

Bayesian Sensitivity Analysis for Measurement Error Models

Fernando A. Quintana, Pilar L. Iglesias * Heleno Bolfarine †

September 2, 2005

Abstract

The problem of outlier and change-point identification has received considerable attention in traditional linear regression models from both, classical and Bayesian standpoints. In contrast, for the case of regression models with measurement errors, also known as error-in-variables models, the corresponding literature is scarce and largely focused on classical solutions for the normal case. The main object of this paper is to propose clustering algorithms for outlier detection and change-point identification in scale mixture of error-in-variables models. We propose an approach based on product partition models (PPMs) which allows to study clustering for the models under consideration. This includes the change-point problem and outlier detection as special cases. The outlier identification problem is approached by adapting the algorithms developed by Quintana and Iglesias (2003) for simple linear regression models. An special algorithm is developed for the change-point problem which can be applied in a more general setup.

*Departamento de Estadística, Facultad de Matemáticas, PUC, CHILE. E-mail: {quintana,pliz}@mat.puc.cl. Partially supported by grants FONDECYT 1020712 and 1030588.

†Departamento de Estatística, IME, Universidade de Sao Paulo, BRASIL. E-mail: hbolfar@ime.usp.br. Partially supported by CNPq-Brasil.

The methods are illustrated with two applications: (i) outlier identification in a problem involving the relationship between two methods for measuring serum kanamycin in blood samples from babies, and (ii) change-point identification in the relationship between the monthly dollar volume of sales on the Boston Stock Exchange and the combined monthly dollar volumes for the New York and American Stock Exchanges.

Keywords: Dirichlet processes; Measurement error model; Outlier and change-point identification; Product partition models.

1 Introduction

Regression diagnostics such as outlier or change-point identification have been extensively studied in the statistical literature. Reviews of the different approaches include Broemeling (1985), Chapter 7 and more recently Barnett and Lewis (1994), Carlstein, Müller and Seigmund (1994), Brown (1993), Stephens (2000), Hawkins and Olive (2002) and Arellano-Valle, Galea-Rojas and Iglesias (2000). However, very few papers are devoted to studying this topic in measurement error models. Concerning change-point problems (for which the underlying cluster structure is given in terms of contiguous blocks), one exception is the paper by Chang and Huang (1997). They present an approach based on large sample theory for likelihood ratio tests. For small and moderate samples, however, the efficacy of such approach is yet to be studied. Similarly, the problem of outlier detection has been treated only from a classical viewpoint by Kelly (1984), Abdullah (1995), Lee and Zhao (1996), Galea, Bolfarine and Vilca (2002) and Galea, Bolfarine and Castro (2002). A common idea in these papers is to use measures of local influence to examine the effect of deleting a given subset of observations. In this paper we introduce a general approach based on the product partition models (PPMs), which allows studying the identification of relevant subsets in the data, such as outliers or change-points.

Measurement error models are important in most research areas, because they involve variables that are measured with error or that can not be directly observed (latent variables). Different examples and applications of these models are considered in the books by Fuller (1987), Carroll, Rupert and Stefanski (1995), and Cheng and Van Ness (1999). Bayesian literature includes the pioneering work by Lindley and El-Sayyad (1968) and some more recent results are found in Stephens and Dellaportas (1992), Dellaportas and Stephens (1995), Bolfarine and Cordani (1993), Richardson (1993), and in Richardson and Gilks (1993), among others. In the literature, the most common measurement error model (MEM) is the additive structural model specified by the equations

$$y_i = \alpha + \beta x_i + e_i, \quad (1.1)$$

$$z_i = x_i + u_i, \quad (1.2)$$

$i = 1, \dots, n$, where (y_i, z_i) are observed variables, (x_i, e_i, u_i) , $i = 1, \dots, n$ are unobserved random quantities, and (α, β) are unknown parameters. Another specification considers the x_i as fixed quantities (parameters) leading to the so-called functional version. From a Bayesian viewpoint however, no distinction is made since a prior distribution is assigned to the parameters x_i , $i = 1, \dots, n$. It is typically postulated that

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix} \right], \quad i = 1, \dots, n, \quad (1.3)$$

and with the further assumption that $x_i \sim N(\mu_x, \sigma_x^2)$, we have an structural additive normal model. This model has been extensively studied (see, e.g. Fuller, 1987). Recent statistical literature, however, has incorporated models based on distributions with heavier tails, which can accommodate outliers or spurious observations. For instance, one can explore the possibility of replacing the normal distribution above by a scale-mixture of normals, which includes the Student- t distribution as a special case (Arellano-Valle and Bolfarine, 1996).

To identify change-points in the mean of a model defined by (1.1) and (1.2), Chang and Huang (1997) proposed introducing change-points in the mean of the *unobserved* x

as follows:

$$x_i \sim N(\mu_1, \sigma_x^2), \quad i = 1, \dots, k, \quad (1.4)$$

$$x_i \sim N(\mu_2, \sigma_x^2), \quad i = k + 1, \dots, n, \quad (1.5)$$

where k is unknown. In other words, it is assumed that the unobserved x_i s are partitioned into two contiguous blocks. In fact, combining (1.1) – (1.5), the following change-point model is induced on the observed $\mathbf{w}_i = (z_i, y_i)^T$:

$$\mathbf{w}_i \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \alpha + \beta\mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 + \sigma_u^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \beta^2\sigma_x^2 + \sigma_e^2 \end{pmatrix} \right], \quad i = 1, \dots, k,$$

$$\mathbf{w}_i \sim N_2 \left[\begin{pmatrix} \mu_2 \\ \alpha + \beta\mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 + \sigma_u^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \beta^2\sigma_x^2 + \sigma_e^2 \end{pmatrix} \right], \quad i = k + 1, \dots, n.$$

Note that the marginal distribution of the observed \mathbf{w}_i s presents a change-point on the mean vector as well. Chang and Huang (1997) dealt with the presence of multiple change-point by splitting the original sequence at the change-point into two subsequences and reapplying the log-likelihood ratio statistic to each part. This continues until no further change-point is detected. Hence, there is no simultaneous (global) classical approach for the detection of multiple change-points. We also point out that it seems unlikely that the method by Chang and Huang (1997) can be extended for testing the presence of change-points in the other parameters. Furthermore, it does not seem feasible to extend their approach to models such as mixtures of normals.

Considering now the problem of outlier detection, a popular way of proceeding consists of deleting cases and assessing how this impacts the estimation process. We consider here an opposite approach, in the sense of accounting for the potential presence of outliers without removing observations at all, which simultaneously yields outlier identification *and* model robustification. Simply, we just consider partitions of the subjects into one group representing the main body of “normal” data points and one or more group of “atypical” observations, i.e. the outliers. More generally, either change-point or outlier detection imply a certain partition of the data points. By explicitly incorporating

partitions in the set of data points we propose a general method that, from a modeling viewpoint, makes essentially no distinction between these two problems. However, the selection of the adequate partition is done by means of a clustering algorithm derived from a decision-theoretic view of the problem constructed according to the objective of the analysis. An additional advantage of this procedure is the robustness of the related parameter estimation.

In this paper we consider a Bayesian approach for detecting the presence of cluster structures on the mean of x with applications to outlier detection and change-point identification. The approach is based on the product partition model (PPM) introduced by Hartigan (1990) and Barry and Hartigan (1992). See also Crowley (1997), and Loschi and Cruz (2005). As such, it is a global approach in the sense that the main target is to identify those groups that are defined by a common, possibly multidimensional parameter. We apply the approach developed in Quintana and Iglesias (2003), who established an important connection between PPMs and Bayesian nonparametric models based on Dirichlet process priors (Ferguson, 1973, Antoniak, 1974), using a decision theoretic approach. This connection facilitates the practical and operational use of PPMs in complicated models such as those involving variables measured with error.

The rest of this paper is organized as follows. Section 2 presents a model for additive linear mixture of normals with error-in-variables which incorporates a cluster structure. Our general formulation includes normal and Student-t models as special cases. We emphasize here that, to our knowledge, no approach has ever been proposed for cluster (outlier and change-point) determination in Student-t measurement error models. Section 3 presents the algorithms for cluster selection. We propose two algorithms, one for each specific type of problem considered. This difference is justified by the intrinsic nature of both problems. Section 4 presents applications to two data sets. One of the data sets was previously analyzed for outlier detection using the classical influence function approach. More general results on the partition structure for this data set stems from the clustering approach considered. On the other hand we apply our change-point detec-

tion method to the Boston Stock exchange dataset which has been previously analyzed in, e.g. Chang and Huang (1997). Some final conclusions are presented in Section 5.

2 The cluster structure in scale-mixtures of normals error-in-variables models

The scale-mixture of normals error-in-variables model is specified by considering latent variables $\mathbf{w}_j = (w_{j1}, \dots, w_{jn})^T$, $j = 1, 2, 3$, and assuming that

$$(y_i, z_i)^T | x_i, \mu_i, \alpha, \beta, \sigma_x^2, \sigma^2 \stackrel{ind}{\sim} N_2 \left[\begin{pmatrix} \alpha + \beta x_i \\ x_i \end{pmatrix}, \begin{pmatrix} w_{1i} \lambda \sigma^2 & 0 \\ 0 & w_{2i} \sigma^2 \end{pmatrix} \right], \quad (2.6)$$

and

$$x_i | \mu_x, \alpha, \beta, \sigma_x^2, \sigma^2 \stackrel{ind}{\sim} N(\mu_x, w_{3i} \sigma_x^2) \quad (2.7)$$

where the components w_{ji} are assumed independent of $(\mu_x, \alpha, \beta, \sigma^2, \sigma_x^2)^T$, for $j = 1, 2, 3$ and $i = 1, \dots, n$, and $\lambda > 0$ is a known constant. The above model includes several interesting special cases. If $w_{ji} = 1$, $j = 1, 2, 3$ and $i = 1, \dots, n$, then the normal error-in-variables follows. On the other hand, if we define

$$w_{ji} \stackrel{ind}{\sim} IG\left(\frac{\nu_j}{2}, \frac{\nu_j}{2}\right), \quad j = 1, 2, 3 \quad \text{and} \quad i = 1, \dots, n,$$

we arrive at the Student- t error-in-variables model, where $IG(a, b)$ denotes the inverse gamma distribution with positive parameters a and b .

To identify outliers or change-points we propose to consider groups or clusters of observations based on the marginal mean of the x_i s. Thus, the model is reformulated by replacing (2.7) with

$$x_i | \mu_i, \alpha, \beta, \sigma^2, \sigma_x^2 \stackrel{ind}{\sim} N(\mu_i, w_{3i} \sigma_x^2), i = 1, \dots, n. \quad (2.8)$$

The cluster structure is introduced as follows. First note that $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$

can be re-expressed as

$$\boldsymbol{\mu} = \sum_{i=1}^n (\mu^{S_i} I_{\{1 \in S_i\}}, \dots, \mu^{S_i} I_{\{n \in S_i\}}),$$

where $\{S_1, S_2, \dots, S_k\}$ is a partition of $S_0 = \{1, 2, \dots, n\}$ and $\mu^{S_i} \in \mathbb{R}$ denote the common value for $\mu_i, i \in S, S \subseteq S_0$. The prior for $\boldsymbol{\mu}$ is specified through the following two steps:

- (i) Assign a prior distribution on \mathcal{P} , the set of all partitions of S_0 . The random object with values on \mathcal{P} , is usually denoted by ρ .
- (ii) Assign a prior distribution for $\boldsymbol{\mu}$ given $\rho = \{S_1, \dots, S_k\}$, which is equivalent to specifying a prior for $\mu^{S_1}, \mu^{S_2}, \dots, \mu^{S_k}$.

The prior specified in step (i) reflects our belief about the groups defined by a common mean. We note that $\boldsymbol{\mu}$ can be also represented as $(\mu^{S_1}, \dots, \mu^{S_{|\rho|}}, \rho)$ where $|\rho|$ denotes the number of groups determined by ρ . Thus, we can rewrite (2.8) in the form

$$x_i | (\mu^{S_1}, \dots, \mu^{S_{|\rho|}}, \rho), \alpha, \beta, \sigma^2, \sigma_x^2 \stackrel{ind}{\sim} N(\mu_i, \omega_{3i} \sigma_x^2), \quad i = 1, \dots, n, \quad (2.9)$$

where $\mu_i = \mu^{S_j}$ if and only if $i \in S_j$.

The prior that we put for $\boldsymbol{\mu}$ given the parameters is of the form

$$\mu^{S_1}, \dots, \mu^{S_{|\rho|}} | \rho, \alpha, \beta, \sigma^2, \sigma_x^2 \stackrel{iid}{\sim} N(m, \tau_0 \sigma_x^2), \quad (2.10)$$

and

$$P(\rho = \{S_1, \dots, S_k\} | \alpha, \beta, \sigma^2, \sigma_x^2) = \mathcal{K} \prod_{i=1}^k c(S_i) \quad (2.11)$$

where $c(S) \geq 0$ is denominated *cohesion*, defined for each $S \in \mathcal{P}$ and not depending on $(\alpha, \beta, \sigma^2, \sigma_x^2)$, and \mathcal{K} is an appropriate normalizing constant. Here, the cohesions represent prior weights on group formation and $c(S)$ can be thought of as formalizing our opinion on how tightly clustered the elements of S would be. By (2.11), the prior

probability of observing a partition including S as one of its subsets increases with the value of $c(S)$.

Equation (2.11) is referred to as the *product distribution* for partitions, and the model specified by (2.9), (2.10) and (2.11) is usually called *parametric partition product model*, a class introduced by Hartigan (1990) (see also Barry and Hartigan, 1992 and Crowley, 1997).

Remark 2.1: Some special cases can be obtained for particular elections of the cohesion. For example, if we choose

$$c(S) = \begin{cases} 1 & \text{if } S = S_0 \\ 0 & \text{if } S \neq S_0 \end{cases}$$

then we obtain a MEM defined by (2.6) and (2.7), without change-points. The MEM considered by Chang and Huang (1997) is obtained when choosing cohesions of the form

$$c(S) = \begin{cases} c_k & \text{if } S = \{S_{1k}, S_{2k}\}, k = 1, \dots, n-1 \\ 0 & \text{otherwise,} \end{cases}$$

where $S_{1k} = \{1, 2, \dots, k\}$ and $S_{2k} = \{k+1, \dots, n\}$.

We note that given $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2, \sigma_x^2)$ and ρ , and integrating out $\mu^{S_1}, \dots, \mu^{S_{|\rho|}}$, model (2.9) - (2.10) becomes

$$p(\mathbf{x} | \rho = \{S_1, \dots, S_k\}, \boldsymbol{\theta}) = \prod_{i=1}^k p_{S_i}(\mathbf{x}_{S_i} | \boldsymbol{\theta}),$$

where $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{x}_S = (x_i, i \in S)$, and $p_S(\mathbf{x}_S | \boldsymbol{\theta})$ is the conditional density of \mathbf{x}_S given $S \in \rho$ and $\boldsymbol{\theta}$.

To define the prior distribution on partition, we need to adopt specific values for the set of all cohesions $c(S)$, for all nonempty $S \subset S_0$. Any specification is theoretically possible. In practice, however, there is $2^n - 1$ such numbers so that it is convenient to choose a functional form for $c(S)$. A simple way of accomplishing this is to use the existing connection between PPMs and nonparametric Bayesian models using Dirichlet

process priors (Ferguson, 1973; Antoniak, 1974) pointed out by Quintana and Iglesias (2003). A key advantage of this connection is that MCMC algorithms such as those by Bush and MacEachern (1996) or MacEachern and Müller (1998) can be adapted from the nonparametric to the parametric case. Essentially, both types of models have identical probability distribution on partitions, provided the cohesions in (2.11) are given by

$$c(S) = c \times (|S| - 1)!, \quad S \subseteq S_0, \quad (2.12)$$

where $|S|$ represents the number of elements in S . In addition to the computational advantages implied by the above choice, the behavior of the implied distribution on partitions can be further understood by noting that the prior mean and variance of the number of clusters are respectively given by

$$\sum_{i=1}^n \frac{c}{c+i-1} \quad \text{and} \quad \sum_{i=1}^n \frac{c(i-1)}{(c+i-1)^2}.$$

These values can be approximated, as n grows, by $c \log(1 + n/c)$ and $c\{\log(1 + n/c) - 1\}$ (Liu, 1996), expressions that can be useful for elicitation purposes if prior information is available on the number of clusters. Parameter c thus plays a key role in controlling the clustering structure. It can also be argued (Quintana and Iglesias, 2003) that for moderate values of c , e.g. $c = 1$ or $c = 2$, these cohesions yield a prior distribution that favors the formation of partitions with a small number of groups. This is a desirable feature of the model, because we do not want a prior probability model that stimulates partitions with too many subsets, i.e., excessively many outliers or change-points. Of course, the posterior distribution will reflect the adequate partition structure as required by the data. See further discussion on this and related issues in Section 5. The actual mechanism we use to select partitions is based on minimizing a loss function and will be described later in Section 3.

The prior distribution on θ is specified as follows. Assume that α is independent of

β and σ_x^2 given σ^2 and σ_x^2 is independent of σ^2 with

$$\alpha|\sigma^2 \sim N(a, \xi_{01}^2 \sigma^2) \quad (2.13)$$

$$\beta|\sigma^2 \sim N(b, \xi_{02}^2 \sigma^2) \quad (2.14)$$

$$\sigma^2 \sim IG(\nu_{02}, \delta_{02}) \quad (2.15)$$

$$\sigma_x^2 \sim IG(\nu_{01}, \delta_{01}) \quad (2.16)$$

Finally the model specification is completed by assigning a distribution to the latent variables $\mathbf{w}_j = (w_{ji}, i = 1, \dots, n)$ for $j = 1, 2, 3$. This is done according to the specific type of error we wish to consider. Typical options are $w_{ji} \equiv 1$ for normal errors, or $w_{ji} \sim IG(\nu_j/2, \nu_j/2)$ for $t(\nu_j)$ errors.

To fit the above normal mixture measurement error model, we adapt the algorithm proposed by Bush and MacEachern (1996) in the context of nonparametric Bayesian inference. Details are given in the Appendix.

3 Selecting a partition

To find the optimal partition, Quintana and Iglesias (2003) proposed a decision theoretic approach based on a loss function that depends on the objective of the analysis. Here we have a double purpose: on one hand we want accurate estimation of the model parameters, particularly α and β , and on the other hand we want to select a partition with the specific purpose of identifying outliers and/or detecting change-points. Let $\boldsymbol{\phi} = (\boldsymbol{\mu}, \mathbf{x}, \alpha, \beta, \sigma^2, \sigma_x^2, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ denote the vector of all parameters and latent variables in the model, and let M_ρ denote the model that corresponds to a given value ρ of the partition. Note that fixing ρ gives a specific meaning to $\boldsymbol{\mu}$: it sets the coordinates on the same subset to be equal, and different across subsets. Of course, this implies that all the remaining parameters change their meanings. Let $\boldsymbol{\phi}_\rho = (\boldsymbol{\mu}_\rho, \mathbf{x}_\rho, \alpha_\rho, \beta_\rho, \sigma_\rho^2, \sigma_{x,\rho}^2, \mathbf{w}_{1,\rho}, \mathbf{w}_{2,\rho}, \mathbf{w}_{3,\rho})$ denote the version of $\boldsymbol{\phi}$ that results when

fixing ρ .

The combined decision problem can be approached by minimizing a single loss function that takes into account all aspects of the problem as detailed earlier. Quintana and Iglesias (2003) considered a two-component loss function, that involved a specific term for parameter estimation, and a second term giving a penalization on model complexity, i.e., on the number of subsets $|\rho|$ in the partition. We now consider an extension of such definition, splitting the parameter vector ϕ into all relevant components and considering separate losses on an individual basis. Thus we define the following loss function:

$$l(M_\rho, \phi_\rho, \phi) = \frac{\kappa_1}{n} \|\boldsymbol{\mu}_\rho - \boldsymbol{\mu}\|^2 + \frac{\kappa_2}{n} \|\mathbf{x}_\rho - \mathbf{x}\|^2 + \kappa_3(\alpha_\rho - \alpha)^2 + \kappa_4(\beta_\rho - \beta)^2 + \kappa_5(\sigma_\rho^2 - \sigma^2)^2 + \kappa_6(\sigma_{x,\rho}^2 - \sigma_x^2)^2 + \frac{\kappa_7}{n} \sum_{j=1}^3 \|\mathbf{w}_{j,\rho} - \mathbf{w}_j\|^2 + \kappa_8|\rho|. \quad (3.17)$$

Here $\|\cdot\|$ is an appropriate norm, the weights $\boldsymbol{\kappa} = (\kappa_j, j = 1, \dots, 8)$ satisfy $\kappa_j \geq 0$ and $\sum_{j=1}^8 \kappa_j = 1$. More generally, it is straightforward to allow each component in (3.17) to be defined in terms of different distances or norms. By simplicity we have chosen to use a single norm, but all the later development can be easily stated in this general context.

The form of (3.17) and the fact that the coefficients add up to 1 allows us to view the loss function as a weighted average of losses corresponding to the various parameters in the model. The $\boldsymbol{\kappa}$ weights control the relative impact of each term when determining the optimal partition. In principle they can be defined arbitrarily, subject to the restrictions noted above. For a common loss function on all parameter components, one is typically interested on the case of unequal weights. For instance, note that κ_8 penalizes the number of clusters and κ_7 penalizes possible heteroskedasticity in the data. Thus, if $\sum_{i=3}^8 \kappa_i$ is close to 0, the selection is focused on cluster identification and if $0 < \kappa_4 < 1$, $\kappa_4 > \sum_{i=3}^8 \kappa_i$, cluster identification is associated to the estimation of β . Consequently, the selection of weights $\boldsymbol{\kappa}$ are directly determined by the purpose of study. When $\|\cdot\|$ is the Euclidean norm, the results in Quintana and Iglesias (2003) imply that minimizing

the expected value of (3.17) is equivalent to choosing ρ^* that minimizes

$$SC_{\kappa}(\rho) = \frac{\kappa_1}{n} \|\hat{\boldsymbol{\mu}}_{\rho} - \hat{\boldsymbol{\mu}}_B\|^2 + \frac{\kappa_2}{n} \|\hat{\boldsymbol{x}}_{\rho} - \hat{\boldsymbol{x}}_B\|^2 + \kappa_3(\hat{\alpha}_{\rho} - \hat{\alpha}_B)^2 + \kappa_4(\hat{\beta}_{\rho} - \hat{\beta}_B)^2 + \kappa_5(\hat{\sigma}_{\rho}^2 - \hat{\sigma}_B^2)^2 + \kappa_6(\hat{\sigma}_{x,\rho}^2 - \hat{\sigma}_{x,B}^2)^2 + \frac{\kappa_7}{n} \sum_{j=1}^3 \|\hat{\boldsymbol{w}}_{j,\rho} - \hat{\boldsymbol{w}}_{j,B}\|^2 + \kappa_8|\rho|, \quad (3.18)$$

where for any parameter η in the model,

$$\hat{\eta}_B = E(\eta|\mathbf{y}, \mathbf{z}) \quad \text{and} \quad \hat{\eta}_{\rho} = E(\eta|\rho, \mathbf{y}, \mathbf{z}).$$

From the earlier discussion it follows that to select the optimal ρ^* we need to exploit the specific underlying structure in the problem at hand. In the upcoming subsections we discuss appropriate algorithms for outlier detection and change-point identification.

3.1 Outlier Detection

Quintana and Iglesias (2003) proposed an algorithm to identify the optimal partition as part of a clustering problem. However, as mentioned in that paper, the algorithm does not necessarily lead to the optimal partition but it comes close to it in the special case of outlier detection. The iterative procedure works by detaching, one by one, the most outlying components from the main body of elements in the Bayes estimate of the target parameter vector $\boldsymbol{\eta}$ encoding the cluster structure. This process continues until no further improvement can be obtained according to (3.18). Concretely, let $S \subset \{1, \dots, n\}$ and denote

$$\bar{\boldsymbol{\eta}}_S = |S|^{-1} \sum_{i \in S} \hat{\boldsymbol{\eta}}_{B,i}(\mathbf{y}),$$

the average of the components of $\hat{\boldsymbol{\eta}}_B$ with index contained in S . For any $i \in S$, we determine the most outlying point using the discrepancy function $d(i, S) = \|\hat{\boldsymbol{\eta}}_{B,i} - \bar{\boldsymbol{\eta}}_S\|^2$, which measures the squared $\|\cdot\|$ -distance between component $i \in S$ and the average $\bar{\boldsymbol{\eta}}_S$. Here, $\|\cdot\|$ will typically refer to the Euclidean norm.

The specific steps in the algorithm are given next:

Step 1: Set $j = 1$, $S_1^1 = \{1, 2, \dots, n\}$, $\rho_1 = \{S_1^1\}$ and evaluate $SC_{\kappa}(\rho_1)$ using (3.18).

Step 2: Let $j = 2$, and obtain $k_1^* \in S_1^1$ such that $d(k_1^*, S_1^1) = \max_{i \in S_1^1} \{d(i, S_1^1)\}$.

Step 3: Increase the number of elements in the partition by 1, setting $S_1^2 = S_1^1 - \{k_1^*\}$, $S_2^2 = \{k_1^*\}$ and $\rho_2 = \{S_1^2, S_2^2\}$. Compute $SC_{\kappa}(\rho_2)$. If $SC_{\kappa}(\rho_2) > SC_{\kappa}(\rho_1)$ choose $\rho^* = \rho_1$ and stop. Otherwise, continue to the next step.

Step 4: Set $j = j + 1$ and find the element $k_{j-1}^* \in S_1^{j-1}$ such that $d(k_{j-1}^*, S_1^{j-1}) = \max_{i \in S_1^{j-1}} \{d(i, S_1^{j-1})\}$.

Step 5: Let $\tilde{\rho}_{j,i} = \{S_1^{j-1} - \{k_{j-1}^*\}, S_2^{j-1}, \dots, S_{i-1}^{j-1}, S_i^{j-1} \cup \{k_{j-1}^*\}, S_{i+1}^{j-1}, \dots, S_{|\rho_{j-1}|}^{j-1}\}$, for $i = 2, \dots, |\rho_{j-1}| + 1$. Define ρ_j implicitly as $SC_{\kappa}(\rho_j) = \min_{1 \leq i \leq |\rho_{j-1}|} \{SC_{\kappa}(\tilde{\rho}_{j,i})\}$. If $SC_{\kappa}(\rho_j) \geq SC_{\kappa}(\rho_{j-1})$ choose $\rho^* = \rho_{j-1}$ and stop. Otherwise go back to step 4, unless $j = n$ in which case $\rho^* = \rho_n$.

A natural choice for $\boldsymbol{\eta}$ is the vector (μ_1, \dots, μ_n) , but a more general approach can be taken by letting $\boldsymbol{\eta}$ represent the matrix with rows $(\mu_i, w_{1i}, w_{2i}, w_{3i})$.

3.2 Change-Point Identification

As mentioned above, the previous algorithm works well for identifying a small number of outlying points, but is not efficient at identifying change-points, because these involve partitions of an entirely different nature. We propose now a more suitable strategy for change-point detection. The basic procedure consists of recursively assessing subsequences of $S_0 = \{1, \dots, n\}$ and identifying change-points by splitting each subsequence into two parts. The steps are given next:

Step 1: Set $\mathcal{C} = \emptyset$, $l = 1$, $u = n$, $\rho^* = \{\{1, \dots, n\}\}$.

Step 2: In the current partition ρ^* , split the set $\{l, \dots, u\}$ into $\{l, \dots, j - 1\}$ and $\{j, \dots, u\}$ for $j = l + 1, \dots, u$. Denote by $\tilde{\rho}_j$ the corresponding partition. Let k^* be defined as $SC_{\kappa}(\tilde{\rho}_{k^*}) = \min_{l+1 \leq j \leq u-1} SC_{\kappa}(\tilde{\rho}_j)$.

Step 3: If $SC_{\kappa}(\tilde{\rho}_{k^*}) < SC_{\kappa}(\rho^*)$ then add k^* to \mathcal{C} , replace ρ^* by $\tilde{\rho}_{k^*}$, and recursively repeat Step 2 for $l = l, u = k^* - 1$ and for $l = k^*, u = u$. Otherwise, stop.

The above algorithm considers partitions that are specifically designed to detect change-points. Once a change-point is identified according to the expected loss criterion, the corresponding set is divided into two parts, and then other potential change-points in each of these two subsets are assessed by using the same procedure. It is worth pointing out that, to the best of our knowledge, no methods for globally detecting change-points in MEM are discussed in the statistical literature.

In the next section we illustrate the two algorithms described here.

4 Applications

4.1 Outlier identification

In this section we consider first an application of the results considered in the previous sections to the data set reported in Kelly (1984). The data consist of measurements of serum kanamycin level in blood samples drawn from twenty babies. One of the measurements was obtained by a heelstick method (X) and the other by using an umbilical catheter (Y). The basic question was whether the catheter values systematically differed from the heelstick values. Given the special nature of the data, it was reasoned that $\lambda = 1.0$ was correct. We present now a reanalysis of these data by using a Student-t structural model with equal variances described earlier.

In this example we used the following hyperparameter values: $a = 0$, $b = 1$, $m = 25$, $\nu_{01} = \nu_{02} = 2.01$, $\delta_{01} = \delta_{02} = 1.01$, $\xi_{01}^2 = \xi_{02}^2 = 0.1$ and $\tau_0^2 = 4$. In addition, we chose $c = 1$ so that the prior mean and variance of the number of clusters are 3.598 and 2.002, respectively.

Some marginal posterior summaries for selected parameters are indicated in Table 1, considering the model with normal errors:

	Median	Mean	MC Error	95% HPD
α	-0.082	-0.083	0.0026	(-1.334,1.177)
β	1.017	1.017	0.0002	(0.937,1.102)
# Clusters	4.000	4.030	0.0049	(2.000,6.000)
σ^2	3.948	4.187	0.0055	(2.060,6.820)
σ_x^2	3.150	3.745	0.0104	(0.885,7.992)

Table 1: *Posterior summaries for selected parameters in the serum kanamycin levels in blood example.*

Next, we ran the clustering algorithm described in Section 3.1. The weights for the loss function in this particular application were chosen to be as follows:

$$\boldsymbol{\kappa} = \frac{1}{37.1} (10, 1, 10, 10, 3, 3, 0.1, 0).$$

As argued earlier, these values assign more importance to the regression parameters α and β , putting no restriction to the formation of clusters. Note also that the partition structure is explicitly encoded in $\boldsymbol{\mu}$ which is why we weigh more heavily the μ_i parameters rather than the x_i s.

Intuitively, the algorithm evaluates partitions, selecting the one with smallest expected loss. In this process, a key role is played by the Bayes estimate $\boldsymbol{\mu}_B$, because partitions to be evaluated are determined by selecting those components in $\boldsymbol{\mu}_B$ that deviate most from the mean. The algorithm applied to this problem converged to the

partition

$$\{\{S_0 - \{1, 2, 4, 11, 12, 17, 18\}\}, \{1, 11, 12\}, \{2, 17, 18\}, \{4\}\},$$

which is depicted in Figure 1. The plot reveals that the three small clusters are formed with the 7 observations with posterior means μ_i departing the most from the main body of points. The specific assignments of these points to the subsets is governed by the above choice of κ weights.

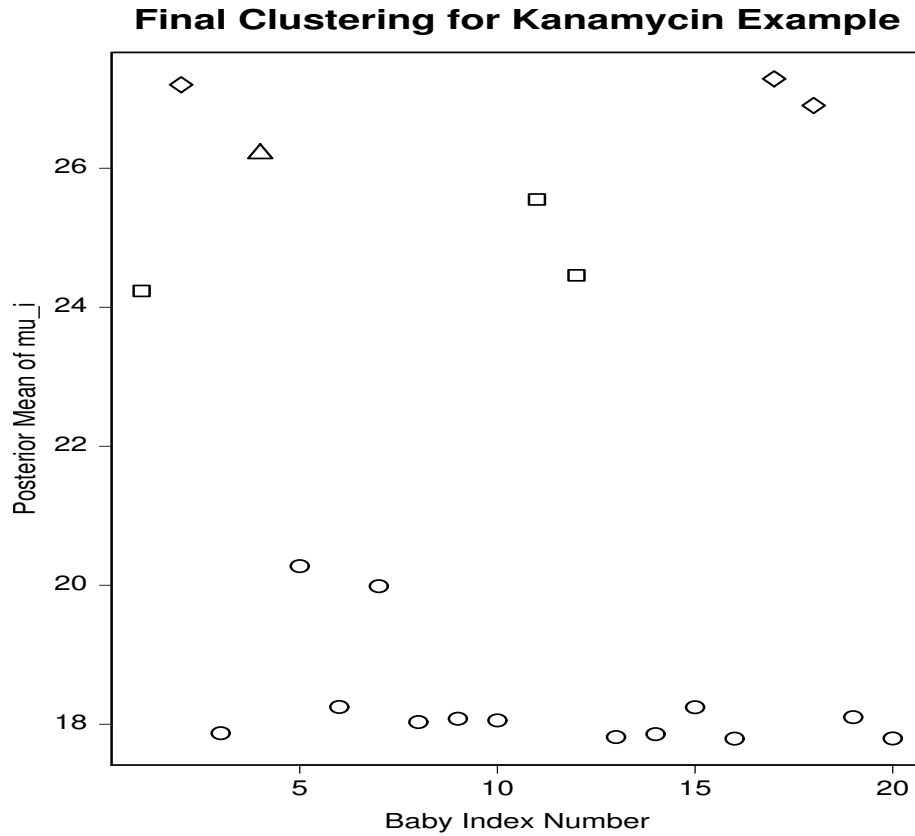


Figure 1: *Optimal clustering for kanamycin example. Here we identified 4 clusters, which correspond to observations 1, 11 and 12 (squares), 2, 17 and 18 (diamond), 4 (triangle) and the rest (circles)*

Finally, we point out that when repeating the analysis with values of λ given by 0.5, 2 and 5, the same partitions were obtained (data not shown). This suggests robustness under different choices of λ .

Remark 4.1: We point out that the approach proposed by Kelly (1984) for outlier detection, requires a more complex structure than the one we have discussed. Nevertheless, we obtained nearly the same results for this particular data set.

4.2 Change-point identification

We consider now a dataset reported in McGee and Carleton (1970) and also analyzed by Holbert (1982) and Chang and Huang (1997). Here, it is of interest to investigate the relationship between the monthly dollar volume of sales on the Boston Stock Exchange (Y) and the combined monthly dollar volumes for the New York and American Stock Exchanges (X). The data consists of $n = 35$ records, corresponding to the period January 1967 – November 1969. To avoid problems related to excessively different scales, we divided the y values by 100.

In this example we used the following hyperparameter values: $a = 0$, $b = 1$, $m = 140$, $\nu_{01} = 2.64$, $\delta_{01} = 656$, $\nu_{02} = 335.33$, $\delta_{02} = 33433.33$, $\xi_{01}^2 = 1$, $\xi_{02}^2 = 0.5$ and $\tau_0^2 = 4$. With this, a priori we have $E(\sigma_x^2) = 400$, $Var(\sigma_x^2) = 250000$, $E(\sigma^2) = 100$ and $Var(\sigma^2) = 30$. In addition, we chose $\lambda = 1$ and $c = 1$, which implies that the prior mean of the number of clusters is 4.147 and the variance is 2.530.

	Median	Mean	MC Error	95% HPD
α	-57.873	-57.960	0.0434	(-74.546,-41.804)
β	1.353	1.354	0.0003	(1.231,1.483)
# Clusters	4.000	4.491	0.0064	(2.000,8.000)
σ^2	123.224	123.459	0.0289	(111.200,137.023)
σ_x^2	516.162	539.488	0.6640	(270.298,947.881)

Table 2: *Posterior summaries for selected parameters in the Boston Stock Exchange example.*

Posterior summaries can be found in Table 2. We began the analysis by applying

the clustering algorithm in Section 3.1, using

$$\kappa = \frac{1}{10122.1}(1, 0.1, 100, 10000, 10, 10, 1, 0),$$

which found no outlier. Next, we applied the change-point version of the algorithm, as explained in Section 3.2 which converged to the partition

$$\{\{1, \dots, 9\}, \{10, \dots, 19\}, \{20, 21\}, \{22, \dots, 35\}\}.$$

In other words, changes occurred at points 10, 20 and 22. Compared to Chang and Huang (1997), our analysis agrees on point 10, but detects points 20 and 22 rather than 19 and 24. The difference is not substantial though. See Figure 2.

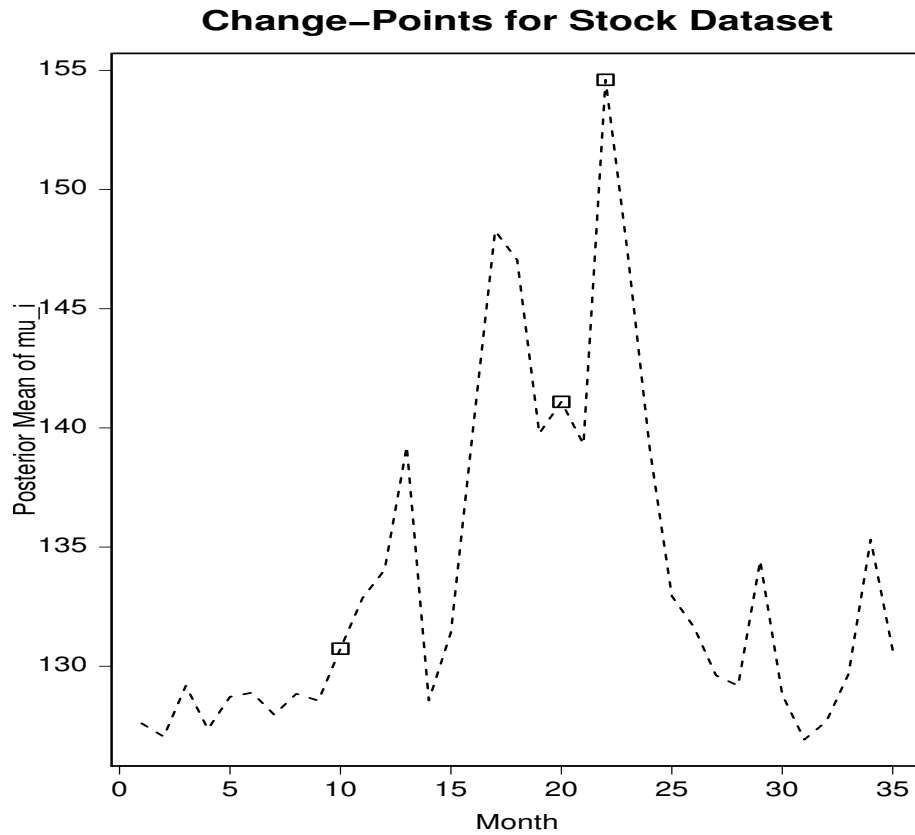


Figure 2: *Change-point identification (marked as a square) for the Boston Stock Exchange dataset.*

5 Conclusions

Our proposal follows a coherent approach from a Bayesian decision theory. Moreover, it robustly estimates model parameters, incorporating the possibility of outlier or change-point. We point out that Chang and Huang (1997) approach is not robust with respect departures of normality. On the contrary our approach naturally incorporates the possibility of non-normal distributions with heavier tails than the normal distributions as for example the student-t distribution considered in this paper. We are currently investigating the possibility of mixing the two algorithms, for cases where it is suspected that the data set contains both outliers and change-points. The approach can also be used, with only a slight modification, for the problem of detecting change-points in α , β and σ_x^2 .

The cohesion functions $c(S)$ govern the partition structure in the context of PPMs, their specification being a key aspect of the modeling process. Of all possible ways to do this, we chose the DP-style case, which has the computational advantages described in Section 2. But such cohesions also imply a simple prior structure that may be advantageous from a prior elicitation viewpoint. Parameter c plays an important role in controlling the number of clusters in the prior PPM. For the purpose of outlier detection and change-point identification our recommendation is to choose a default value of $c = 1$ or $c = 2$. Additional runs with $c = 5$ in the outlier detection example from Section 4.1 converged to the same partition, which suggest the robustness of the procedure (data not shown). An alternative approach would consider a prior distribution on c . Learning about c may be quite useful for parameter estimation or other inference problems. For our clustering application, however, the robustness mentioned earlier suggests that a prior on c will most likely not change the selected partitions. Additionally, setting $c \sim p(c)$ for some convenient specification of $p(c)$ complicates the simple interpretations in terms of prior mean and variance of the number of clusters. Therefore, we do not recommend it.

The loss function (3.17) combines all the aspects of the decision problem underlying the selection of a partition, as desired. Of course other alternative specifications can be used. However, note that the additive structure in (3.17) allows handling the various parameters separately, which simplifies the computational and theoretical approach. The choice of Euclidean norm provides further analytical simplifications, but this is not too critical in the end, as the MCMC scheme allows deriving virtually any required posterior summaries, such as medians or percentiles.

Appendix: MCMC details

We now give a detailed account of all the conditional distributions required for MCMC implementation. From (2.6) to (2.16) we obtain the following joint posterior density

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{x} | \mathbf{y}, \mathbf{z}) \propto \exp \left\{ -\frac{\delta_{01}}{\sigma_x^2} - \frac{\delta_{02}}{\sigma^2} - \frac{1}{2\lambda\sigma^2} \sum_{i=1}^n \frac{(y_i - \alpha - \beta x_i)^2}{w_{1i}} - \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(z_i - x_i)^2}{w_{2i}} \right. \\
- \frac{1}{2\sigma_x^2} \sum_{j=1}^k \sum_{i \in S_j} \frac{(x_i - \mu^{S_j})^2}{w_{3i}} + \frac{1}{2\tau_0^2 \sigma_x^2} \sum_{j=1}^{|\rho|} (\mu^{S_j} - m)^2 - \frac{1}{2\xi_{01}^2 \sigma^2} (\alpha - a)^2 - \frac{1}{2\xi_{02}^2 \sigma^2} (\beta - b)^2 \\
\left. - \frac{d}{2} \sum_{j=1}^3 \sum_{i=1}^n \frac{1}{w_{ji}} \right\} (\sigma)^{-\{2+\nu_{02}+\frac{2n}{2}\}} \times (\sigma_x^2)^{\{1+\nu_{01}+\frac{n+|\rho|}{2}\}} \prod_{j=1}^3 \prod_{i=1}^n \frac{p(w_{ji})}{w_{ji}^{1/2}} \prod_{i=1}^k c(S_i).
\end{aligned}$$

We adapt here the algorithm by Bush and MacEachern (1996). Thus, for given starting values $\boldsymbol{\mu}_0$, \mathbf{x}_0 , α_0 , β_0 , σ_{x0}^2 , and σ_0^2 , we need to sample from the following conditional distributions:

$$\beta | \mathbf{z}, \mathbf{y}, \alpha, \sigma_x^2, \sigma^2, \mathbf{x}, \boldsymbol{\mu}, \mathbf{w} \sim N \left(\frac{\sum_{i=1}^n x_i (y_i - \alpha) / w_{1i} \lambda + b / \xi_{02}^2}{\sum_{i=1}^n x_i^2 / \lambda w_{1i} + 1 / \xi_{02}^2}, \frac{\sigma^2}{\sum_{i=1}^n x_i^2 / \lambda w_{1i} + 1 / \xi_{02}^2} \right),$$

$$\alpha | \mathbf{z}, \mathbf{y}, \beta, \sigma_x^2, \sigma^2, \mathbf{x}, \boldsymbol{\mu} \sim N \left(\frac{\sum_{i=1}^n (y_i - \beta x_i) / w_{1i} \lambda + a / \xi_{01}^2}{\sum_{i=1}^n 1 / \lambda w_{1i} + 1 / \xi_{01}^2}, \frac{\sigma^2}{\sum_{i=1}^n 1 / w_{2i} \lambda + 1 / \xi_{01}^2} \right),$$

$$\sigma^2 | \mathbf{z}, \mathbf{y}, \alpha, \sigma_x^2, \mathbf{x}, \mu \sim GI \left(n + \nu_{02} + 1, \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 / 2w_{1i} \lambda + \sum_{i=1}^n (z_i - x_i)^2 / 2w_{2i} + (\alpha - a)^2 / 2\xi_{01}^2 + (\beta - b)^2 / 2\xi_{02}^2 + \delta_{02} \right),$$

$$\sigma_x^2 | \mathbf{z}, \mathbf{y}, \alpha, \sigma^2, \mathbf{x}, \mu \sim GI \left(n/2 + \rho/2 + \nu_{01}, \frac{1}{2} \sum_{j=1}^{|\rho|} \sum_{i \in S_j} (x_i - \mu^{S_j})^2 / w_{3i} + \frac{1}{2\tau_0^2} \sum_{j=1}^{|\rho|} (\mu^{S_j} - m)^2 + \delta_{01} \right),$$

$$x_i | \mathbf{z}, \mathbf{y}, \beta, \mathbf{x}_{-i}, \mu, \sigma_x^2, \sigma^2 \sim N \left(\frac{(y_i - \alpha)\beta / (w_{1i}\lambda\sigma^2) + z_i / w_{2i}\sigma^2 + \mu_i / w_{3i}\sigma_x^2}{\beta^2 / (w_{1i}\lambda\sigma^2) + 1/w_{2i}\sigma^2 + 1/w_{3i}\sigma_x^2}, \frac{1}{\beta^2 / (w_{1i}\lambda\sigma^2) + 1/w_{2i}\sigma^2 + 1/w_{3i}\sigma_x^2} \right),$$

$$\mu_i | \mu_{-i}, \sigma_x^2, \mathbf{z}, \alpha, \sigma^2, \mathbf{x} \propto \sum_{j \neq i} \exp \left\{ -\frac{1}{\sigma_x^2} (x_i - \mu_j)^2 / w_{3i} \right\} \delta_{\mu_j}(\mu_i) + \frac{c}{\sqrt{1 + \tau_0^2 / w_{3i}}} \exp \left\{ -\frac{(m - x_i)}{2\sigma_x^2 (w_{3i} + \tau_0^2)} \right\} N \left(\frac{x_i / w_{3i} + m / \tau_0^2}{1/w_{3i} + 1/\tau_0^2}, \frac{\sigma_x^2}{1/w_{3i} + 1/\tau_0^2} \right),$$

$$w_{1i} | \mathbf{w}_{-1i}, \mathbf{y}, \mathbf{z}, \beta, \sigma_x^2, \sigma^2, \mathbf{x}, \mu \propto \exp \left\{ \frac{(y_i - \alpha - \beta x_i)^2}{2\lambda\sigma^2 w_{1i}} \right\} w_{1i}^{-1/2} p(w_{1i}),$$

$$w_{2i} | \mathbf{w}_{-2i}, \mathbf{y}, \mathbf{z}, \beta, \sigma_x^2, \sigma^2, \mathbf{x}, \mu \propto \exp \left\{ \frac{(z_i - x_i)^2}{2\sigma^2 w_{2i}} \right\} w_{2i}^{-1/2} p(w_{2i})$$

$$w_{3i} | \mathbf{w}_{-3i}, \mathbf{y}, \mathbf{z}, \beta, \sigma_x^2, \sigma^2, \mathbf{x}, \mu \propto \exp \left\{ \frac{(x_i - \mu_i)^2}{2\sigma_x^2 w_{3i}} \right\} w_{3i}^{-1/2} p(w_{3i}),$$

where $\mathbf{a}_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)^T$, for any vector $\mathbf{a} = (a_1, \dots, a_n)^T$.

In the special case

$$w_{ji} \stackrel{ind}{\sim} IG\left(\frac{d_j}{2}, \frac{d_j}{2}\right), \quad j = 1, 2, 3,$$

then

$$\begin{aligned} w_{1i} | \boldsymbol{\theta}, \mathbf{x} &\sim IG\left(\frac{d_1 + 1}{2}, \frac{d_1}{2} + \frac{(y_i - \alpha - \beta x_i)^2}{2\lambda\sigma^2}\right), \\ w_{2i} | \boldsymbol{\theta}, \mathbf{x} &\sim IG\left(\frac{d_2 + 1}{2}, \frac{d_2}{2} + \frac{(z_i - x_i)^2}{2\sigma^2}\right), \\ w_{3i} | \boldsymbol{\theta}, \mathbf{x} &\sim IG\left(\frac{d_3 + 1}{2}, \frac{d_3}{2} + \frac{(x_i - \mu_i)^2}{2\sigma_x^2}\right). \end{aligned}$$

Finally, to update the unique values μ^{S_j} we use

$$\mu^{S_j} | \boldsymbol{\mu}_{-i}, \sigma_x^2, \mathbf{z}, \alpha, \sigma^2, \mathbf{x} \sim N\left(\frac{\sum_{i \in S_j} x_i/w_{3i} + m/\tau_0^2}{\sum_{i \in S_j} 1/w_{3i} + 1/\tau_0^2}, \frac{\sigma_x^2}{\sum_{i \in S_j} 1/w_{3i} + 1/\tau_0^2}\right),$$

for $i = 1, \dots, n$.

References

Abdullah, M.B. (1995). Detection of influential observations in functional error in variables model. *Communications in Statistics, Theory and Methods*, **24**, 1585-1595.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, **2**, 1152-1174.

Arellano-Valle, R.B. and Bolfarine, H. (1996). Elliptical functional models. *Communication in Statistics*, **25**, 2319-2341.

Arellano-Valle, R., Galea-Rojas, M. and Iglesias, P. (2000). Bayesian sensitivity analysis in elliptical linear regression models. *Journal of Statistical Planning and Inference*, **86**, 175-199.

Barnett, V. and Lewis, T. (1994) *Outliers in statistical data, Third Edition*. New York: John Wiley.

Barry, D. and Hartigan, J.A. (1992). Product Partition Models for Change Point Problems. *The Annals of Statistics*, **20**, 260–279.

Bolfarine, H. and Cordani, L. (1993). Estimation of a structural regression model with known reliability ratio. *Annals of the Institute of Statistical Mathematics*, **45**, 531-540.

Broemeling, L.D. (1995). *Bayesian analysis of linear models*. New York: Marcel Dekker.

Brown, P. (1993). *Measurement, regression and calibration*. Oxford.

Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs, *Biometrika*, **83**, 275–285.

Carroll, R., Rupert, D. and Stefanski, L. (1995). *Nonlinear measurement error models*. New York: Wiley.

Chang, Y.-P. and Huang, W.-T. (1997). Inference for the linear error-in variables with changepoint models. *Journal of the American Statistical Association*, **92**, 171-178.

Carlstein, E., Müller, H.G. and Seigmund, D. (1994). Change-point problem, IMS lecture Notes-Monograph Series, Volume 23, 28-56.

Cheng, C.-L. and Van Ness (1999). *Statistical regression with measurement error*. Arnold, London.

Crowley, E.M. (1997). Product partition models for normal means, *Journal of the American Statistical Association*, **92**, 192-198.

Dellaportas, P. and Stephens, D.A. (1995). Bayesian analysis of error-in-variables regression models. *Biometrics*, **51**, 1085-1095.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Fuller, W. (1987). *Measurement error models*. Wiley, N.Y.

Galea, M., Bolfarine, H. and Vilca, F. (2002). Influence and diagnostic for structural error-in-variable models under the student t distribution. *Journal of Applied Statistics*, **29**, 1191 - 1204.

Galea, M., Bolfarine, H. and de Castro, M. (2002). Local influence in comparative calibration models. *Biometrical Journal*, **41**, 59-81.

Hartigan, J. (1990). Product Partition Models, *Communications in Statistics – Theory and Methods*, **19**, 2745–2756.

Hawkins, D.M. and Olive, D.J. (2002) Inconsistency of high breakdown estimators for high breakdown regression estimators and a new algorithm (with discussion). *Journal of the American Statistical Association*, **97**, 136-148, rejoinder 156-159.

Holbert, D. (1982), A Bayesian analysis of a switching linear model, *Journal of Econometrics*, **19**, 77-87.

Kelly, G. (1984). The influence function in the error-in-variables problems. *The Annals of Statistics*, **12**, 87-100.

Lee, A.H. and Zhao, Y. (1996). Assessing local influence in measurement error models, *Biometrical Journal*, **38**, 829- 841.

Lindley, D.V. and El-Sayyad, G.M.(1968). The Bayesian estimation of a linear functional relationship. *Journal of the Royal statistical Society, Series B*, **30**, 190- 202.

Liu, J. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, **24**, 911–930.

Loschi, R. and Cruz, F.R.B. (2005). Extensions to the product partition model:

Computing the probability of a change. *Computational Statistics and Data Analysis*, **48**(2), 255–268.

MacEachern, S.N. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

McGee, V.E. and Carleton, W.T. (1970). Piecewise regression, *Journal of the American Statistical Association*, **65**, 1109-1124.

Quintana, F.A. and Iglesias, P.L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B*, **65**, 557-574.

Richardson, S. and Gilks, W.R. (1993). A Bayesian approach to measurement error problems in epidemiological studies with covariate measurement error. *Statistics in Medicine*, **12**, 1703-1722.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components; An alternative to reversible jump methods. *The Annals of Statistics*, **28**, 40-74.

Stephens, M. and Dellaportas, P. (1992). Bayesian analysis of generalized linear measurement error models. In *Bayesian statistics 4*. J.M. Bernardo, J. O. Berger, A. P. Dawid and A.F.M. Smith (eds), pp 813-820. Oxford University Press.