Bayesian density estimation from grouped continuous data

Philippe Lambert*,a, Paul H.C. Eilersb,c

[a] *Université de Liège, Institut des sciences humaines et sociales, Méthodes quantitatives en sciences sociales, Boulevard du Rectorat 7 (B31), B-4000 Liège, Belgium*
[b] *Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands*
[c] *Data Theory Group, Leiden University, Leiden, The Netherlands*

Abstract

Grouped data occur frequently in practice, either because of limited resolution of instruments, or because data have been summarized in relatively wide bins. A combination of the composite link model with roughness penalties is proposed to estimate smooth densities from such data in a Bayesian framework. A simulation study is used to evaluate the performances of the strategy in the estimation of a density, of its quantiles and first moments. Two illustrations are presented: the first one involves grouped data of lead concentrations in the blood and the second one the number of deaths due to tuberculosis in The Netherlands in wide age classes.

Key words: Histogram smoothing, Density estimation, Grouped data, Symbolic data, P-splines, Langevin-Hastings algorithm.

1. Introduction

Descriptive statistics are easily computed when data are available with high precision. Unfortunately, this is not always the case. Measured concentrations may be below the detection limit, or an instrument may have poor resolution. It may also be the case that precise data have been summarized with a small number of wide intervals. An obvious quick-and-dirty solution replaces the observations in each interval by its midpoint and handles them as if they were actually observed. This can work quite well when computing the mean and the variance, but it fails for the estimation of quantiles.

An improvement is to assume a parametric model for the underlying distribution, and fit it to the grouped data with the EM (estimation-maximization) algorithm. If an approximation to the distribution is available, one distributes the counts in each wide interval to pseudo-counts on a grid of narrow intervals, proportionally to the current approximation (the E step). The midpoints of the narrow intervals are used as "precise data", with weights proportional to their pseudo-counts, to estimate distribution parameters by, say, maximum likelihood (the M step). This process is repeated till convergence.

When the observed coarse distribution has a simple shape, the parametric method can work quite well. But when it is skewed or multi-modal, or both, a lot of trial-and-error may be needed to find the right model. A non-parametric model will be more attractive then.

Histosplines are a popular non-parametric method. The idea is to compute a smooth spline under the condition that integrals over the given wide intervals are

*Tel: +32-4-366.59.90; fax: +32-4-366.47.51.
Email address: p.lambert@ulg.ac.be (Philippe Lambert)
URL: http://www.statsoc.ulg.ac.be/ (Philippe Lambert)

equal to the observed counts. Early work in this area was reported by Boneva et al. (1971). Although the computations are not complicated, histosplines are not without disadvantages. There is no guaranty that the estimated distribution will be non-negative everywhere. Especially when there are wide intervals with zero observations next to intervals with positive counts, the spline, maintaining its smoothness, may undershoot the horizontal axis. A more fundamental objection is that the sampling variation in the observed counts is ignored. Generally one sees that a histospline shows a number of unrealistic wiggles.

Other non-parametric methods have been proposed. Braun et al. (2005) use an EM-type algorithm to generalize the kernel density estimate for interval censored data and propose to select the bandwidth using cross-validation. The described strategy is partially implemented in an R package named `ICE` where the crucial choice of the bandwidth is left to the user.

We present here a non-parametric spline-based model: 1) the logarithm of a smooth latent distribution is modelled with P-splines on a fine grid, 2) expected counts are obtained from integrals over the wide intervals, and 3) observed counts are modeled with a multinomial distribution, having the given expectations. The result is a penalized composite link model (Eilers, 2007).

We introduce a Bayesian variant of the model, allowing uncertainties to be quantified for model components and derived quantities, like quantiles. As is common for complex Bayesian models, analytic results are not obtainable, so we use a simulation algorithm. A combination of the Langevin-Hastings algorithm and rotation of the P-spline parameters allows fast computation, making the model a practical tool for everyday use.

2. Density estimation from grouped data

Assume that one is interested in estimating a discrete representation of a continuous density $f_Y(\cdot)$ of a random variable Y on a fine grid on (a,b). Fine here means that the grid spacing is small enough to give an accurate description of the density, for plotting it or for computing quantiles or other statistics accurately. The fine grid may consist of, say, 100 or more grid points, partitioning (a,b) into I consecutive intervals $\mathcal{I}_i = (x_{i-1}, x_i)$ of equal width Δ with midpoints u_i $(i = 1 \ldots I)$, $x_0 = a$ and $x_I = b$. Then, the quantities of interest are $\pi_i = \int_{\mathcal{I}_i} f_Y(t)dt \approx f_Y(u_i)\Delta$.

Let m_j $(j = 1 \ldots J)$ be the number of observations belonging to each of (say 10 or less) given non-overlapping wide bins $\mathcal{J}_j = (L_j, U_j)$ partitioning (a,b). For simplicity, assume for now that the limits of these (wide) bins make a subset of $\{x_0, \ldots, x_I\}$. The relationship between the wide bins and the initial grid is coded by the J by I matrix $C = [c_{ji}]$, where $c_{ji} = 1$ if $\mathcal{I}_i \subset \mathcal{J}_j$ and 0 otherwise. More sophisticated settings will be discussed in Section 4.

If γ_j denotes the probability $\int_{\mathcal{J}_j} f_Y(t)dt$ that Y belongs to wide bin $\mathcal{J}_j$, then one has $\gamma_j = \sum_{i=1}^{I} c_{ji}\pi_i$; in vector-matrix notation: $\boldsymbol{\gamma} = C\boldsymbol{\pi}$. If the only available data are the frequencies associated to the wide bins, then the estimation of the π_i's is an ill-conditioned problem.

Therefore, we demand π to be smooth. Our strategy has two components: 1) model the logarithm of π with a generous number of B-splines, and 2) add a discrete roughness penalty on the B-spline coefficients. This the P-spline approach advocated by Eilers and Marx (1996) and Eilers (2007).

More specifically, consider a basis $\{b_k(\cdot) : k = 1, \ldots, K\}$ of B-splines associated to equidistant knots on (a,b). Denote by $(B)_{ik} = b_{ik} = b_k(u_i)$ the $I \times K$ matrix giving the basis functions evaluated at the midpoint u_i $(i = 1, \ldots, I)$ of I_i. Then, we model π_i by

$$\pi_i = \pi_i(\boldsymbol{\phi}) = \frac{\mathrm{e}^{\eta_i}}{\mathrm{e}^{\eta_1} + \mathrm{e}^{\eta_2} + \ldots + \mathrm{e}^{\eta_I}} \ ,$$

with $\boldsymbol{\eta} = B\boldsymbol{\phi}$ and the identifiability constraint $\sum_{k=1}^{K} \phi_k = 0$.

The P-spline penalty is based on rth order differences, $\Delta^r \boldsymbol{\phi}$, of the splines coefficients $\boldsymbol{\phi}$. In a Bayesian setting, it translates into a prior distribution on the splines coefficients (Lang and Brezger, 2004):

$$\Delta^r \phi_k \sim N(0, \tau^{-1}).$$

This idea has been used successfully in several papers in various contexts (see e.g. Berry et al. (2002) in normal regression models and Lang and Brezger (2004) in additive models, Lambert and Eilers (2005) in survival analysis, Lambert (2007) in copula estimation). Consequently, the following (improper) prior for the B-splines coefficients is assumed

$$p(\boldsymbol{\phi}|\tau) \propto \tau^{\mathcal{R}(P)/2} \exp\left\{-\frac{\tau}{2}\ \boldsymbol{\phi}' P \boldsymbol{\phi}\right\} \ , \tag{1}$$

where $\phi_K = 1 - \sum_{k=1}^{K-1} \phi_k$ (identifiability constraint), $\mathcal{R}(P)$ denotes the rank of P and $P = D'D$ is the matrix such that

$$\sum_k (\Delta^r \phi_k)^2 = \boldsymbol{\phi}' P \boldsymbol{\phi} \ .$$

For example, with $r = 2$, one has

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & -2 & 1 \end{bmatrix},$$

and $\mathcal{R}(P) = K - 2$.

A gamma prior $\mathcal{G}\,(a,\ b)$ with a large variance (as obtained by taking $a = b = .0001$, say) is usually advocated (Lang and Brezger, 2004) to express prior ignorance about suitable values for τ. Alternative robust priors can be found in Jullion and Lambert (2007).

Apart from the penalty, the model for $\boldsymbol{\pi}$ is a familiar generalized linear model (GLM), but the model for $\boldsymbol{\gamma}$ is a composite link model (CLM), as introduced by Thompson and Baker (1981).

3. Inference

3.1. Frequentist estimation

Thompson and Baker (1981) showed how to estimate the parameters of the CLM in a maximum likelihood setting. It boils down to a polytomous logistic regression of $\mathbf{m} = (m_1, \ldots, m_J)'$ on a "working" matrix $X = W^{-1}CHB$, with

$$H = \mathrm{diag}(m_+ \boldsymbol{\pi}(1 - \boldsymbol{\pi})), W = \mathrm{diag}(m_+ \boldsymbol{\gamma}(1 - \boldsymbol{\gamma})) \text{ and } m_+ = \sum_{j=1}^{J} m_j \ .$$

The well-known scoring algorithm leads, at iteration $(t + 1)$, to

$$X_t' W_t X_t \boldsymbol{\phi}_{t+1} = X_t'(\boldsymbol{m} - m_+ \boldsymbol{\gamma} + W_t X_t \boldsymbol{\phi}_t),$$

$\boldsymbol{\phi}_t$ denoting the current approximation.

The penalized log-likelihood is

$$l_{pen} = \sum_j m_j \log \gamma_j - \frac{\tau}{2} \; \boldsymbol{\phi}' D' D \boldsymbol{\phi} \ ,$$

The penalty modifies the scoring algorithm slightly:

$$(X_t' W_t X_t + \tau D' D) \boldsymbol{\phi}_{t+1} = X_t' (\boldsymbol{m} - m_+ \boldsymbol{\gamma} + W_t X_t \boldsymbol{\phi}_t), \tag{2}$$

Different strategies can be used to select the penalty parameter τ: cross-validation or information criteria like the AIC or BIC (Eilers and Marx, 1996) are possible guidelines. For a given τ, the variance-covariance of the MLE is, at convergence, given by

$$(X_t' W_t X_t + \tau D' D)^{-1}. \tag{3}$$

3.2. Posterior distribution

Combining (using Bayes' formula) the priors defined in Section 2 with the multinomial likelihood resulting from the observed frequencies for the wide bins, one obtains the following joint proper posterior for $\boldsymbol{\theta} = (\boldsymbol{\phi}, \tau)$:

$$p(\boldsymbol{\phi}, \tau | \mathcal{D}) \propto \left(\prod_{j=1}^{J} \gamma_j^{m_j} \right) \tau^{a+0.5\mathcal{R}(P)-1} \exp\left\{ -\tau (b + 0.5 \; \boldsymbol{\phi}' P \boldsymbol{\phi}) \right\}, \tag{4}$$

where $\boldsymbol{\gamma} = C\boldsymbol{\pi}(\boldsymbol{\phi})$, $\mathcal{D}$ stands for the observed frequencies associated to the (wide) bins, $\{(m_j, \mathcal{J}_j) : j = 1, \ldots, J\}$, and $\phi_K = 1 - \sum_{k=1}^{K-1} \phi_k$ (identifiability constraint).

Under that constraint, the conditional posterior distributions can be shown to be

$$\begin{aligned} (\tau | \boldsymbol{\phi}, \mathcal{D}) \quad &\sim \quad \mathcal{G}\left(a + 0.5 \; \mathcal{R}(P), \; b + 0.5 \; \boldsymbol{\phi}' P \boldsymbol{\phi} \right) \ , && (5) \\ p(\boldsymbol{\phi} | \tau, \mathcal{D}) \quad &\propto \quad \left(\prod_{j=1}^{J} \gamma_j^{m_j} \right) \exp\left\{ -\frac{\tau}{2} \; \boldsymbol{\phi}' P \boldsymbol{\phi} \right\} \ . && (6) \end{aligned}$$

3.3. Exploring the posterior distribution

Markov chain Monte Carlo (MCMC) methods can be used to draw a sample, $\{(\boldsymbol{\phi}^{(m)}, \tau^{(m)}) : m = 1, \ldots, M\}$, from the posterior. Here, we use a 'Metropolis-within-Gibbs' strategy with a Gibbs step for τ (see Eq. (5)) and a Metropolis step through the (modified) Langevin-Hastings algorithm (see Section 3.3.1) for $\boldsymbol{\phi}$ with Eq. (6) . Thus, the proposed algorithm for building a chain $\{(\boldsymbol{\phi}^{(m)}, \tau^{(m)}) : m = 1, \ldots, M\}$ (after an appropriate burn-in) is

```
For $m = 1, \ldots, M$:
    -1- Draw $\boldsymbol{\phi}^{(m)}$ from $p(\boldsymbol{\phi} | \tau^{m-1}, \mathcal{D})$ using Langevin-Hastings.
    -2- Generate $\tau^{(m)}$ from $\mathcal{G}\left( a + 0.5 \; \mathcal{R}(P), \; b + 0.5 \; \boldsymbol{\phi}^{(m)'} P \boldsymbol{\phi}^{(m)} \right)$ .
```

To each element of the so-obtained Monte Carlo sample, $\boldsymbol{\phi}^{(m)}$, corresponds a density $f^{(m)}(y)$ for which any summary measure $\xi^{(m)}$ of interest such as the mean, the standard deviation or quantiles can be computed. Point estimates and credible intervals for ξ can be derived from the so-obtained sample $\{\xi^{(m)} : m = 1, \ldots, M\}$.

Specific properties such as unimodality or log-concavity can be imposed on the estimated density by excluding, through the prior, the configurations of $\boldsymbol{\phi}$ corresponding to non-desirable densities. If Φ denotes the set of $\boldsymbol{\phi}$ meeting the identifiability constraint and ensuring the desired properties for the corresponding density $\boldsymbol{\pi}(\boldsymbol{\phi})$, then the following equation will be substituted to the prior for $\boldsymbol{\phi}$ in Eq. (1):

$$p(\boldsymbol{\phi}|\tau) \propto \tau^{\mathcal{R}(P)/2} \exp\left\{-\frac{1}{2}\ \tau\ \boldsymbol{\phi}' P \boldsymbol{\phi}\right\}\ I_\Phi(\boldsymbol{\phi})\ , \tag{7}$$

where $I_\Phi(\boldsymbol{\phi})$ is 1 if $\boldsymbol{\phi} \in \Phi$, and 0 otherwise. Eqs (4) and (6) will be multiplied by the same indicator function.

3.3.1. The modified Langevin-Hastings algorithm

Several algorithms can be set up to make a proposal for $\boldsymbol{\phi}$ for a given $\tau^{(m)}$ in the Metropolis step of the MCMC algorithm. One strategy relies on a Bayesian version of the IWLS algorithm in generalized linear models (see e.g. Gamerman, 1997; Brezger and Lang, 2006). Here, we suggest to use a modified version of the Metropolis-adjusted Langevin algorithm (MALA, Roberts and Tweedie, 1996).

The basic MALA algorithm builds MCMC chains with proposals relying on the gradient of the log posterior distribution at the current state. More precisely, if $p(\boldsymbol{\theta}|\mathcal{D})$ is the posterior distribution (used as a shorthand notation for conditional posterior $p(\boldsymbol{\phi}|\tau^{(m-1)}, \mathcal{D})$) and $\boldsymbol{\theta}^{(m-1)} \in \mathrm{I\!R}^K$ the state of the chain at iteration $(m-1)$, then the proposal for $\boldsymbol{\theta}$ at iteration m is obtained by a random generation from the K-variate normal distribution

$$N_K(\boldsymbol{\theta}^{(m-1)} + 0.5\ \delta\ \nabla \log p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}), \delta\mathcal{I}_K)\ ,$$

where $\delta > 0$ and $\mathcal{I}_K$ is the K dimensional identity matrix. This proposal is accepted with probability

$$\alpha(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta}) = \min\left\{1, \frac{p(\boldsymbol{\theta}|\mathcal{D})}{p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D})} \frac{q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)})}{q(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta})}\right\}\ ,$$

where

$$q(\boldsymbol{x}, \boldsymbol{z}) = (2\pi\delta)^{-K/2}\ \exp\left[-\frac{1}{2\delta}\left\|\boldsymbol{z} - \boldsymbol{x} - 0.5\delta\ \nabla \log p(\boldsymbol{x}|\mathcal{D})\right\|^2\right]\ ,$$

i.e. $\boldsymbol{\theta}^{(m)}$ is set equal to $\boldsymbol{\theta}$ if accepted and to $\boldsymbol{\theta}^{(m-1)}$ otherwise.

Roberts and Rosenthal (1998) have shown that the relative efficiency of the algorithm can be characterized by its overall acceptance rate, independently of the target distribution. The asymptotic optimal value for that last quantity is 0.57 with acceptance probabilities in the range $(0.40, 0.80)$ still reasonable. The parameter δ must be tuned to have an acceptance rate in that range.

The mixing of the so-generated chain can be improved and the cross-correlation lowered with a slight modification of the proposal step by the introduction of a carefully chosen positive definite matrix Σ of size K. Then the proposal $\boldsymbol{\theta}$ for the next state is generated from

$$N_K(\boldsymbol{\theta}^{(m-1)} + 0.5\ \delta\Sigma\ \nabla \log p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}), \delta\Sigma)\ ,$$

with acceptance probability $\alpha(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta})$ where

$$\frac{q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)})}{q(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta})} = \exp\left\{-\frac{1}{2}\left(G + G^{(m-1)}\right)'\left((\boldsymbol{\theta} - \boldsymbol{\theta}^{(m-1)}) + \frac{\delta\Sigma}{4}(G - G^{(m-1)})\right)\right\}\ ,$$

with

$$G = \nabla \log p(\boldsymbol{\theta}|\mathcal{D}) \quad \text{and} \quad G^{(m-1)} = \nabla \log p(\boldsymbol{\theta}^{(m-1)}|\mathcal{D}) \ .$$

Note that this is equivalent to using the original Langevin-Hastings algorithm on a posterior reparametrized in terms of $\boldsymbol{\psi}$ with $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + L\boldsymbol{\psi}$ where $\boldsymbol{\theta}_0 \in \mathbb{R}^K$ and L denotes the lower triangular matrix resulting from the Cholesky decomposition of Σ.

Matrix Σ is ideally an approximation to the 2nd order dependence structure of the conditional posterior. It can obtained by an estimation of the B-splines parameters using, for example, a frequentist method for a fixed and reasonably chosen value of the roughness penalty parameter, see Section 3.1. Then, Eq. (3) can be used to define Σ and Eq. (2) to select $\boldsymbol{\theta}_0$.

3.3.2. Automatic tuning of the Langevin algorithm

An automatic tuning of δ targetting the optimal 0.57 rate can be achieved (Haario et al., 2001; Atchadé and Rosenthal, 2005): at the end of iteration m, set the value of δ to be used at iteration $(m+1)$ to δ_{m+1} where

$$\sqrt{\delta_{m+1}} = h\left(\sqrt{\delta_m} + \gamma_m \left(\alpha(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta}) - 0.57\right)\right) \ ,$$

with

$$h(x) = \begin{cases} \epsilon & \text{if} \quad x < \epsilon \\ x & \text{if} \quad x \in (\epsilon, A) \\ A & \text{if} \quad x > A \end{cases} \ ,$$

ϵ being a small number (say 10^{-4}) and A a large one (say $A = 10^4$). These two constants must be modified if the targeted acceptance rate is not attained. The series $\{\gamma_m\}$ is a non-increasing sequence of positive real numbers such that $|\gamma_m - \gamma_{m-1}| \leq m^{-1}$. A possible choice for γ_m is $\gamma_m = m^{-1}$.

In practice, we recommend to use the adaptive (modified) Langevin-Hastings algorithm for a few hundreds iteration with $\delta = 1.65^2/K^{1/3}$ as starting value for the tuning parameter. That value can be derived from the equations in Roberts and Rosenthal (1998, see Section 2) when the target posterior is the multivariate normal with identity variance-covariance matrix. The last value of δ_m in the so-generated sequence can then be used for the tuning parameter in the non-adaptive version of the modified Langevin-Hastings algorithm to produce the long chain(s) used for inference.

4. More general bin patterns

To simplify the presentation, it was assumed in Section 2 that the wide bins $\mathcal{J}_j$ do not overlap, and that their limits are a subset of the grid points.

However, the model has much wider application if the role of the initial grid is interpreted as a device for numerical integration of the underlying continuous density. Indeed, for given coefficients $\boldsymbol{\phi}$ associated to the B-splines basis, one can evaluate the density at the midpoints of the (fine) regular grid $\{x_i : i = 0, \ldots, I\}$ on (a, b) (with $x_0 = a$ and $x_I = b$). Then, setting $dx = (x_i - x_{i-1})$, one has

$$f_{\boldsymbol{\phi}}(x)dx = \frac{\exp\left(\sum_{k=1}^K \phi_k b_k(u_i)\right)}{\sum_{i=1}^I \exp\left(\sum_{k=1}^K \phi_k b_k(u_i)\right)} = \pi_i \quad \text{if} \quad x \in (x_{i-1}, x_i) \ , \tag{8}$$

where u_i denotes the midpoint of (x_{i-1}, x_i). At the limit, when $dx \rightarrow 0$, one obtains

$$f_{\boldsymbol{\phi}}(x) = \frac{\exp\left(\sum_{k=1}^{K} \phi_k b_k(u_i)\right)}{\int_a^b \exp\left(\sum_{k=1}^{K} \phi_k b_k(u)\right) du} \quad \text{if} \quad x \in (x_{i-1}, x_i) \ .$$

The probability γ_j that Y belongs to the jth wide bin $J_j = (L_j, U_j)$ can be obtained from Eq. (8) using numerical integration:

$$\gamma_j = \int_{L_j}^{U_j} f_{\boldsymbol{\phi}}(x) dx \approx \sum_{i=1} c_{ji} \pi_i \ ,$$

where the weights c_{ji} follow from, say, the rectangle method. When the wide bins boundaries are part of the regular grid, then the c_{ji}'s are simply the 0-1 quantities described in Section 2.

Wide bins may also overlap. In the limit each raw observation may have its own interval with an associated frequency, m_j, equal to one. This leads to a quite general procedure for density estimation from arbitrarily individually (interval) censored observations.

The chosen domain (a, b) spanning the fine grid should be chosen to respect the context of the data. In our examples (see below) we add a wide bin with a zero frequency at the right tail, because no raw observations were found there, although the possibility of a non zero count could not be excluded a priori. This forces the estimated density to go smoothly to zero. On the other hand, our example data concern chemical concentrations and human age, for which we know that they cannot be negative. It would not be right to add bins with zero counts below zero, because that would imply that we allow negative concentration (or age).

Of course our strategy will also work when there is essentially no grouping. In that case the bins and the grid will coincide and C becomes the identity matrix.

Hybrid situations may also occur, when only a part of the raw observations is censored. An interesting situation of this kind occurs in some types of genetic research, in which only individuals with extreme phenotypes (e.g. weight) are genotyped. There we have "central censoring": narrow bins in the tails, with one wide bin in the center of the distribution.

5. Simulation

A simulation study was performed to assess the performances of the Bayesian CLM for varying degrees of coarsening when the mean, the standard deviation and selected quantiles are estimated using the obtained fitted density. The data were simulated using a gamma distribution with mean 5 and variance $\sigma^2 = 2.5$. The compact interval (0,15.18) was taken as the support of the target gamma distribution: this is a reasonable approximation as it contains 99.999% of the probability mass. That interval was subdivided into $I = 48$ (small) bins of equal width $(= 0.2\sigma)$. Seven levels of coarsening corresponding to wide bins of respective width (`bw`) $0.4\sigma, 0.6\sigma, 0.8\sigma, 1.0\sigma, 1.2\sigma, 1.4\sigma, 1.6\sigma$ were considered by grouping consecutive small bins, larger widths corresponding to coarser data. $S = 500$ data sets of size $n = 200$ (and $1,000$) were simulated. For each data set, the posterior of the spline parameters in the Bayesian CLM was re-parameterized using the frequentist maximum likelihood estimate of these parameters with a roughness penalty parameter selected using BIC in a pre-defined grid. A first chain of length $M = 2,000$ (including a burn-in of 500) was built to derive a more reliable estimate of the dependence structure of the posterior. Then, a final chain of length $M = 3,500$ (following a burn-in of 500) was constructed to explore the posterior distribution. The fitted

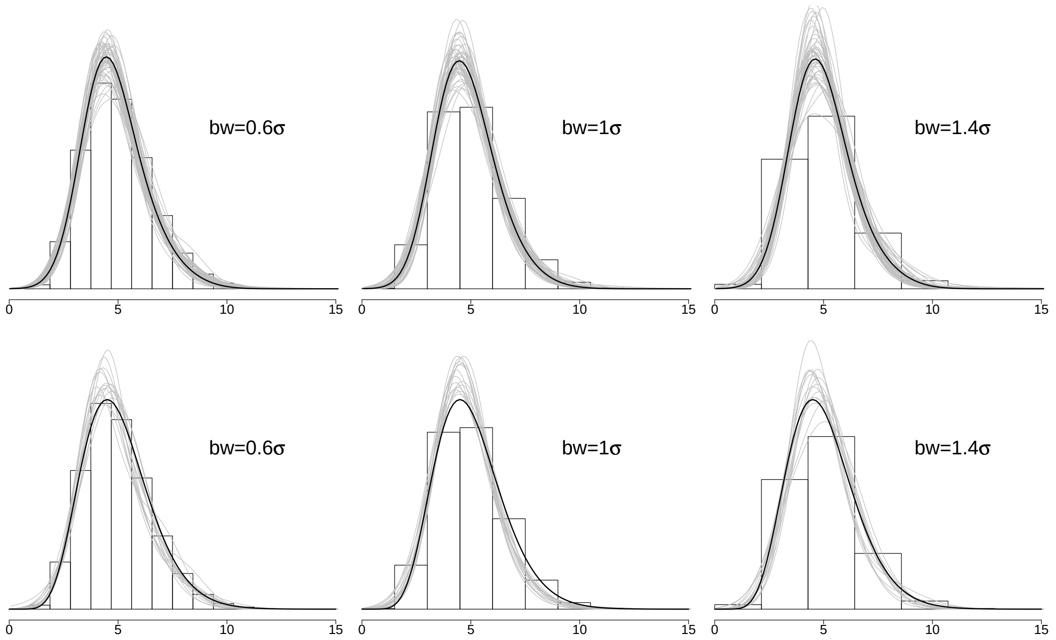

Figure 1: Row 1: Plot of 50 of the M=3,500 generated densities (grey curves) together with the fitted density $f^{(s)}$ (black curve) for one of the 500 simulated datasets of size $n = 200$. Row 2: Plot of 15 of the $S = 500$ fitted densities $f^{(s)}$ (grey curves) together with the target gamma density (black curve) of size $n = 200$.

density $f^{(s)}$ for the sth data set corresponds to the MCMC estimate $\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\phi}^{(m)}$ of the posterior mean of the spline parameters $\boldsymbol{\phi}$. It can be used to derive a point estimate for the mean, standard deviation and selected quantiles of the unknown density. The strategy is illustrated on Fig. 1.

The integrated squared error (ISE) defined by

$$\text{ISE} = \int \Big(f(x) - \hat{f}(x)\Big)^2 \ f(x)\ dx \ ,$$

was computed for each of the fitted densities $f^{(s)}$. The boxplots of the logarithm of this global measure of performance can be viewed on Fig. 2 for different wide bin widths. These can be compared to the values obtained with a frequentist approach where the penalty parameter is selected to minimize the AIC or the BIC.

For each sample size and provided that the wide bin width `bw` is at most 1.2σ, the Bayesian density estimate (BDE), followed by the frequentist density estimate (FDE) with τ selected using BIC (FDE-BIC), tends to perform better than the FDE with τ selected using AIC (FDE-AIC). For coarser data, the FDE-BIC is slightly better. The AIC criteria was often found to select smaller penalty parameters than the BIC and, hence, to under-smooth the observed histogram. Note that the same qualitative conclusions are obtained with the integrated L_1 norm.

The kernel density estimate (KDE: strategy -4- on Fig. 2) with a bandwidth selected using the Sheather and Jones (1991) method was also computed on the ungrouped data: it is clearly defeated by the BDE and the FDE-BIC computed from grouped data with wide bins of width 0.4σ.

The mean, the standard deviation and quantiles of the target density can be estimated by computing these quantities from the Bayesian density estimate. These posterior estimates for the mean and the standard deviation are plotted in Fig. 3 for different amounts of grouping together with the sample estimates computed from

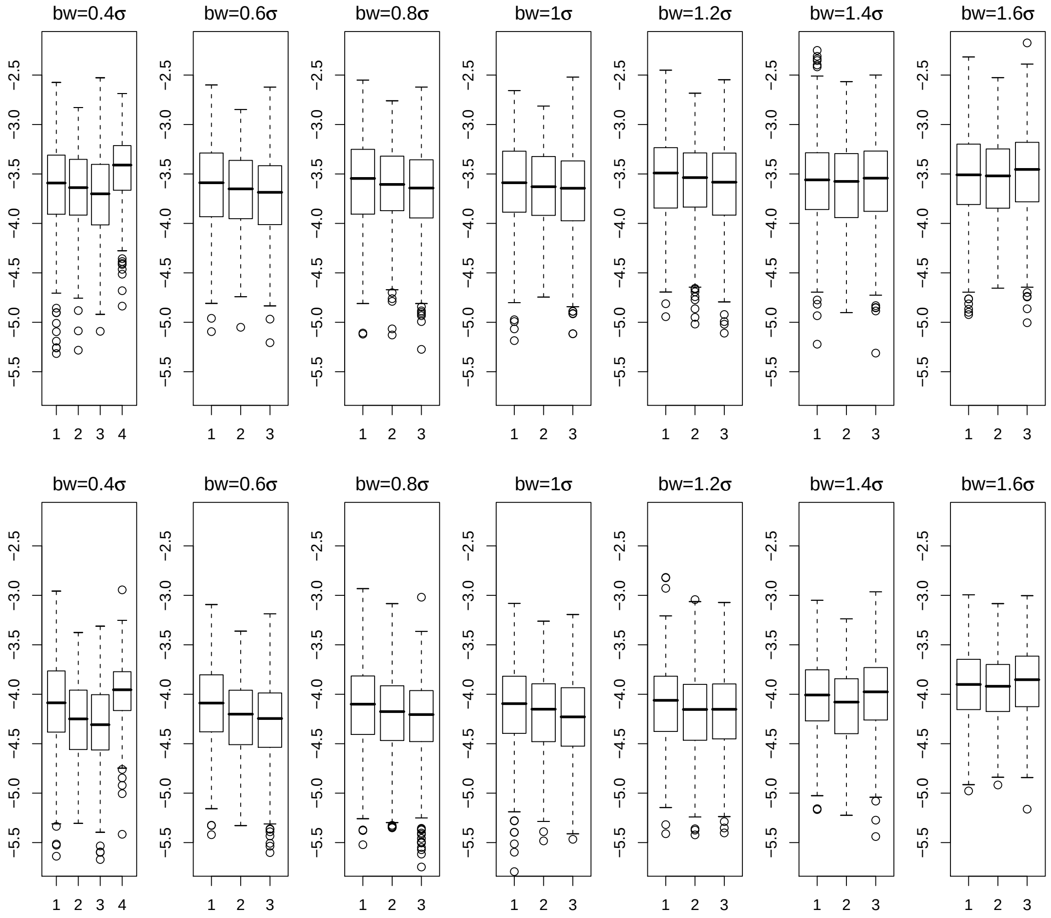

Figure 2: Boxplots of $\log_{10} \mathrm{ISE}(\hat{f})$ for different values of the (wide) bins width `bw` and different estimation strategies when $n = 200$ (Row 1) or $n = 1000$ (Row 2), $S = 500$ and the simulated density is a gamma with mean 5 and variance 2.5. The strategies are either frequentist (with τ selected using -1- AIC or -2- BIC) or -3- Bayesian. Strategy -4- corresponds to kernel density estimation (KDE) from the ungrouped data.

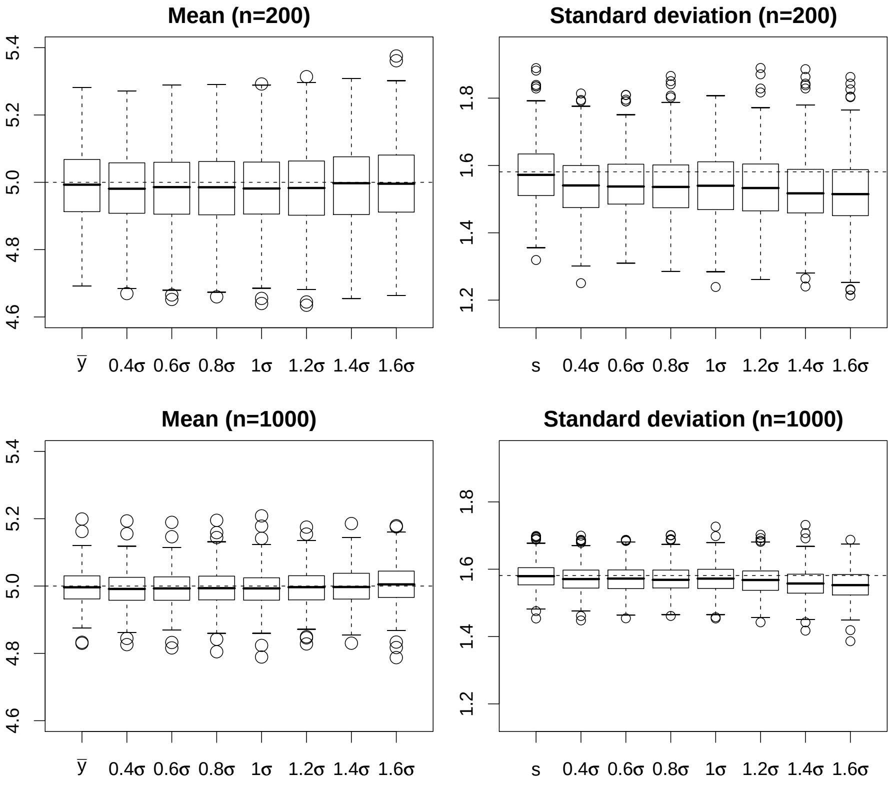

Figure 3: Boxplots of the posterior point estimate of the mean (left panel) or of the standard deviation (right panel) under different amounts of grouping (with the width of the wide bins varying from 0.4σ to 1.6σ) or directly using the sample mean $\bar{y}$ or standard deviation s computed on the ungrouped data; the horizontal dashed lines show the exact value for the mean and standard deviation. Row 1: $n = 200$, $S = 500$; Row 2: $n = 1000$, $S = 500$.

the un-grouped data. It shows that, whatever the sample size, grouping hardly affects the (negligible) bias and the sampling variance of the estimator for μ: these are comparable to the corresponding properties of the sample mean computed from the un-grouped data. When $n = 200$, the estimator of the standard deviation shows a non negligible bias ; when $n = 1,000$, the bias is moderate provided that `bw` does not exceed 1.2σ.

The root mean squared errors (RMSE, expressed as a percentage of the standard deviation) are reported for quantile estimates in Table 1. These RMSEs slowly increase with the amount of grouping with an acceleration when `bw` is larger than 1.2σ. The values corresponding to `bw` equal to 0.4σ can be compared to what one obtains when the quantiles are estimated from the kernel density estimate (KDE) or using the sample quantiles (SQ) computed from the un-grouped data. The Bayesian quantile estimate is, most of the time, the best performer: its superiority over KDE is particularly marked in the lower tail, while it is systematically better than SQ. The preference for the Bayesian estimate over SQ (computed from the un-grouped data) persists provided that `bw` does not exceed 1.2σ, a remarkable result. This is probably attributable to the joint estimation of quantiles in our approach.

RMSEs can be compared (see Table 2) to what one obtains for these when a gamma distribution is (correctly) assumed for the unknown density with parameters estimated using the EM algorithm (see Section 1 for some more details). Not surprisingly, the (nonparametric) Bayesian method is less efficient than the para-

	Quantiles ($n = 200$)										
`bw`	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
1.6σ	10.2	8.8	8.5	8.5	8.3	8.1	8.1	8.5	9.8	13.7	19.2
1.4σ	9.6	8.1	7.5	7.3	7.3	7.5	7.8	8.3	9.7	13.5	18.9
1.2σ	8.7	7.1	6.9	7.0	7.3	7.6	7.9	8.4	9.6	13.1	18.1
1.0σ	8.9	6.9	6.4	6.5	6.9	7.3	7.7	8.4	9.7	13.3	18.2
0.8σ	8.1	6.5	6.4	6.8	7.0	7.3	7.7	8.3	9.4	12.7	17.5
0.6σ	7.8	6.3	6.1	6.4	6.6	6.9	7.2	7.9	9.1	12.4	16.8
0.4σ	7.5	6.1	6.0	6.3	6.5	6.8	7.2	7.9	9.1	12.2	16.7
KDE	13.2	10.1	7.3	6.4	6.3	6.9	7.8	8.9	10.3	12.6	16.0
SQ	8.6	7.7	7.1	6.8	7.1	7.8	8.4	9.4	10.7	13.0	16.9

	Quantiles ($n = 1000$)										
J	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
1.6σ	4.5	4.7	5.1	4.8	4.2	3.7	3.6	4.0	5.0	6.7	8.3
1.4σ	4.5	4.5	4.1	3.6	3.3	3.4	3.7	4.2	4.8	5.9	7.7
1.2σ	3.9	3.5	3.3	3.2	3.3	3.4	3.6	4.0	4.6	6.0	7.6
1.0σ	3.7	3.2	3.1	3.2	3.2	3.3	3.5	3.8	4.4	5.8	7.5
0.8σ	3.5	3.1	3.2	3.1	3.2	3.2	3.4	3.7	4.3	5.7	7.3
0.6σ	3.3	2.9	3.0	3.1	3.1	3.2	3.4	3.7	4.3	5.6	7.2
0.4σ	3.1	2.8	2.9	3.0	3.0	3.2	3.3	3.6	4.2	5.5	7.2
KDE	7.0	5.2	3.7	3.1	3.0	3.2	3.6	3.9	4.6	6.1	7.7
SQ	3.7	3.4	3.4	3.1	3.3	3.5	3.8	4.0	4.7	6.3	8.4

Table 1: $100 \times \sqrt{\text{MSE}}/\sigma$ for selected quantiles when estimation is performed under different amounts of grouping (with the width `bw` of the wide bins varying from 0.4σ to 1.6σ) or directly using the sample quantiles (SQ) computed on the ungrouped data ; $n = 200$ or 1000 and $S = 500$.

	Quantiles ($n = 200$)									
bw	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
1.6σ	1.23	1.03	1.05	1.11	1.13	1.11	1.08	1.05	1.03	1.09	1.22
1.4σ	1.21	1.01	1.03	1.08	1.09	1.08	1.05	1.02	1.00	1.07	1.19
1.2σ	1.15	0.97	1.01	1.09	1.12	1.12	1.09	1.05	1.02	1.07	1.20
1.0σ	1.21	0.99	0.98	1.05	1.10	1.11	1.08	1.03	1.00	1.05	1.17
0.8σ	1.12	0.96	0.99	1.04	1.09	1.10	1.07	1.02	0.99	1.06	1.19
0.6σ	1.19	1.05	1.04	1.05	1.06	1.06	1.03	1.00	1.00	1.09	1.25
0.4σ	1.11	1.01	1.06	1.10	1.11	1.08	1.04	1.00	0.99	1.09	1.23

	Quantiles ($n = 1000$)										
bw	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
1.6σ	1.12	1.21	1.42	1.39	1.23	1.07	1.00	1.03	1.10	1.16	1.17
1.4σ	1.15	1.20	1.18	1.07	1.01	1.03	1.08	1.12	1.09	1.05	1.10
1.2σ	1.07	0.99	0.99	1.01	1.04	1.07	1.08	1.08	1.09	1.11	1.14
1.0σ	1.11	0.98	1.03	1.07	1.09	1.10	1.08	1.06	1.06	1.10	1.18
0.8σ	1.08	1.00	1.07	1.09	1.09	1.09	1.09	1.09	1.10	1.14	1.20
0.6σ	1.07	1.00	1.08	1.11	1.12	1.12	1.11	1.10	1.09	1.12	1.20
0.4σ	1.08	1.00	1.08	1.11	1.13	1.12	1.11	1.09	1.08	1.14	1.22

Table 2: $\sqrt{\text{MSE}_1}/\sqrt{\text{MSE}_2}$ for selected quantiles when estimation is performed under different amounts of grouping (with the width bw of the wide bins varying from 0.4σ to 1.6σ) where subscript '1' refers to the Bayesian method and '2' to an estimation performed under the correct assumption of a gamma distribution with parameters estimated using the EM algorithm; $n = 200$ or 1000 and $S = 500$.

metric EM when the right parametric guess is made for the density ; this is more marked in the tails.

The coverages of the 80% and 90% credible intervals were also investigated for selected quantiles, see Tables 3 and 4. For the sth data set, the $(1-\alpha)\times 100\%$ credible interval of a $\beta-$quantile can be estimated using the $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of the set of $\beta-$quantiles associated to each of the $M = 3,500$ generated densities. The reported coverages correspond to the proportions of the so-defined $S = 500$ credible intervals containing the corresponding quantile of the target gamma density. The bold values in Tables 3 and 4 indicate where coverages are significantly different from their large sample frequentist nominal values. One can see that values close to the nominal levels are obtained provided that bw does not exceed 1.2σ.

6. Illustrations

6.1. Air pollution

Hasselblad et al. (1980) studied interval censored concentrations of lead in the blood of New-Yorkers in the 1970-1976 period. The dataset of interest concerns young Puerto Ricans aged 1-12 years in 1974:

	Lead concentration (in μg/dl)						
Wide interval	(0,15)	(15,25)	(25,35)	(35,45)	(45,55)	(55,65)	65+
Freq. m_j	27	71	32	6	3	0	0

The data were recorded in broad intervals for screening purposes. A blood lead level over 30 μg/dl is sometimes described as an unacceptable medical risk. Therefore, one quantity of scientific interest is the proportion of persons with a blood lead level

		Quantiles ($n = 200$)						
`bw`	Credib.	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1.6σ	80%	78	**75**	**74**	**75**	76	78	79
	90%	89	87	**85**	**85**	87	89	89
1.4σ	80%	81	78	78	78	80	81	80
	90%	91	88	87	88	89	91	90
1.2σ	80%	82	77	77	77	76	78	78
	90%	91	88	88	87	88	89	89
1.0σ	80%	84	82	79	78	77	76	**74**
	90%	93	92	90	88	88	88	87
0.8σ	80%	82	79	76	77	78	76	**75**
	90%	91	88	87	88	88	88	89
0.6σ	80%	83	79	78	78	78	80	77
	90%	92	90	89	89	89	89	90
0.4σ	80%	84	79	77	79	80	78	77
	90%	92	89	89	89	89	89	89

Table 3: Estimated coverages of 80% and 90% credible intervals for selected quantiles when estimation is performed under different amounts of grouping (with the width `bw` of the wide bins varying from 0.4σ to 1.6σ) ; $n = 200$ and $S = 500$.

		Quantiles ($n = 1000$)						
`bw`	Credib.	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1.6σ	80%	**65**	**65**	**70**	**72**	79	80	78
	90%	**82**	**81**	**82**	**85**	87	89	90
1.4σ	80%	**71**	**72**	76	79	81	81	79
	90%	**84**	**84**	87	89	91	91	90
1.2σ	80%	77	78	81	83	82	79	80
	90%	87	89	89	89	91	90	89
1.0σ	80%	80	78	78	78	79	81	79
	90%	89	88	88	87	88	89	90
0.8σ	80%	77	77	77	78	80	79	80
	90%	87	87	88	89	90	90	89
0.6σ	80%	77	78	78	78	78	79	79
	90%	89	88	88	89	90	90	90
0.4σ	80%	79	77	79	80	79	79	81
	90%	88	87	88	89	91	91	89

Table 4: Estimated coverages of 80% and 90% credible intervals for selected quantiles when estimation is performed under different amounts of grouping (with the width `bw` of the wide bins varying from 0.4σ to 1.6σ) ; $n = 1000$ and $S = 500$.

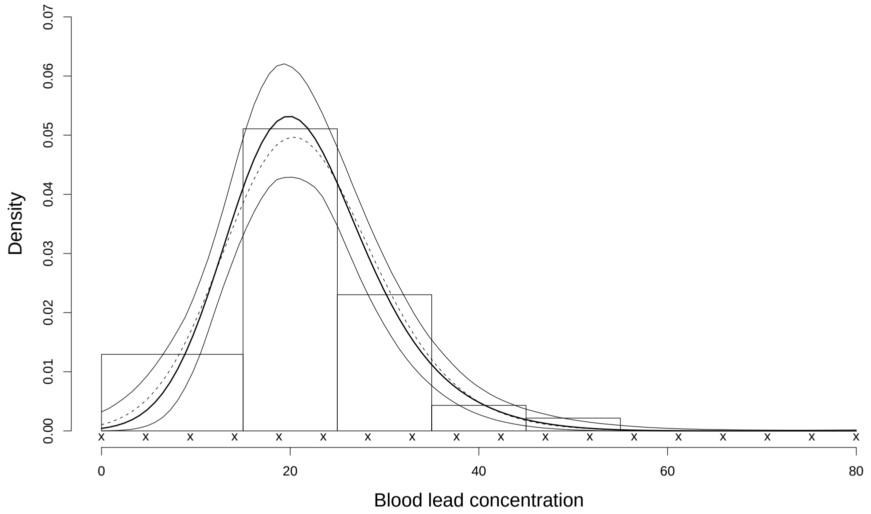

Figure 4: Histogram of the grouped data of lead concentration (in $\mu g/100ml$) in the blood of young Puerto Ricans in New-York in 1974. Frequentist (dashed line) and Bayesian (thick solid line) estimates of the latent density with knots positioned at the crosses and a 3rd order penalty. The thin solid lines delimit the (Bayesian) 90% pointwise credible intervals for the density.

over that threshold. We shall use our strategy for density estimation to evaluate that proportion.

The last wide bin, 65+, was arbitrarily replaced by $(65, 80)$. Larger bins were also tried: these were not found to affect the final results. A grid of narrow intervals of width $\Delta = 1$ on the domain $(0, 80)$ was considered. Twenty ($= K$) cubic B-splines associated to equi-spaced knots on $(0, 80)$ were used in the basis.

A frequentist estimate of the density f of the blood lead level can be derived using Eq. (2) with τ selected in a grid to minimize the BIC: this yields the dashed curve on Figure 4. The Bayesian estimate for the latent density (thick solid line) is also shown, with the 90% pointwise credible interval computed from a chain of length $M = 10,000$. In the Bayesian framework, unimodality was forced using Eq. (7) by giving a zero prior to spline parameters $\boldsymbol{\phi}$ not yielding a unimodal density $f_{\boldsymbol{\phi}}$.

Point estimates and 90% credible intervals for $\Pr(Y > 30)$, the mean, the standard deviation and two quantiles were also derived:

				Quantile	
	$\Pr(Y > 30)$	μ_Y	σ_Y	0.20	0.80
Posterior mean	0.14	21.8	8.3	14.6	27.8
90% cred. intervals	(0.10,0.19)	(20.6,23.0)	(7.3,9.6)	(13.1, 15.9)	(26.1,29.5)

This can be done for any function of the density.

6.2. Mortality table

Our second example comes from an (unpublished) study on historical trends of Tuberculosis mortality in The Netherlands. Yearly data, starting in 1900, are available on numbers of deaths, classified according to the IDC (International Disease Classification). Unfortunately, in the early years, very wide age groups have been used. In contrast, yearly population numbers are available in one-year age intervals. So, if we estimate density on an age grid with a spacing of one year, we can construct a detailed mortality table.

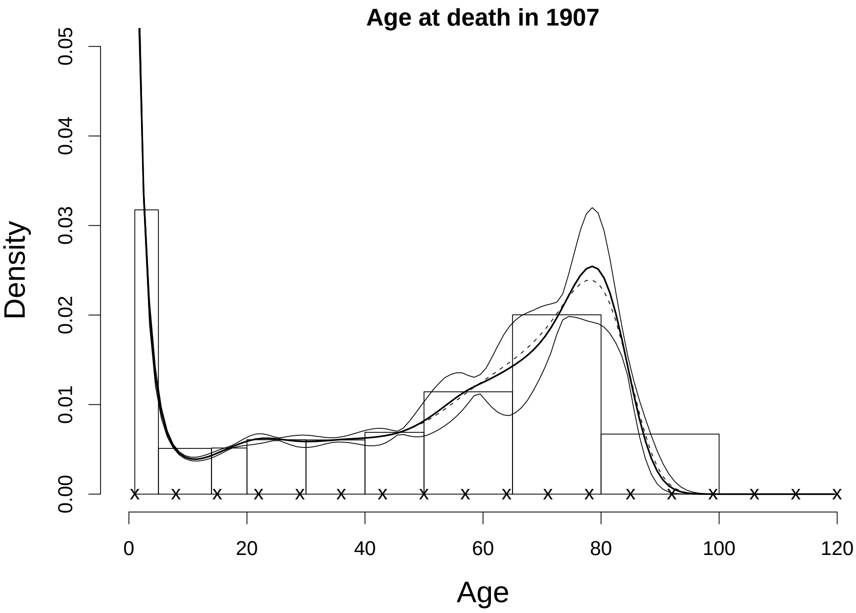

Figure 5: Histogram of age at death (for all-cause deaths) in The Netherlands in 1907. Frequentist (dashed line) and Bayesian (thick solid line) estimates of the latent density with knots positioned at the crosses and a 3rd order penalty. The thin solid lines delimit the (Bayesian) 90% pointwise credible intervals for the density.

Figure 5 shows the histogram of age at death (for all-cause deaths) in The Netherlands in 1907. Over 128,000 people died in 1907. The available data stopped at age 99. An extra interval from 100 to 119 was added, with zero count. Figure 6 shows results for counts of deaths by TB in the same year. The total number is 9,440.

In each case, a basis with 20 cubic splines was used with a difference penalty of the third order. The sampler was run for 30,000 iterations, which took 3 seconds on a standard model 2007 MacBook computer. The Langevin-Hastings sampler was programmed in C to enhance speed.

The fitted density is given on both figures by the thick solid lines: it is smooth as expected. The credible envelope contains, at each age, 90% of the M=30,000 densities generated by MCMC: these are plausible densities given the limited information provided by the grouped data. The shape of these envelopes shows that some of the generated densities are somewhat wiggly within a wide bin. Indeed, given the very large frequencies, the area under one of these densities over a given wide bin should be very close to the proportion of deaths corresponding to that bin (as given by the area of the associated rectangle). Therefore, a density that is much larger (smaller) than the rectangle height at the beginning of the wide bin should take a smaller (larger) value by the end of the interval. The fitted density and the width of the credible envelope are not sensitive to the number of bins and knots (provided that this number is large enough).

Using Bayes theorem and these two densities, one can estimate the probability that a death at a given age is due to TB. Indeed, one has

$$\mathrm{Pr(TB\ death|Age = k\ \&\ died)} = \frac{\mathrm{Pr(Age = k\ |TB\ death)}}{\mathrm{Pr(Age = k\ |died)}}\ \mathrm{Pr(TB\ death|died)}$$

An estimate of the numerator (resp. denominator) can be derived from the first (resp. second) estimated density, while the last factor in the preceding equation is

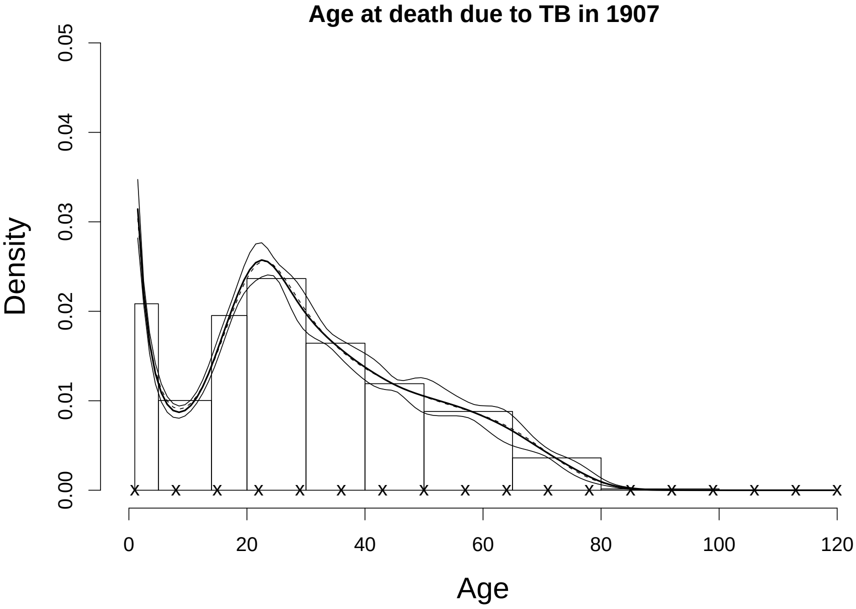

Figure 6: Histogram of the grouped counts of deaths by Tuberculosis in The Netherlands in 1907. Frequentist (dashed line) and Bayesian (thick solid line) estimates of the latent density with knots positioned at the crosses and a 3rd order penalty. The thin solid lines delimit the (Bayesian) 90% pointwise credible intervals for the density.

simply the proportion of deaths due to TB (in 1907). The so-obtained estimate is given by the thick solid line on Fig. 7. A 90% credible envelope can be obtained from the repeated computation of the preceding expression for each of the M (pairs of) densities generated by MCMC: it provides a set of plausible values for the probability of interest at a given age.

7. Discussion

Grouped data occur in many places. There is a need for good tools for density estimation on real data. The combination of P-splines and the composite link model provides such a tool. We have documented its performance with extensive simulations and illustrated it on real data. We provide two variants, one based on penalized likelihood, the other fully Bayesian. The latter takes more computing time (but just a few seconds on a modern computer), but it delivers uncertainty estimates of all sorts of derived statistics, like moments and quantiles.

The prior distribution on the spline parameters can be used to enforce (or only to encourage) desired shape properties of the density like monotonicity, unimodality or log-concavity. A 3rd order penalty was used during the simulations and in the illustrations as it yields a normal density at the limit for large values of the penalty.

We have advocated, following the arguments in Eilers and Marx (1996), to put many equidistant knots on the compact interval thought to contain most of the probability mass. In specific circumstances such as a positive random variable Y with a peaked long-tailed distribution, this is not optimal. One needs light smoothing near the peak and strong smoothing in the tails, which is impossible to obtain with one smoothing parameter. In such cases it is advisable to first transform the data (e.g., taking the logarithm or the square root) or to use adaptive penalties (Jullion and Lambert, 2007).

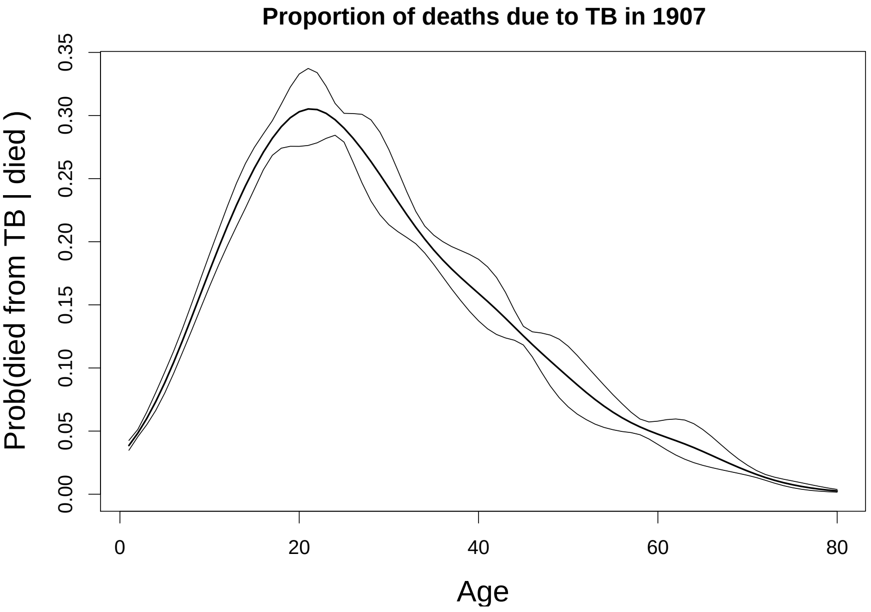

Figure 7: Probability of death due to Tuberculosis at a given age in The Netherlands in 1907 (thick solid line). The thin solid lines delimit the (Bayesian) 90% pointwise credible intervals for these probabilities.

Several useful generalizations are possible. Extension to two dimensions is one of them. Intervals become rectangles and we use tensor product of P-splines to model the logarithm of the smooth density. Expected counts are given by integrals of this density over the rectangles. Our initial results look very promising; we shall report on this elsewhere.

All computation were done with R. A R package will soon be available from the first author.

Acknowledgements

Financial support from the FSR research grant nr FSRC-08/42 from the University of Liège is gratefully acknowledged.

References

Atchadé, Y. F., Rosenthal, J. S., 2005. On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11, 815–828.

Berry, S. M., Carroll, R. J., Ruppert, D., 2002. Bayesian smoothing and regression splines for measurement error problems. Journal of the American Statistical Association 97, 160–169.

Boneva, L. I., Kendall, D. G., Stefano, I., 1971. Spline transformations. three new diagnostic aids for the statistical data-analyst. Journal of the Royal Statistical Society, Series B 33, 1–71.

Braun, J., Duchesne, T., Stafford, J. E., 2005. Local likelihood density estimation for interval censored data. The Canadian Journal of Statistics 33, 39–60.

Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis 50, 967–991.

Eilers, P. H. C., 2007. Ill-posed problems with counts, the composite link model, and penalized likelihood. Statistical Modelling 7, 239–254.

Eilers, P. H. C., Marx, B. D., 1996. Flexible smoothing with B-splines and penalties (with discussion). Statistical Science 11, 89–121.

Gamerman, D., 1997. Efficient sampling from the posterior distribution in generalized linear models. Statist. Comput. 7, 57–68.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli 7, 223–242.

Hasselblad, V., Stead, A. G., Galke, W., 1980. Analysis of coarsely grouped data from the lognormal distribution. Journal of the American Statistical Association 75, 771–778.

Jullion, A., Lambert, P., 2007. Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian P-splines models. Computational Statistics and Data Analysis 51, 2542–2558.

Lambert, P., 2007. Archimedean copula estimation using Bayesian splines smoothing techniques. Computational Statistics and Data Analysis 51, 6307–6320.

Lambert, P., Eilers, P. H., 2005. Bayesian proportional hazards model with time varying regression coefficients: a penalized Poisson regression approach. Statistics in Medicine 24, 3977–3989.

Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics 13, 183–212.

Roberts, G. O., Rosenthal, J. S., 1998. Optimal scaling of discrete approximations to Langevin diffusions. Journal of the Royal Statistical Society, Series B 60, 225–268.

Roberts, G. O., Tweedie, R. L., 1996. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 24, 341–363.

Sheather, S. J., Jones, M. C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B 53, 683–690.

Thompson, R., Baker, R. J., 1981. Composite link functions in generalized linear models. Applied Statistics 30, 125–131.