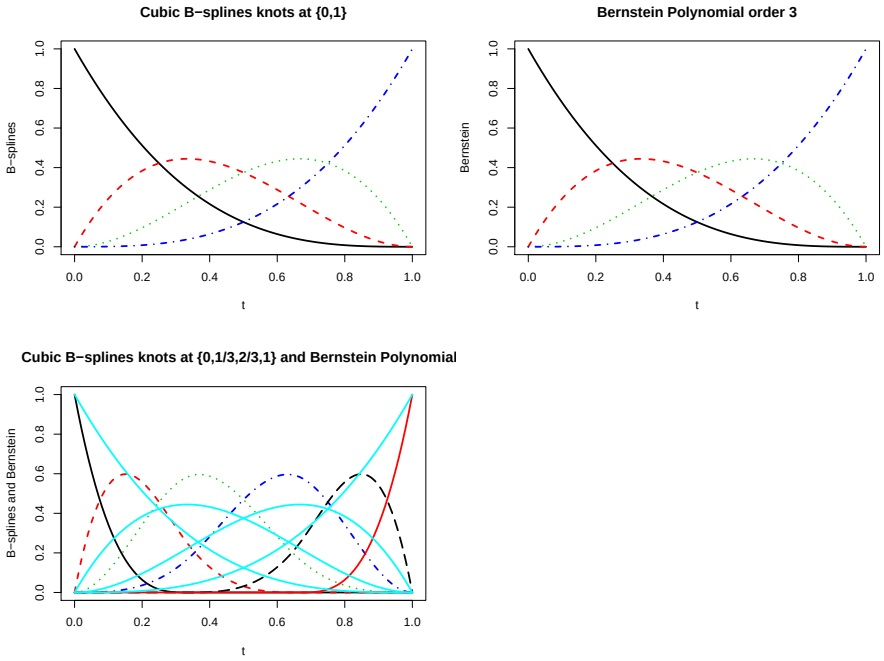
3. Inference

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed data from a density $g(x|\Psi)$, where

$$g(x|\Psi) = \delta \sum_{j=1}^{J} \alpha_j f_j(x|\theta_j) + (1-\delta) \sum_{k=1}^{K} \beta_k B_k^q(x) \tag{1}$$

where $\alpha_j \geqslant 0$ and $\sum_{j=1}^{J} \alpha_j = 1$, $\beta_k \geqslant 0$ and $\sum_{k=1}^{K} \beta_k = 1$, $0 \leqslant \delta \leqslant 1$, $f_j(x|\theta_j)$ are parametric models for $j = 1, \ldots, J$, and $B_k^q(y)$ are normalized B-splines of order q for $k = 1, \ldots, K$. Note that this spline approach naturally forms a mixture of a very flexible densities given by the linear combination of the normalized B-splines. Thus, 1 combines a mixture of parametric densities and a mixture of normalized B-splines. By doing so it provides a very adaptive and easy to understand model. Finally, our parameter vector is given by $\Psi = (\delta, \alpha_1, \ldots, \alpha_J, \theta_1, \ldots, \theta_J, \beta_1, \ldots, \beta_K)$.

Figure 2
B-spline basis and Bernstein polynomials



Let λ be the number of equally spaced knots for the B-splines, $x_{(1)} = \min\{x_1, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, \ldots, x_n\}$, so $u = (x_{(n)} - x_{(1)})/(\lambda + 1)$ is the width of the sub-intervals. Therefore, the knots for $q = 4$ are placed at $x_{(1)}, x_{(1)}, x_{(1)}, x_{(1)} + u, \ldots, x_{(n)} - u, x_{(n)}, x_{(n)}, x_{(n)}$. The interior knots can also be placed at the order statistics. This well known procedure in non-parametric density fit avoids trying to solve a difficult problem of optimizing the knot positions. Any other procedure has to take into account the fact that changes in the knot positions might cause considerable change in the function $h(.)$ and consequently in the function $g(.)$.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed data from a density $g(x|\Psi)$. Then, the likelihood function is given by

$$L(\Psi|\mathbf{x}) = \prod_{i=1}^{n} g(x_i|\Psi) = \prod_{i=1}^{n} \left[\delta \sum_{j=1}^{J} \alpha_j f_j(x_i|\theta_j) + (1-\delta) \sum_{k=1}^{K} \beta_k B_k^q(x_i) \right] \quad (2)$$

The Expectation-Maximization (EM) algorithm is a general technique for maximum likelihood (ML) or maximum a posteriori (MAP) estimation. The EM algorithm has become a very popular computational method in statistics and in econometrics. In the numerical optimization, the implementation of the E-step and M-step is easy for many real situations particularly because of the expression of the complete-data likelihood function. Moreover, in general, solutions of the M-step exist in closed form. In many cases the M-step can be performed with a standard statistical software thus saving a considerable amount of programming time. Another reason to prefer EM is that it does not require large storage space and so it is especially attractive to be used with small computers. In the appendices A and B we show how to compute the vector of parameters θ, α and β, by using a standard EM algorithm.

3.1 Posterior analysis via MCMC

3.1.1 Prior distribution

Let $\pi(\Psi)$ be the prior distribution of the vector Ψ. We will assume that

$$\pi(\Psi) = \pi(\delta)\pi(\alpha)\pi(\beta) \prod_{j=1}^{J} \pi(\theta_j) \quad (3)$$

where $\alpha = (\alpha_1, \ldots, \alpha_J)$ and $\beta = (\beta_1, \ldots, \beta_K)$. Moreover, $\alpha \sim \text{Dirichlet}(a_1, \ldots, a_J)$, $\beta \sim \text{Dirichlet}(b_1, \ldots, b_K)$ and $\delta \sim \text{Beta}(d_1, d_2)$. The prior distribution of θ_j will usually depend on the functional form $f_j(x|\theta_j)$ with natural conditionally conjugate priors a common choice. The prior for δ must take into account the plausibility of the parametric model.

3.1.2 Posterior distribution

For $\mathbf{x} = (x_1, \ldots, x_n)$,

$$\pi(\Psi|\mathbf{x}) \propto \prod_{i=1}^{n} \left[\delta \sum_{j=1}^{J} \alpha_j f_j(x_i|\theta_j) + (1-\delta) \sum_{k=1}^{K} \beta_k B_k^q(x_i) \right] \pi(\delta)\pi(\alpha)\pi(\beta) \prod_{j=1}^{J} \pi(\theta_j) \tag{4}$$

which is analytically intractable. However, by letting $(Z_1, W_1), \ldots, (Z_n, W_n)$ be latent random indicators such that $W_i \sim \text{Bernoulli}(\delta)$, $P(Z_i = j|W_i = 1) = \alpha_j$ for $j = 1, \ldots, J$ and $P(Z_i = k|W_i = 0) = \beta_k$ for $k = 1, \ldots, K$, we can rewrite $g(x_i|\Psi)$ by

$$\begin{aligned} g(x_i|\Psi) &= \sum_{j=1}^{J} \left[g(x_i|\Psi, Z_i = j) P(Z_i = j|W_i = 1) \right] P(W_i = 1) \\ &\quad + \sum_{k=1}^{K} \left[g(x_i|\Psi, Z_i = k) P(Z_i = k|W_i = 0) \right] P(W_i = 0) \\ &= \sum_{j=1}^{J} \left[f_j(x_i|\theta_j) P(Z_i = j|W_i = 1) \right] P(W_i = 1) \\ &\quad + \sum_{k=1}^{K} \left[B_k^q(x_i) P(Z_i = k|W_i = 0) \right] P(W_i = 0). \end{aligned}$$

Now, we can rewrite the posterior distribution by

$$\pi(\Psi, \mathbf{z}, \mathbf{w}|\mathbf{x}) \propto \prod_{i=1}^{n} \left[g(x_i|\Psi, z_i) p(z_i|w_i) p(w_i) \right] \pi(\delta)\pi(\alpha)\pi(\beta) \prod_{j=1}^{J} \pi(\theta_j) \tag{5}$$

where

$$g(x_i|\Psi, z_i) = \begin{cases} B_k^q(x_i) & \text{if } w_i = 0 \\ f_j(x_i|\theta_j) & \text{if } w_i = 1. \end{cases}$$

In the Appendix C we present an algorithm for sampling the posterior through Gibbs sampling methodology.

4. Model Assessment

We briefly summarize alternative measures of model selection that can be used in our mixture models using Bayesian approach, even though some posterior distributions may not be close to the normal distribution.

Let $L(\Psi_k|\mathbf{x}, \mathcal{M}_k)$ be the likelihood function given by (2) for the $\mathcal{M}_k$ competing model in the finite set $\{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$. Now, define

$$D(\Psi_k) = -2logL(\Psi_k|\mathbf{x}, \mathcal{M}_k),$$

as the deviance function. So, the deviance information criterion (DIC) is given by

$$DIC(\mathcal{M}_k) = 2E[D(\Psi_k)|\underline{x}, \mathcal{M}_k] - D(E[\Psi_k|\underline{x}, \mathcal{M}_k]). \quad (6)$$

Also, define d_k as the number of parameters of the $\mathcal{M}_k$ model and c_n a penalization constant. Alternatively, we can define a set of model selection criteria using the penalized deviance (PD) and its expected version, which are given by

$$PD(\mathcal{M}_k) = D(E[\Psi_k|\underline{x}, \mathcal{M}_k]) + c_n d_k \quad \text{and} \quad EPD(\mathcal{M}_k) = E[D(\Psi_k)|\underline{x}, \mathcal{M}_k] + c_n d_k.$$

For instance, we have for $c_n = 2$ the Akaike information criterion (AIC), for $c_n = \log(n)$ the Bayesian information criterion (BIC), and for $c_n = \log(\log(n))$ the Hannan and Quinn information criterion (HQIC).

Further details on all these measures can be found in Spiegelhalter et al. (2002). For computational purposes, suppose that $\{\Psi_k^{(1)}, \ldots, \Psi_k^{(L)}\}$ corresponds to a sample from $g(\Psi_k|\mathbf{x}, \mathcal{M}_k)$. Then, Monte Carlo approximations are given by

$$E[D(\Psi_k)|\underline{x}, \mathcal{M}_k] \simeq \frac{1}{L}\sum_{\ell=1}^{L} D(\Psi_k^{\ell}) \text{ and } E[\Psi_k|\underline{x}, \mathcal{M}_k] \simeq \frac{1}{L}\sum_{\ell=1}^{L} \Psi_k^{\ell},$$

which in turn can be used to approximate the DIC, AIC, BIC, HQIC, EAIC, EBIC and EHQIC. Indeed, MCMC reversible jump scheme could also be implemented. However, due to its very high computational cost we decide not to use it.

5. Applications

5.1 Simulated data exercise

In this section we illustrate the flexibility and robustness of our density estimation strategy by a series of examples. They emphasize particular characteristics of the mixture of parametric and nonparametric components, such as multimodality, heavy-tail and parsimony. In order to clarify notation throughout the exercise, we define n_k as the vector of knots, k_1 as the number of mixtures and k_2 as the number of basis used to fit the data.

5.1.1 Mixture of three normals

In this first exercise a three-component mixture of normal is entertained. Figure 3 shows the true model along with a sample of size $n = 1000$. The means, standard deviations and weights of the three normal components are $\mu = (-0.5, 2, 4)$, $\sigma = (0.5, 0.5, 0.5)$ and $\alpha = (0.6, 0.3, 0.1)$, respectively.

Figure 3
Simulation exercise I: Mixture of three univariate normal densities. $\mu = (-0.5, 2, 4)$, $\sigma = (0.5, 0.5, 0.5)$ and $\pi = (0.6, 0.3, 0.1)$. $n = 1000$

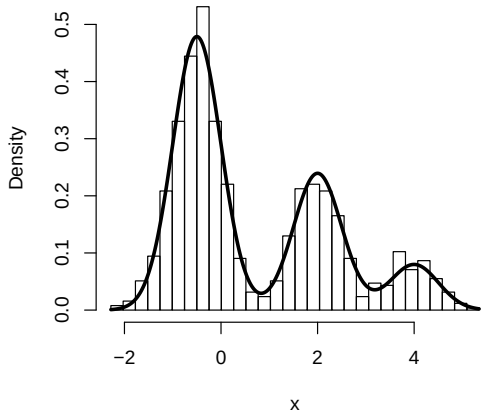


Table 1 presents Akaike's (AIC) and Schwarz (BIC) information criteria for several models for various values of k_1 (number of mixture components) and k_2 (number of basis). More specifically, $k_1 = (0, 1, 2, 3, 4, 5)$ and $k_2 = (0, 4, 8, 10, 12)$. Recalling, the number of basis, k_2, equals the number of knots, n_k, plus the order of the polynomial spline, which is fixed at four in all our simulations. These criteria are also shown in figures 4 and 5. Figure 6 shows both maximum likelihood and posterior inference for the density estimation problem.

In figures 4 and 5, one may notice that moving from 0 to 4 knots the AIC and BIC criteria drop significantly. This might indicate that the nonparametric component substantially contributes to improve a misspecified model with $k_1 = 1$.

5.1.2 Bivariate mixture of five normals

Although B-spline fitting (regression splines) has advantages such as flexibility and low computational cost, in multivariate case it also suffers of what it is known as *curse of dimensionality*. In other words, the computational cost increases at polynomial rate as the number of covariates increases. (See Scott (1992) for more details). Our procedure tries to overcome this problem by allowing the parametric component to act properly. In multivariate data, parametric models are easier to compute and in the misspecification case, a nonparametric component would help to find an appropriate model with with small number of knots in each direction.

In this exercise a five-component mixture of bivariate normal is entertained.

Table 1
AIC and BIC for the best-fit mixture models for the mixture of three normals

k1	nk	δ	AIC	BIC
0	0	0.0000000	0.000	0.000
0	4	0.0000000	3455.716	3494.978
0	6	0.0000000	3394.275	3443.353
0	8	0.0000000	3212.341	3271.234
1	0	1.0000000	3882.420	3916.775
1	4	0.5622872	3312.014	3365.999
1	6	0.5111636	3234.038	3297.838
1	8	0.4673966	3203.936	3277.552
2	0	1.0000000	3341.081	3390.159
2	4	0.7931247	3212.582	3281.290
2	6	0.7705431	3203.114	3281.638
2	8	0.6835683	3206.144	3294.484
3	0	1.0000000	3201.054	3264.855
3	4	0.9071151	3201.701	3285.132
3	6	0.8311305	3203.408	3296.655
3	8	0.7481505	3206.496	3309.559
4	0	1.0000000	3202.759	3281.283
4	4	0.9114228	3207.676	3305.831
4	6	0.8361831	3209.405	3317.376
4	8	0.7531043	3212.540	3330.326
5	0	1.0000000	3208.357	3301.605
5	4	0.9133282	3213.663	3326.541
5	6	0.8408506	3215.404	3338.098
5	8	0.7568929	3218.514	3351.023

Figure 4
Simulation exercise I: log AIC. $k_1 = (0, 1, 2, 3, 4, 5)$ and $n_k = (0, 4, 6, 8)$

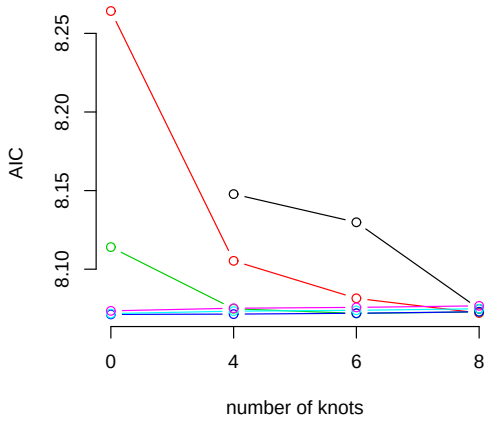


Figure 5
Simulation exercise I: log BIC

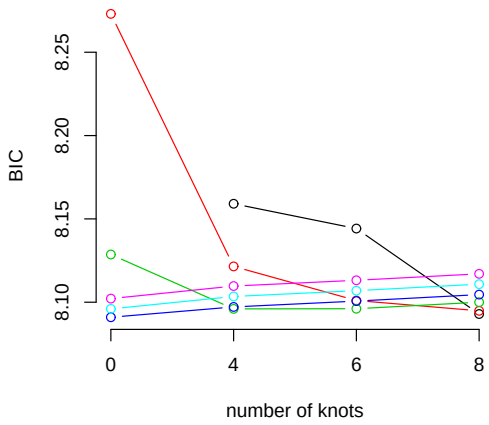


Figure 6
Simulation exercise I: Maximum likelihood and posterior inference for model with $k_1 = 3$ and $n_k = 0$. `mle:` $p(x|\hat{\theta}_{mle})$ via EM algorithm. `Bayes1:` $p(x|\tilde{\theta})$, where $\tilde{\theta} = E(\theta|data)$. `Bayes2:` $p(x) = \int p(x|\theta)p(\theta|data)d\theta$.

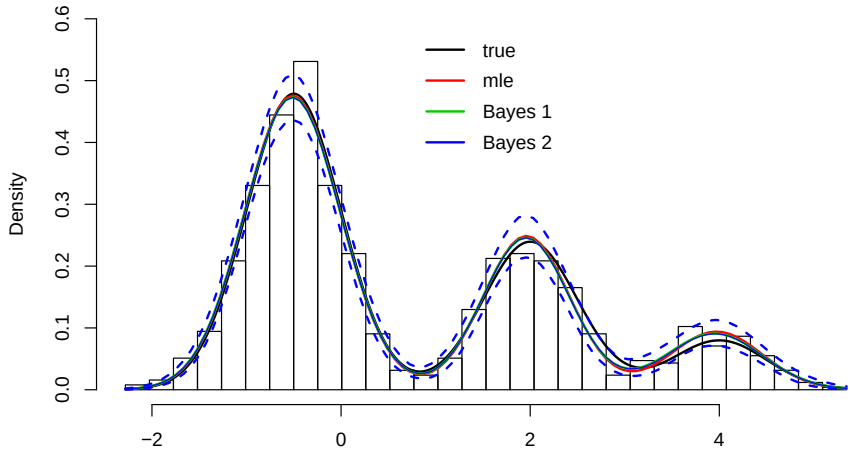


Figure 7 shows the true model along with a sample of size $n = 1000$.

The weights, means and variances of the three normal components are

$$\begin{aligned} \alpha &= (0.007, 0.106, 0.00002, 0.044, 0.34, 0.246, 0.258) \\ \mu &= (-11.40, -5.24, -9.84, 1.51, -0.65, 0.53, -2.36) \\ \sigma^2 &= (5.80, 2.61, 5.18, 0.17, 0.64, 0.34, 1.26) \end{aligned}$$

respectively.

Figure 7
Simulation exercise IV: Mixture of 5 bivariate normal densities. $n = 1000$ observations

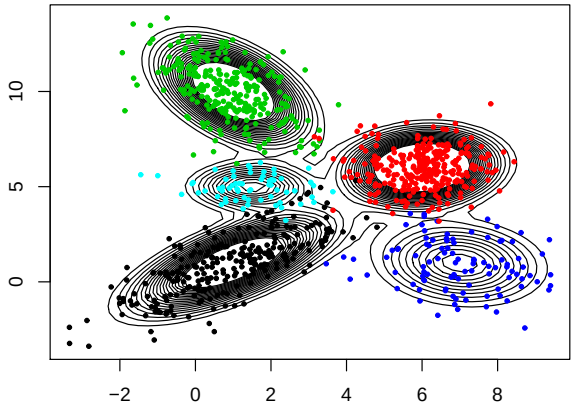


Table 2 presents AIC and BIC for several models for $k_1 = (2, \ldots, 7)$ and $n_k = (4, 6, 8)$. The BIC criterion is also shown in figures 8. Figure 9 presents the fit of the model with smallest BIC, i.e. $k_1 = 3$ and $n_k = 4$.

Note that the true model have 5-mixture components of a bivariate normal densities. The fitted models is able to recover 4 components precisely and show strong evidence about the location of the fifth one.

5.2 Real data exercise

5.2.1 Income data

The data in Figure 10 consists of 7121 random samples of yearly net income in the United Kingdom (Family Expenditure Survey, 1968-1983). The income data is considerably large and so it is more of a challenge to computing resources. The data has been standardized to induce unit variance.

Figure 8
Simulation exercise IV: BIC

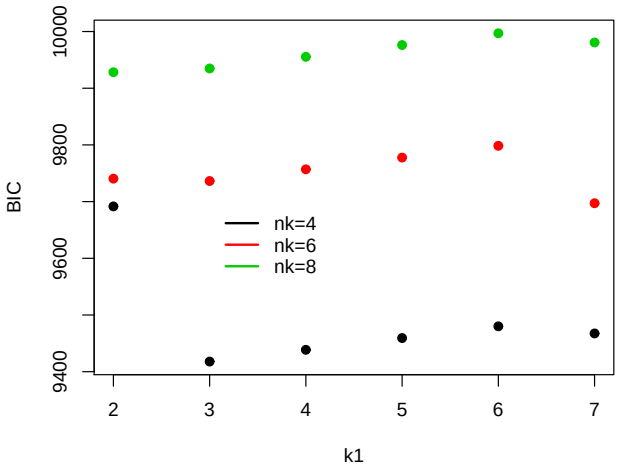


Figure 9
Simulation exercise IV: Model with $k_1 = 3$ and $n_k = 4$ knots

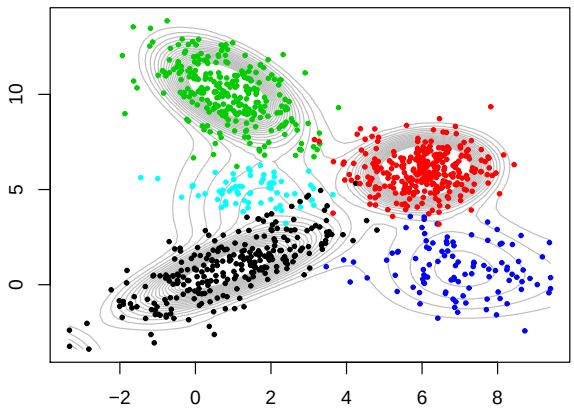


Table 2
AIC and BIC for the best-fit 5-normal mixture components

k1	nk	AIC	BIC
2	4	9348.027	9691.569
2	6	9220.287	9740.509
2	8	9192.024	9928.187
3	4	9059.593	9417.860
3	6	9201.361	9736.306
3	8	9183.936	9934.823
4	4	9065.597	9438.586
4	6	9207.292	9756.961
4	8	9189.775	9955.385
5	4	9071.599	9459.312
5	6	9213.292	9777.684
5	8	9195.829	9976.162
6	4	9077.597	9480.033
6	6	9219.309	9798.424
6	8	9201.845	9996.902
7	4	9050.327	9467.486
7	6	9103.326	9697.165
7	8	9170.895	9980.674

Figure 10
Income data: 7121 random samples of yearly net income in the United Kingdom
(Family Expenditure Survey, 1968-1983)

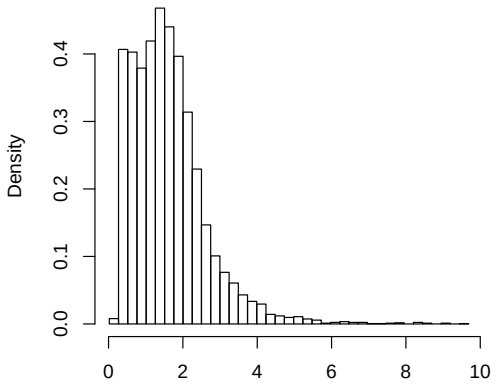


Table 3 presents AIC and BIC for several models for $k_1 = (0, 1, \ldots, 4)$ and $n_k = (0, 4, 6, 8)$. Figure 11 presents the BIC for these models, while the fits for models $k_1 = 4$ and $n_k = 0$ and $k_1 = 1$ and $n_k = 4$ appear in Figure 12.

Table 3
AIC and BIC for the best-fit for income data

k1	nk	δ	AIC	BIC
0	4	0.0000	17892.83	17947.79
0	6	0.0000	17777.38	17846.09
0	8	0.0000	17793.61	17876.06
1	0	1.0000	20221.52	20269.62
1	4	0.1544	17749.95	17825.53
1	6	0.2321	17761.41	17850.73
1	8	0.2478	17770.62	17873.68
2	0	1.0000	18522.55	18591.26
2	4	0.4809	17506.09	17602.28
2	6	0.3591	17776.41	17886.34
2	8	0.3983	17791.60	17915.28
3	0	1.0000	17747.77	17837.09
3	4	0.5800	17514.17	17630.98
3	6	0.4500	17783.71	17914.26
3	8	0.4836	17797.00	17941.29
4	0	1.0000	17557.97	17667.90
4	4	0.6142	17503.71	17641.13
4	6	0.4650	17788.40	17939.56
4	8	0.4969	17802.12	17967.02

Our procedure indicates that for this dataset the underline density has 2-mixture components.

5.2.2 Old Faithful Geyser data

> Old Faithful erupts every 35-120 minutes for 1.5-5 minutes to a height of 90-184 feet. The rangers say that 90% of their predictions are within +/- 10 minutes. The time to the next eruption is predicted using the duration of the current eruption. The longer the eruption lasts, the longer the interval until the next eruption. For instance, a 2 minute eruption results in an interval of about 50 minutes whereas a 4.5 minute eruption results in an interval of about 85 minutes. It is not possible to predict more than one eruption in advance. Old Faithful is deceiving. The benches around the geyser are over 300 feet from the geyser but with nothing to judge the distance by, I rarely realize just how big the geyser is until I get further away. I like the view from Geyser Hill. As with any geyser, watch the wind direction or you may only see steam. (From `http://www.geocities.com/Yosemite/1407/geysers.html`)

Figure 11
Income data: BIC for several models. $k_1 = (0, 1, \ldots, 4)$ and $n_k = (0, 4, 6, 8)$

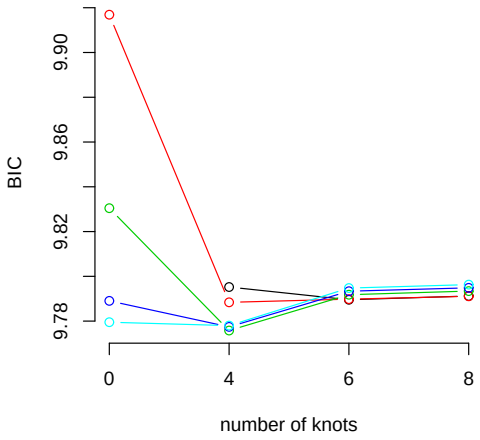


Figure 12
Income data: `Mixture` $k_1 = 4$ and $n_k = 0$, `Mixture+Bspline` $k_1 = 1$ and $n_k = 4$

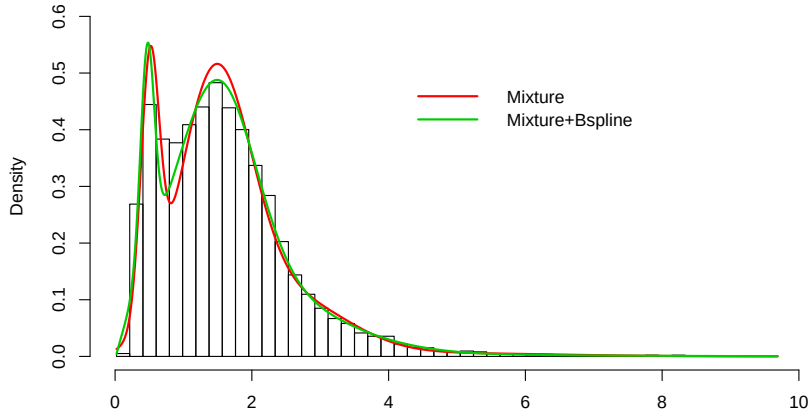


Figure 13
Old faithful data: Data along with contours of mixture model with $k_1 = 2$ and $n_k = 4$

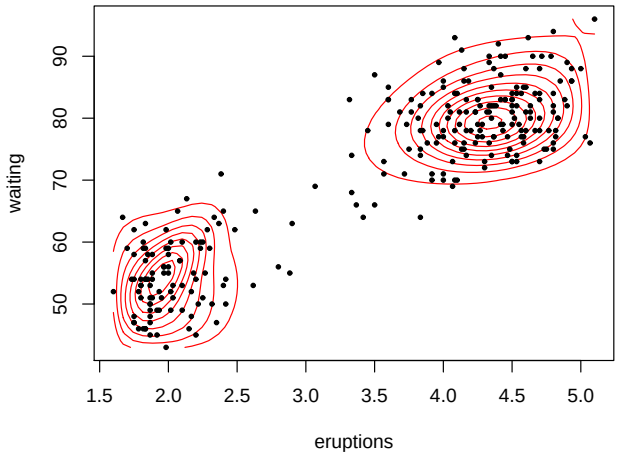


Table 4
AIC and BIC for the best-fit for Old Faithful data

k1	nk	AIC	BIC
2	4	2348.862	2601.268
2	6	2402.981	2785.196
2	8	2478.296	3019.166
3	4	2348.669	2611.893
3	6	2402.087	2795.119
3	8	2475.910	3027.598
4	4	2336.907	2610.948
4	6	2401.799	2805.649
4	8	2474.973	3037.478
5	4	2335.737	2620.596
5	6	2399.663	2814.330
5	8	2478.017	3051.339
6	4	2338.513	2634.189
6	6	2399.510	2824.994
6	8	2482.367	3066.507

6. Concluding Remarks

The mixture model presented is able to determine when a parametric model is being misspecified through the δ parameter. Moreover, prior information can always be incorporated in the model. For instance, if we believe that the parametric model is correct, we can pass this information to the model by setting adequately the hyperparameter of δ.

The presented examples illustrate the idea of model misspecification and how the nonparametric component helps to identify and measure, in some sense, the effect of model misspecification. In addition, real and simulated exercises show that the proposed method works very well in handling this situation.

References

Antoniadis, A. (1994). Wavelet methods for smoothing noisy data. In *Wavelets, images, and surface fitting (Chamonix-Mont-Blanc, 1993)*, pages 21–28. A K Peters, Wellesley, MA.

Bodin, P., Villemoes, L. F., & Wahlberg, B. (2000). Selection of best orthonormal rational basis. *SIAM J. Control Optim.*, 38(4):995–1032 (electronic).

Cai, B. & Meyer, R. (2011). Bayesian semiparametric modeling of survival data based on mixtures of b-spline distributions. *Computational Statistics and Data Anlaysis*, 55:1260–1272.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer Verlag, New York.

De Vore, R., Petrova, G., & Temlyakov, V. (2003). Best basis selection for approximation in L_p. *Found. Comput. Math.*, 3(2):161–185.

Dias, R. (1998). Density estimation via hybrid splines. *Journal of Statistical Computation and Simulation*, 60:277–294.

Dias, R. (2000). A note on density estimation using a proxy of the Kullback-Leibler distance. *Brazilian Journal of Probability and Statistics*, 13(2):181–192.

Dias, R. & Garcia, N. L. (2004). A spline approach to nonparametric test of hypotheses. *Brazilian Journal of Probability and Statistics.*, 18(1).

Dias, R. & Garcia, N. L. (2007). Consitent estimator for basis selection based on a proxy of the kullbakc-leibler distance. *Journal of Econometrics*, 141(1).

Dierckx, P. (1993). *Curve and surface fitting with splines*. Monographs on Numerical Analysis. The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.*, 11(2):89–121. With comments and a rejoinder by the authors.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *J. of the Amer. Stat'l. Assn.*, 88:495–504.

Kohn, R., Marron, J. S., & Yau, P. (2000). Wavelet estimation using Bayesian basis selection and basis averaging. *Statist. Sinica*, 10(1):109–128.

Koo, J.-Y. & Kooperberg, C. (2000). Logspline density estimation for binned data. *Statist. Probab. Lett.*, 46(2):133–147.

Kooperberg, C. & Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics and Data Analyis*, 12:327–347.

Lenk, P. J. (2003). Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *J. Comput. Graph. Statist.*, 12(3):548–565.

Luo, Z. & Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association*, 92:107–116.

McLachlan, G. & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Olkin, I. & Spiegelman, C. H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.*, 82(399):858–865.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. on Scientific and Stat'l. Computing*, 9:363–379.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. of Mathematical Stat.*, 33:1065–1076.

Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.*, 27(1):105–126.

Petrone, S. & Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(1):79–100.

Schumaker, L. L. (1972). *Spline Functions and Aproximation theory*. Birkhauser.

Schumaker, L. L. (1993). *Spline functions: basic theory*. Robert E. Krieger Publishing Co. Inc., Malabar, FL. Correlated reprint of the 1981 original.

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice, and Visualization*. John Wiley and Sons (New York, Chichester).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–639.

Estimation via EM

A. EM Algorithm for Mixtures

At iteration m,

$$\tau_{il}^{(m)} = \frac{\alpha_l^{(m)} p(x_i|\theta_l^{(m)})}{\sum_{j=1}^{\kappa_1} \alpha_j^{(m)} p(x_i|\theta_j^{(m)})}$$

Then,

$$\begin{aligned}
\alpha_l^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n} \tau_{il}^{(m)} \\
\mu_l^{(m+1)} &= \frac{\sum_{i=1}^{n} \tau_{il}^{(m)} x_i}{\sum_{i=1}^{n} \tau_{il}^{(m)}} \\
\sigma_l^{2(m+1)} &= \frac{\sum_{i=1}^{n} \tau_{il}^{(m)} (x_i - \mu_l^{(m+1)})^2}{\sum_{i=1}^{n} \tau_{il}^{(m)}}
\end{aligned}$$

B. EM Algorithm for B-Splines

$$\tau_{il}^{(m)} = \frac{\alpha_l^{(m)} B_l(x_i)}{\sum_{j=1}^{\kappa_2} \alpha_j^{(m)} B_j(x_i)}$$

Then,

$$\alpha_l^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{il}^{(m)}$$

B.1 EM algorithm for combination of mixtures and B-splines

At iteration m,

$$\begin{aligned}
\tau_{il}^{(m)} &= \frac{\alpha_l^{(m)} p(x_i|\theta_l^{(m)})}{\sum_{j=1}^{\kappa_1} \alpha_j^{(m)} p(x_i|\theta_j^{(m)})}, \quad l = 1, \ldots, k_1 \\
\tau_{il}^{(m)} &= \frac{\alpha_l^{(m)} B_{l-k_1}(x_i)}{\sum_{j=1}^{\kappa_2} \alpha_j^{(m)} B_j(x_i)}, \quad l = k_1+1, \ldots, k_1+k_2.
\end{aligned}$$

Then,

$$\begin{aligned}
\alpha_l^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n} \tau_{il}^{(m)}, \quad l = 1, \ldots, k_1+k_2 \\
\mu_l^{(m+1)} &= \frac{\sum_{i=1}^{n} \tau_{il}^{(m)} x_i}{\sum_{i=1}^{n} \tau_{il}^{(m)}}, \qquad l = 1, \ldots, k_1 \\
\sigma_l^{2(m+1)} &= \frac{\sum_{i=1}^{n} \tau_{il}^{(m)} (x_i - \mu_l^{(m+1)})^2}{\sum_{i=1}^{n} \tau_{il}^{(m)}} \quad l = 1, \ldots, k_1
\end{aligned}$$

C. Full Conditionals for the Gibbs Sampling

- Sampling from $(\mathbf{z}, \mathbf{w}|\Psi, \mathbf{x})$:

$$\pi(\mathbf{z}, \mathbf{w}|\Psi, \mathbf{x}) \propto \prod_{i=1}^{n} g(x_i|z_i)p(z_i|w_i)p(w_i).$$

  We can sample from

$$P(Z_i = \ell, W_i = \omega|\Psi, \mathbf{x}) \propto \begin{cases} (1-\delta)\beta_\ell B_\ell(x_i) & \text{if } \omega = 0 \text{ and } \ell = 1, \ldots, K \\ \delta\alpha_\ell f_\ell(x_i|\theta_\ell) & \text{if } \omega = 1 \text{ and } \ell = 1, \ldots, J. \end{cases}$$

- Sampling from $(\delta|\mathbf{w}) \sim \text{Beta}(d_1 + \sum_{i=1}^{n} w_i, d_2 + n - \sum_{i=1}^{n} w_i)$.

- Sampling from $(\alpha|\mathbf{z}, \mathbf{w}) \sim \text{Dirichlet}(a_1+n_1, \ldots, a_J+n_J)$ where $n_j = \sum_{i=1}^{n} I(z_i = j, w_i = 1)$ for $j = 1, \ldots, J$.

- Sampling from $(\beta|\mathbf{z}, \mathbf{w}) \sim \text{Dirichlet}(b_1 + m_1, \ldots, b_K + m_K)$ where $m_k = \sum_{i=1}^{n} I(z_i = k, w_i = 0)$ for $k = 1, \ldots, K$.

- Sampling θ_j when $f_j(x_i|\theta_j = (\mu_j, \sigma_j^2)) = \phi((x_i - \mu_j)/\sigma_j)$ where $\phi(.)$ is the probability density function of a standard normal distribution. Let the prior for μ_j and σ_j^2 be conditionally conjugate,

$$\mu_j \sim \text{N}(\mu_{0j}, \tau_{0j}^2) \text{ and } \sigma_j^2 \sim \text{IG}(\nu_{0j}/2, nu_{0j}s_{0j}^2/2),$$

  for $j = 1 \ldots, J$. Then,

  $(\mu_j|.) \sim \text{N}(\mu_{1j}, \tau_{1j}^2)$, where

$$\tau_{1j}^{-2} = \tau_{0j}^{-2} + n_j\sigma_j^{-2} \text{ and } \mu_{1j} = \tau_{1j}^2 \left[\mu_{0j}\tau_{0j}^{-2} + \sigma_j^{-2} \sum_{i=1}^{n} x_i I(z_i = j, w_i = 1)\right]$$

  $(\sigma_j^2|.) \sim \text{IG}(\nu_{1j}/2, \nu_{1j}s_{1j}^2)$, where

$$\nu_{1j} = \nu_{0j} + n_j \text{ and } \nu_{1j}s_{1j}^2 = \nu_{0j}s_{0j}^2 + \sum_{i=1}^{n} (x_i - \mu_j)^2 I(z_i = j, w_i = 1)$$

- Sampling θ_j when $f_j(x_i|\theta_j = (\mu_j, \sigma_j^2)) = \psi(x_i|\mu_j, \sigma_j)$ where $\psi(.)$ is the probability density function of a gamma distribution with mean μ_j/σ_j. We assume a Jeffreys' prior for μ_j and σ_j^2, which is given by

$$\pi(\mu_j, \sigma_j) \propto \sigma_j^{-1}(\mu_j\varphi^{(1)}(\mu_j) - 1)^{1/2}$$

where $\varphi^{(1)}(.)$ being the trigamma function and $j = 1 \ldots, J$. Then,

$$\log(\pi(\mu_j|.)) = c + n_j\mu_j \log(\sigma_j) - n_j \log(\Gamma(\mu_j)) + \frac{1}{2}\log(\mu_j\varphi^{(1)}(\mu_j) - 1)$$
$$+ \mu_j \sum_{i=1}^{n} \log(x_i) I(z_i = j, w_i = 1)$$
$$(\sigma_j|.) \sim \mathrm{G}(n_j\mu_j, \tau_j), \text{ where } \tau_j = \sum_{i=1}^{n} x_i I(z_i = j, w_i = 1)$$