Li, Feng; Villani, Mattias; Kohn, Robert

Working Paper

Modeling conditional densities using finite smooth mixtures

Sveriges Riksbank Working Paper Series, No. 245

Provided in Cooperation with:
Central Bank of Sweden, Stockholm

SVERIGES RIKSBANK
WORKING PAPER SERIES 245

Modeling Conditional Densities Using Finite Smooth Mixtures

Feng Li, Mattias Villani and Robert Kohn

AUGUST 2010

WORKING PAPERS ARE OBTAINABLE FROM

Sveriges Riksbank • Information Riksbank • SE-103 37 Stockholm
Fax international: +46 8 787 05 26
Telephone international: +46 8 787 01 00
E-mail: info@riksbank.se

The Working Paper series presents reports on matters in
the sphere of activities of the Riksbank that are considered
to be of interest to a wider public.
The papers are to be regarded as reports on ongoing studies
and the authors will be pleased to receive comments.

MODELING CONDITIONAL DENSITIES USING FINITE SMOOTH MIXTURES

FENG LI, MATTIAS VILLANI, AND ROBERT KOHN

Sveriges Riksbank Working Paper Series No. 245

August 2010

Abstract. Smooth mixtures, i.e. mixture models with covariate-dependent mixing weights, are very useful flexible models for conditional densities. Previous work shows that using too simple mixture components for modeling heteroscedastic and/or heavy tailed data can give a poor fit, even with a large number of components. This paper explores how well a smooth mixture of symmetric components can capture skewed data. Simulations and applications on real data show that including covariate-dependent skewness in the components can lead to substantially improved performance on skewed data, often using a much smaller number of components. Furthermore, variable selection is effective in removing unnecessary covariates in the skewness, which means that there is little loss in allowing for skewness in the components when the data are actually symmetric. We also introduce smooth mixtures of gamma and log-normal components to model positively-valued response variables.
Keywords: Bayesian inference, Markov chain Monte Carlo, Mixture of Experts, Variable selection

Li: Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. Villani: Research Division, Sveriges Riksbank and Department of Statistics, Stockholm University. Kohn: Australian School of Business, University of New South Wales, UNSW, Sydney 2052, Australia. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

1. Introduction

Finite smooth mixtures, or *mixtures of experts* (ME) as they are known in the machine learning literature, are increasingly popular in the statistical literature since their introduction in Jacobs et al. (1991). A smooth mixture is a mixture of regression models where the mixing probabilities are functions of the covariates, leading to a partitioned covariate space with stochastic (soft) boundaries. The first applications of smooth mixtures focused on flexible modeling of the mean function $\mathrm{E}(y|x)$, but more recent works explore their potential for nonparametric modeling of conditional densities $p(y|x)$. A smooth mixture models $p(y|x)$ non-parametrically for any given x, but is also flexible across different covariate values.

Smooth mixtures are capable of approximating a large class of conditional distributions. For example, Jiang and Tanner (1999a,b) show that smooth mixtures with sufficiently many (generalized) linear regression mixture components can approximate any density in the exponential family with arbitrary smooth mean function. More recently, Norets (2010) proves results for a mixture of Gaussian components under fairly general regularity conditions. See also Zeevi and Meir (1997) for additional results along these lines.

Like any mixture model, a smooth mixture may have a fairly complex multimodal likelihood surface. The choice of estimation method is therefore a key ingredient for successfully implementing smooth mixture models. Jordan and Jacobs (1994) employ the expectation maximization (EM) algorithm for the ME model, and similar optimization algorithms are popular in the machine learning field. Some recent approaches to smooth mixtures are Bayesian, with the computation implemented by Markov Chain Monte Carlo (MCMC) methods. The first Bayesian paper on smooth mixtures is Peng et al. (1996) who use the random walk Metropolis algorithm to sample from the posterior. More sophisticated algorithms are proposed by Wood et al. (2002), Geweke and Keane (2007) and Villani et al. (2009).

The initial work on smooth mixtures in the machine learning literature advocated what may be called a *simple-and-many* approach with very simple mixture components (constants or linear homoscedastic regressions), but many of them. This practice is partly because estimating complicated component models was somewhat difficult in the pre and early days of MCMC, but probably also reflects an underlying divide-and-conquer philosophy in the machine learning literature. More recent implementations of smooth mixtures with access to MCMC technology successively introduce more flexibility within the components. This *complex-and-few* strategy tries to model nonlinearities and non-Gaussian features within the components and relies less on the mixture to generate the required flexibility, i.e. mixtures are used only when needed. For example, Wood et al. (2002) and Geweke and Keane (2007) use basis expansion methods (splines and polynomials) to allow for nonparametric component regressions. Further progress is made in Villani et al. (2009) who propose the Smooth Adaptive Gaussian Mixture (SAGM) model as a flexible model for regression density estimation. Their model is a finite mixture of Gaussian densities with the mixing probabilities, the component means and component variances modeled as (spline) functions of the covariates. Li et al. (2010) extend this model to asymmetric student's t components with the location, scale, skewness and degrees of freedom all modeled as functions of covariates. Villani et al. (2009) and Li et al. (2010) show that a single complex component can often give a better and numerically more stable fit in substantially less computing time than a model with many simpler components. As an example, simulations and real applications in Villani et al. (2009) show that a mixture of homoscedastic regressions can fail to fit heteroscedastic data even

with a very large number of components. Having heteroscedastic components in the mixture is therefore crucial for accurately modeling heteroscedastic data. The empirical stock returns example in Li et al. (2010) shows that including heavy-tailed components in the mixture can improve on the SAGM model when modeling heteroscedastic heavy-tailed distributions. This finding is backed up by the theoretical results in Norets (2010).

This chapter further explores the simple-and-many vs complex-and-few issue by modeling regression data with a skewed response variable. A simulation study shows that it may be difficult to model a skewed conditional density by a smooth mixture of heteroscedastic Gaussian components (like SAGM). Introducing skewness within the components can improve the fit substantially.

We use the efficient Markov chain Monte Carlo (MCMC) method in Villani et al. (2009) to simulate draws from the posterior distribution in smooth mixture models; see Section 3.1. This algorithm allows for Bayesian variable selection in all parameters of the density, and in the mixture weights. Variable selection mitigates problems with over-fitting, which is particularly important in models with complex mixture components. The automatic pruning effect achieved by variable selection in a mixture context is illustrated in Section 4.2 on the LIDAR data. Reducing the number of effective parameters by variable selection also helps the MCMC algorithm to converge faster and mix better.

Section 4.3 uses smooth mixtures of Gaussians and split-t components to model the electricity expenditure of households. To take into account that expenditures are positive, and more generally to handle positive dependent variables, we also introduce two smooth mixtures for strictly positively valued data: a smooth mixture of gamma densities and smooth mixture of log normal densities. In both cases we use an interpretable re-parametrized density where the mean and the (log) variance are modeled as functions of the covariates.

2. The model and prior

2.1. **Smooth mixtures.** Our model for the conditional density $p(y|x)$ is a finite mixture density with weights that are smooth functions of the covariates,

$$p(y|x) = \sum_{k=1}^{K} \omega_k(x) p_k(y|x), \tag{1}$$

where $p_k(y|x)$ is the kth component density with weight $\omega_k(x)$. The next subsection discusses specific component densities $p_k(y|x)$. The weights are modeled by a multinomial logit function

$$\omega_k(x) = \frac{\exp(x'\gamma_k)}{\sum_{r=1}^{K} \exp(x'\gamma_r)}, \tag{2}$$

with $\gamma_1 = 0$ for identification. The covariates in the components can in general be different from the covariates in the mixture weights.

To simplify the MCMC simulation, we express the mixture model in terms of latent variables as in Diebolt and Robert (1994) and Escobar and West (1995). Let $s_1, ..., s_n$ be unobserved indicator variables for the observations in the sample such that $s_i = k$ means that the ith observation belongs to the kth component, $p_k(y|x)$. The model in (1) and (2) can then be written as

$$\Pr(s_i = k|x_i, \gamma) = \omega_k(x_i)$$
$$y_i|(s_i = k, x_i) \sim p_k(y_i|x_i).$$

Conditional on $s = (s_1, ..., s_n)'$, the mixture model decomposes into K separate component models $p_1(y|x), ..., p_K(y|x)$, with each data observation being allocated to one and only one component.

2.2. **The component models.** The component densities in SAGM (Villani et al.; 2009) are Gaussian with both the mean and variance functions of covariates,

$$y|x, s = k \sim \mathrm{N}\left[\mu_k(x), \sigma_k^2(x)\right],$$

where

$$\mu_k(x) = \beta_{\mu_0,k} + x'\beta_{\mu,k} \quad \ln \sigma_k^2(x) = \beta_{\sigma_0,k} + x'\beta_{\sigma,k} \tag{3}$$

Note that each mixture components has its own set of parameters. We will suppress the component subscript k in the remainder of this section, but, unless stated otherwise, all parameters are component-specific. SAGM uses a linear link function for the mean and log link for the variance, but any smooth link function can equally well be used in our MCMC methodology. Additional flexibility can be obtained by letting a subset of the covariates be a non-linear basis expansions, e.g. additive splines or splines surfaces (Ruppert et al.; 2003) as in Villani et al. (2009); see also the LIDAR example in Section 4.2.

SAGM is in principle capable of capturing heavy-tailed and skewed data. In line with the complex-and-few approach it may be better however to use mixture components that allow for skewness and excess kurtosis. Li et al. (2010) extend the SAGM model to components that are split-t densities according to the following definition.

Definition 1 (Split-t distribution)**.** *The random variable y follows a* split-t distribution *with $\nu > 0$ degrees of freedom, if its density function is of the form*

$$p\left(y; \mu, \phi, \lambda, \nu\right) = c \cdot \kappa(y; \mu, \phi, \nu)\mathbf{1}_{y \leq \mu} + c \cdot \kappa(y; \mu, \lambda\phi, \nu)\mathbf{1}_{y > \mu},$$

where

$$\kappa(y; \mu, \phi, \nu) = \left[1 + \left(\frac{y - \nu}{\phi}\right)^2 \nu^{-1}\right]^{-\frac{\nu+1}{2}},$$

is the kernel of a student's t *density with variance $\phi^2\nu/(\nu-2)$ and $c = 2[(1+\lambda)\phi\sqrt{\nu}Beta(\nu/2, 1/2)]^{-1}$ is the normalization constant.*

The location parameter μ is the mode, $\phi > 0$ is the scale parameter, and $\lambda > 0$ is the skewness parameter. When $\lambda < 1$ the distribution is skewed to the left, when $\lambda > 1$ it is skewed to the right, and when $\lambda = 1$ it reduces to the usual symmetric student's t density. The split-t distribution reduces to the split-normal distribution in Gibbons and Mylroie (1973) and John (1982) as $\nu \to \infty$. Any other asymmetric t density can equally well be used in our MCMC methodology, see Section 3.1.

Each of the four parameters μ, ϕ, λ and ν are connected to covariates as

$$\begin{array}{ll} \mu = \beta_{\mu_0} + x'\beta_\mu, & \ln \phi = \beta_{\phi_0} + x'\beta_\phi, \\ \ln \nu = \beta_{\nu_0} + x'\beta_\nu, & \ln \lambda = \beta_{\lambda_0} + x'\beta_\lambda, \end{array} \tag{4}$$

but, as mentioned above, any smooth link function can equally well be used in the MCMC methodology.

Section 4.3 applies smooth mixtures in a situation where the response is non-negative. Natural mixture components are then Gamma and log-normal densities. The Gamma components are of the form

$$y|s,x \sim \text{Gamma}\left(\frac{\mu^2}{\sigma^2}, \frac{\sigma^2}{\mu}\right),$$

where

$$\ln \mu(x) = \beta_{\mu_0} + x'\beta_\mu \quad \ln \sigma^2(x) = \beta_{\sigma_0} + x'\beta_\sigma \ , \tag{5}$$

where we have again suppressed the component labels. Note that we use an interpretable parametrization of the Gamma distribution where μ and σ^2 are the mean and variance, respectively.

Similarly, the log-normal components are of the form

$$y|s,x \sim \text{LogN}\left(\ln \mu - \frac{1}{2}\ln\left(1+\frac{\sigma^2}{\mu^2}\right), \sqrt{\ln\left(1+\frac{\sigma^2}{\mu^2}\right)}\right),$$

where

$$\ln \mu(x) = \beta_{\mu_0} + x'\beta_\mu, \quad \ln \sigma^2(x) = \beta_{\sigma_0} + x'\beta_\sigma \ . \tag{6}$$

Again, the two parameters, μ and σ^2, are the mean and variance.

A smooth mixture of complex densities is a model with a large number of parameters, however, and is therefore likely to over-fit the data unless model complexity is controlled effectively. We use Bayesian variable selection on all the component's parameters, and in the mixing function. This can lead to important simplifications of the mixture components. Not only does this control complexity for a given number of components, but it also simplifies the existing components if an additional component is added to the model (the LIDAR example in 4.2 illustrates this well). Increasing the number of components can therefore in principle even reduce the number of effective parameters in the model. It may nevertheless be useful to put additional structure on the mixture components before estimation. One particularly important restriction is that one or more component parameters are common to all components. A component parameter (e.g. ν in the split-t model in 4) is said to be *common* to the components when only the intercepts in (4) are allowed to be different across components. The unrestricted model is said to have *separate* components.

The regression coefficient vectors, e.g. β_μ, β_ϕ, β_ν and β_λ in the split-t model, are all treated in a unified way in the MCMC algorithm. Whenever we refer to a regression coefficient vector without subscript, β, the argument applies to any of the regression coefficient vector of the split-t parameters in (4).

2.3. **The prior.** We now describe an easily specified prior for smooth mixtures, proposed by Villani et al. (2010) that builds on Ntzoufras et al. (2003) and depends only on a few hyperparameters. Since there can be a large number of covariates in the model, the strategy in Villani et al. (2010) is to incorporate available prior information via the intercepts, and to use a unit-information prior that automatically takes the model geometry and link function into account.

We standardize the covariates to have zero mean and unit variance, and assume prior independence between the intercept and the remaining regression coefficients. The intercepts then have the interpretation of being the (possibly transformed) density parameters at the mean of the original covariates. The strategy in Villani et al. (2010) is to specify priors directly on the parameters of the mixture component, e.g. the degrees of freedom ν in the split-t components, and then back out the implied on the intercept β_{ν_0}. For example, a normal prior for a parameter with identity link (e.g. μ in the split-t model) trivially implies a normal

prior on $\beta_{\mu 0}$; a log-normal prior with mean m^* and variance s^{*2} for a parameter with log link (e.g. ϕ in the split-t model) implies a normal prior $N(m_0, s_0^2)$ for β_{ϕ_0} where

$$m_0 = \ln m^* - \frac{1}{2}\ln\left[\left(\frac{s^*}{m^*}\right)^2 + 1\right] \text{ and } s_0^2 = \ln\left[\left(\frac{s^*}{m^*}\right)^2 + 1\right].$$

The regression coefficients vectors are assumed to be independent a priori. We allow for Bayesian variable selection by augmenting each parameter vector β by a vector of binary covariate selection indicators $\mathcal{I} = (i_1, ..., i_p)$ such that $\beta_j = 0$ if $i_j = 0$. Let $\beta_{\mathcal{I}}$ denote the subset of β selected by $\mathcal{I}$. In a Gaussian linear regression one can use a g-prior (Zellner; 1986) $\beta \sim N[0, \tau_\beta^2 (X'X)^{-1}]$ on the full β and then condition on the restrictions imposed by $\mathcal{I}$. Setting $\tau^2 = n$, where n is the number of observations, gives the unit-information prior, i.e. a prior that carries information equivalent to a single observation from the model. More generally, the unit information prior is $\beta \sim N[0, \tau_\beta^2 \mathcal{I}^{-1}(\beta)]$ where

$$\mathcal{I}(\beta) = -\,\mathrm{E}\left[\left.\frac{\partial^2 \ln p(\beta|y)}{\partial\beta\partial\beta'}\right|_{\beta=\bar{\beta}}\right]$$

and $\bar{\beta} = (\beta_0, 0, ..., 0)'$ is the prior mean of β. When the analytical form of the expected Hessian matrix is not available in closed form, we simulate replicated data sets from a model with parameter vector β_0, and approximate the expected Hessian by the average Hessian over the simulated data sets.

The variable selection indicators are assumed to be independent Bernoulli variables with probability π_β a priori, but more complicated distributions are easily accommodated, see e.g. the extension in Villani et al. (2009) for splines in a mixture context, or a prior which is uniform on the variable selection indicators for a given model size in Denison et al. (2002). It is also possible to estimate π_β as proposed in Kohn et al. (2001) with an extra Gibbs sampling step. Note also that π_β may be different for each parameter in the mixture components. Our default prior has $\pi_\beta = 0.5$.

The prior on the mixing function decomposes as

$$p(\gamma, \mathcal{Z}, s) = p(s|\gamma, \mathcal{Z})p(\gamma|\mathcal{Z})p(\mathcal{Z}),$$

where $\mathcal{Z}$ is the $p \times (K-1)$ matrix with variable selection indicators for the p covariates in the mixing function (recall that $\gamma_1 = 0$ for identification). The variable indicators in $\mathcal{Z}$ are assumed to be *iid* Bernoulli(ω_γ). Let $\gamma_{\mathcal{Z}}$ be the prior on $\gamma = (\gamma_2', ..., \gamma_m')'$ of the form

$$\gamma_{\mathcal{Z}}|\mathcal{Z} \sim N(0, \tau_\gamma^2 I),$$

and $\gamma_{\mathcal{Z}^c} = 0$ with probability one. We use $\tau_\gamma^2 = 10$ as the default value. Finally, $p(s|\gamma, \mathcal{Z})$ is given by the multinomial logit model in (2). To reduce the number of parameters and to speed up the MCMC algorithm we restrict the columns of $\mathcal{Z}$ to be identical, i.e. we make the assumption that a covariate is either present in the mixing function in all components, or does not appear at all, but the extension to general $\mathcal{Z}$ is straightforward; see Villani et al. (2009).

3. Inference methodology

3.1. **The general MCMC scheme.** We use MCMC methods to sample from the joint posterior distribution, and draw the parameters and variable selection indicators in blocks.

The algorithm below is the preferred algorithm from the experiments in Villani et al. (2009). The number of components is determined by a Bayesian version of cross-validation discussed in Section 3.3.

The MCMC algorithm is very general, but for conciseness we describe it for the smooth mixture of split-t components. The algorithm is a Metropolis-within-Gibbs sampler that draws parameters using the following six blocks:

(1) $\{(\beta_\mu^{(k)}, \mathcal{I}_\mu^{(k)})\}_{k=1,...,K}$
(2) $\{(\beta_\phi^{(k)}, \mathcal{I}_\phi^{(k)})\}_{k=1,...,K}$
(3) $\{(\beta_\lambda^{(k)}, \mathcal{I}_\lambda^{(k)})\}_{k=1,...,K}$
(4) $\{(\beta_\nu^{(k)}, \mathcal{I}_\nu^{(k)})\}_{k=1,...,K}$
(5) $s = (s_1, ..., s_n)$
(6) γ and $\mathcal{I}_Z$.

The parameters in the different components are independent conditional on s. This means that each of the first four blocks split up into K independent updating steps. Each updating step in the first four blocks is sampled using highly efficient tailored MH proposals following a general approach described in the next subsection. The latent component indicators in s are independent conditional on the model parameters and are drawn jointly from their full conditional posterior. Conditional on s, Step 6 is a multinomial logistic regression with variable selection, and γ and $\mathcal{I}_Z$ are drawn jointly using a generalization of the method used to draw blocks 1-4; see Villani et al. (2009) for details.

It is well known that the likelihood function in mixture models is invariant with respect to permutations of the components, see e.g. Celeux et al. (2000), Jasra et al. (2005) and Frühwirth-Schnatter (2006). The aim here is to estimate the predictive density, so label switching is neither a numerical nor a conceptual problem (Geweke; 2007). If an interpretation of the mixture components is required, then it is necessary to impose some identification restrictions on some of the model parameters, e.g. an ordering constraint (Jasra et al.; 2005). Restricting some parameters to be common across components is clearly also helpful for identification.

3.2. **Updating β and $\mathcal{I}$ using variable-dimension finite-step Newton proposals.** Nott and Leonte (2004) extend the method which was introduced by Gamerman (1997) for generating MH proposals in a generalized linear model (GLM) to the variable selection case. Villani et al. (2009) extend the algorithm to a general setting not restricted to the exponential family. We first treat the problem without variable selection. The algorithm in Villani et al. (2009) only requires that the posterior density can be written as

$$p(\beta|y) \propto p(y|\beta)p(\beta) = \prod_{i=1}^{n} p(y_i|\varphi_i)p(\beta), \tag{7}$$

where $\varphi_i = x_i'\beta$ and x_i is a covariate vector for the ith observation. Note that $p(\beta|y)$ may be a conditional posterior density and the algorithm can then be used as a step in a Metropolis-within-Gibbs algorithm. The full conditional posteriors for blocks 1–4 in Section 3.1 are clearly all of the form in (7). Newton's method can be used to iterate R steps from the current point β_c in the MCMC sampling toward the mode of $p(\beta|y)$, to obtain $\hat{\beta}$ and the Hessian at $\hat{\beta}$. Note that $\hat{\beta}$ may not be the mode but is typically close to it already after a few Newton iterations, so setting $R = 1, 2$ or 3 is usually sufficient. This makes the algorithm fast, especially when the gradient and Hessian are available in closed form, which is the case here, see Appendix A.

Having obtained good approximations of the posterior mode and covariance matrix from the Newton iterations, the proposal β_p is now drawn from the multivariate t-distribution with $g > 2$ degrees of freedom:

$$\beta_p|\beta_c \sim t\left[\hat{\beta}, -\left(\frac{\partial^2 \ln p(\beta|y)}{\partial\beta\partial\beta'}\right)^{-1}\Bigg|_{\beta=\hat{\beta}}, g\right],$$

where the second argument of the density is the covariance matrix.

In the variable selection case we propose β and $\mathcal{I}$ simultaneously using the decomposition

$$g(\beta_p, \mathcal{I}_p|\beta_c, \mathcal{I}_c) = g_1(\beta_p|\mathcal{I}_p, \beta_c)g_2(\mathcal{I}_p|\beta_c, \mathcal{I}_c),$$

where g_2 is the proposal distribution for $\mathcal{I}$ and g_1 is the proposal density for β conditional on $\mathcal{I}_p$. The Metropolis-Hasting acceptance probability is

$$a[(\beta_c, \mathcal{I}_c) \to (\beta_p, \mathcal{I}_p)] = \min\left(1, \frac{p(y|\beta_p, \mathcal{I}_p)p(\beta_p|\mathcal{I}_p)p(\mathcal{I}_p)g_1(\beta_c|\mathcal{I}_c, \beta_p)g_2(\mathcal{I}_c|\beta_p, \mathcal{I}_p)}{p(y|\beta_c, \mathcal{I}_c)p(\beta_c|\mathcal{I}_c)p(\mathcal{I}_c)g_1(\beta_p|\mathcal{I}_p, \beta_c)g_2(\mathcal{I}_p|\beta_c, \mathcal{I}_c)}\right).$$

The proposal density at the current point $g_1(\beta_c|\mathcal{I}_c, \beta_p)$ is a multivariate t-density with mode $\widetilde{\beta}$ and covariance matrix equal to the negative inverse Hessian evaluated at $\widetilde{\beta}$, where $\widetilde{\beta}$ is the point obtained by iterating R steps with the Newton algorithm, this time starting from β_p. A simple way to propose $\mathcal{I}_p$ is to randomly select a small subset of $\mathcal{I}_c$ and then always propose a change of the selected indicators. It is important to note that β_c and β_p may now be of different dimensions, so the original Newton iterations no longer apply. We will instead generate β_p using the following generalization of Newton's method. The idea is that when the parameter vector β changes dimensions, the dimension of the functionals $\varphi_c = x'\beta_c$ and $\varphi_p = x'\beta_p$ stay the same, and the two functionals are expected to be quite close. A generalized Newton update is

$$\beta_{r+1} = A_r^{-1}(B_r\beta_r - s_r), \qquad (r = 0, ..., R-1), \tag{8}$$

where $\beta_0 = \beta_c$, and the dimension of β_{r+1} equals the dimension of β_p, and

$$\begin{aligned} s_r &= X'_{r+1}d + \frac{\partial \ln p(\beta)}{\partial\beta} \\ A_r &= X'_{r+1}DX_{r+1} + \frac{\partial^2 \ln p(\beta)}{\partial\beta\partial\beta'} \\ B_r &= X'_{r+1}DX_r + \frac{\partial^2 \ln p(\beta)}{\partial\beta\partial\beta'}, \end{aligned} \tag{9}$$

where d is an n-dimensional vector with gradients $\partial \ln p(y_i|\varphi_i)/\partial\varphi_i$ for each observation currently allocated to the component being updated. Similarly, D is a diagonal matrix with Hessian elements

$$\frac{\partial^2 \ln p(y_i|\varphi_i)}{\partial\varphi_i\partial\varphi_i'},$$

X_r is the matrix with the covariates that have non-zero coefficients in β_r, and all expressions are evaluated at $\beta = \beta_r$. For the prior gradient this means that $\partial \ln p(\beta)/\partial\beta$ is evaluated at β_r, including all zero parameters, and that the sub-vector conformable with β_{r+1} is extracted from the result. The same applies to the prior Hessian (which does not depend on β however, if the

prior is Gaussian). Note that we only need to compute the scalar derivatives $\partial \ln p(y_i|\phi_i)/\partial\phi_i$ and $\partial^2 \ln p(y_i|\phi_i)/\partial\phi_i^2$.

3.3. **Model comparison.** The number of components is assumed known in our MCMC scheme above. A Bayesian analysis via mixture models with an unknown number of components is possible using e.g., Dirichlet process mixtures (Escobar and West; 1995), reversible jump MCMC (Richardson and Green; 1997) and birth-and-death MCMC (Stephens; 2000). The fundamental quantity determining the posterior distribution of the number of components is the marginal likelihood of the models with different number of components. It is well-known, however, that the marginal likelihood is sensitive to the choice of prior, and this is especially true when the prior is not very informative, see e.g. Kass (1993) for a general discussion and Richardson and Green (1997) in the context of density estimation.

Following Geweke and Keane (2007) and Villani et al. (2009), we therefore compare and select models based on the out-of-sample Log Predictive Density Score (LPDS). By sacrificing a subset of the observations to update/train the vague prior we remove much of the dependence on the prior, and obtain a better assessment of the predictive performance that can be expected for future observations. To deal with the arbitrary choice of which observations to use for estimation and model evaluation, we use B-fold cross-validation of the log predictive density score (LPDS):

$$\frac{1}{B}\sum_{b=1}^{B} \ln p(\tilde{y}_b|\tilde{y}_{-b}, x),$$

where $\tilde{y}_b$ is an n_b-dimensional vector containing the n_b observations in the bth test sample and $\tilde{y}_{-b}$ denotes the remaining observations used for estimation. If we assume that the observations are independent conditional on θ, then

$$p(\tilde{y}_b|\tilde{y}_{-b}, x) = \int \prod_{i\in\mathcal{T}_b} p(y_i|\theta, x_i)p(\theta|\tilde{y}_{-b})\mathrm{d}\theta,$$

where $\mathcal{T}_b$ is the index set for the observations in $\tilde{y}_b$, and the LPDS is easily computed by averaging $\prod_{i\in\mathcal{T}_b} p(y_i|\theta, x_i)$ over the posterior draws from $p(\theta|\tilde{y}_{-b})$. This requires sampling from each of the B posteriors $p(\theta|\tilde{y}_{-b})$ for $b = 1, ..., B$, but these MCMC runs can all be run in isolation from each other and are therefore ideal for straight-forward parallel computing on widely available multi-core processors. Cross-validation is less appealing in a time series setting since it is typically false that the observations are independent conditional on the model parameters for time series data. A more natural approach is to use the most recent observations in a single test sample, see Villani et al. (2009).

4. Applications

4.1. **A small simulation study.** The simulation study in Villani et al. (2009) explores the out-of-sample performance of a smooth mixture of homoscedastic Gaussian components for heteroscedastic data. The study shows that a smooth mixture of heteroscedastic regressions is likely to be a much more effective way of modelling heteroscedastic data. This section uses simulations to explore how different smooth mixture models cope with skewed and heavy-tailed data. We generate data from the following models:

(1) A one-component normal with mean $\mu = 0$ and variance $\phi^2 = 1$ at $x = \bar{x}$.

(2) A split-normal with mean $\mu = 0$, variance $\phi^2 = 0.5^2$ and skewness parameter $\lambda = 5$ at $x = \bar{x}$.
(3) A student-t with mean $\mu = 0$, variance $\phi^2 = 1$ and $\nu = 5$ degrees of freedom at $x = \bar{x}$.
(4) A split-t with mean $\mu = 0$, variance $\phi^2 = 1$, $\nu = 5$ degrees of freedom, and skewness parameter $\lambda = 5$ at $x = \bar{x}$.

Each of the parameters μ, ϕ, ν and λ are connected to four covariates (drawn independently from the $\mathrm{N}(0, 1)$ distribution) as in (4). Two of the covariates have non-zero coefficients in the data generating process, the other two have zero coefficients. The number of observations in each simulated data set is 1000. We generate 30 data sets for each model and analyze them with both SAGM and a smooth mixture of split-t components using 1-5 mixture components. The priors for the parameters in the estimated models are set as in Table 1.

TABLE 1. Priors in the simulation study

	μ	ϕ	ν	λ
Mean	0	1	10	1
Std	10	1	7	0.8

We analyze the relative performance of SAGM and split-t by comparing the estimated conditional densities $q(y|x)$ with the true data-generating densities $p(y|x)$ using estimates of both the Kullback–Leibler divergence and the L_2 distance, defined respectively as

$$D_{\mathrm{KL}}(p, q) = \sum_{i=1}^{n} p(y_i|x_i) \ln \frac{p(y_i|x_i)}{q(y_i|x_i)},$$

$$D_{L2}(p, q) = 100 \cdot \left(\sum_{i=1}^{n} \left(q(y_i|x_i) - p(y_i|x_i) \right)^2 \right)^{\frac{1}{2}},$$

where $\{y_i, x_i\}_{i=1}^n$ is the estimation data.

Table 2 shows that when the true data is normal (DGP 1), both SAGM and Split-t do well with a single component. The extra coefficients in the degrees of freedom and skewness in the split-t are effectively removed by variable selection. SAGM improves a bit when components are added, while the split-t gets slightly worse.

When the DGP also exhibits skewness (DGP 2), SAGM(1) performs much worse than split-t(1). SAGM clearly improves with more components, but the fit of SAGM(5) is still much worse than the one-component split-t. Note how variable selection makes the performance of the split-t deteriorate only very slowly as we add unnecessary components.

The same story as in the skewed data situation holds when the data are heavy tailed (DGP 3), and when the data are both skewed and heavy tailed (DGP 4).

In conclusion, smooth mixtures with a few complex components can greatly outperform smooth mixtures with many simpler components. Moreover, variable selection is effective in down-weighting unnecessary aspects of the components and makes the results robust to mis-specification of the number of components, even when the components are complex.

4.2. **LIDAR data .** Our first real data set comes from a technique that uses laser-emitted light to detect chemical compounds in the atmosphere (LIDAR, LIght Detection And Ranging). The response variable (logratio) consists of 221 observations on the log ratio of recieved

TABLE 2. Kullback–Leibler and L_2 distance between estimated models and the true DGPs

	Split-t					SAGM				
K	1	2	3	4	5	1	2	3	4	5
					DGP 1 - Normal					
D_{KL}	1.06	1.40	1.54	1.79	2.19	1.31	1.03	0.90	0.95	1.05
D_{L2}	1.73	2.64	3.18	6.11	8.33	2.21	1.52	1.34	1.46	1.71
					DGP 2 - Split-normal					
D_{KL}	3.67	3.67	4.76	4.74	5.57	51.05	14.16	7.30	7.33	8.01
D_{L2}	6.05	6.82	9.51	9.55	13.11	106.13	31.49	16.46	16.20	17.59
					DGP 3 - Student-t					
D_{KL}	1.12	1.72	1.79	2.05	2.20	13.30	1.94	1.78	2.16	2.65
D_{L2}	2.14	4.82	4.70	5.72	5.42	35.79	4.33	3.91	4.70	6.61
					DGP 4 - Split-t					
D_{KL}	3.99	3.24	4.24	4.66	5.67	75.80	21.02	8.89	7.35	7.36
D_{L2}	9.02	8.22	11.78	13.13	16.90	199.99	59.54	27.06	22.43	22.63

light from two laser sources: one at the resonance frequency of the target compound, and the other from a frequency off this target frequency. The predictor is the distance travelled before the light is reflected back to its source (range). The original data comes from Holst et al. (1996) and has been analyzed by for example Ruppert et al. (2003) and Leslie et al. (2007). Our aim is to model the predictive density $p(\mathsf{logratio} \mid \mathsf{range})$.

Leslie et al. (2007) show that a Gaussian model with nonparametric mean and variance can capture this data set quite well. We will initially use the SAGM model in Villani et al. (2009) with the mean, variance and mixing functions all modelled nonparametrically by thin plate splines (Green and Silverman; 1994). Ten equidistant knots in each component are used for each of these three aspects of the model. We use a version of SAGM where the variance functions of the components are proportional to each other, i.e. only the intercepts in the variance functions are allowed to be different across components. The more general model with completely separate variance functions gives essentially the same LPDS, and the posterior distributions of the component variance functions (identified by order-restrictions) are largely over-lapping. We use the variable selection prior in Villani et al. (2009) where the variable selection indicator for a knot κ in the kth mixture component is distributed as $Bernoulli\left[\pi_\beta \cdot \omega_k(\kappa)\right]$. This has the desirable effect of down-weighting knots in regions where the corresponding mixture component has small probability. We compare our results to the smoothly mixing regression (SMR) in Geweke and Keane (2007) which is a special case of SAGM where the components' variance functions are independent of the covariates and any heteroscedasticity is generated solely by the mixture. We use a prior with $m^* = 0$ and $s^{*2} = 10$ in the mean function, and $m^* = 1$ and $s^{*2} = 1$ in the variance function (see Section

2.3). Given the scale of the data, these priors are fairly non-informative. As documented in Villani et al. (2009) and Li et al. (2010), the estimated conditional density and the LPDS are robust to variations in the prior.

TABLE 3. Log predictive density score (LPDS) over the five cross-validation samples for the LIDAR data.

	Linear components			Thin plate components		
	$K=1$	$K=2$	$K=3$	$K=1$	$K=2$	$K=3$
SMR	26.564	59.137	63.162	48.399	61.571	62.985
SAGM	30.719	61.217	64.223	64.267	64.311	64.313

Table 3 displays the five-fold cross-validated LPDS for the SMR and SAGM models, both when the components are linear in covariates and when they are modelled by thin plate splines. The three SAGM models with splines have roughly the same LPDS. The SMR model needs three components to come close the LPDS of the SAGM(1) model with splines, and even then does not quite reach it. All the knots in the variance function of the SAGM models have posterior inclusion probabilities smaller than 0.1, suggesting strongly that the (log) variance function is linear in range. Figure 1 plots the LIDAR data and the 68% and 95% Highest Posterior Density (HPD) regions in the predictive distribution $p(\mathsf{logratio} \mid \mathsf{range})$ from the SMR(3) and the SAGM models with 1, 2 and 3 components. Perhaps the most interesting result in Table 3 and Figure 1 is that SAGM models with more than one component do not seem to overfit. This is quite remarkable since the one-component model fit the data well, and additional components should therefore be a source of over-fitting. This is due to the self-adjusting mechanism provided by the variable/knot selection prior where the already present components automatically becomes simpler (more linear) as more components are added to the model. The estimation results for the SAGM(3) model with spline components (not shown) reveals that the SAGM(3) model with spline components is in fact reduced to essentially a model with linear components. Figure 1 also shows that the fit of the SAGM(3) models with linear components (bottom row, second column) and spline components (second row, second column) are strikingly similar. The same holds for the LPDS in Table 3. Finally, Figure 1 also displays the fit of the split-t model with one component. The estimation results for this model shows that only two knots are really active in the mean function, all of the knots in the scale, degrees of freedom and skewness have posterior probabilities smaller than 0.3. The degrees of freedom are roughly 43 for the smallest values of range and then decreases smoothly toward 7 when range is 720. The skewness parameter λ is roughly 0.5 for all values of range, a sizeable skewness which is also visible in Figure 1. The LPDS of the one-component split-t model is 64.014, which is only slightly worse than SAGM(1).

4.3. **Electricity expenditure data.** Our second example uses a data set with electricity expenditures in 1602 households from South Australia (Bartels et al.; 1996). Leslie et al. (2007) analyze this data set and conclude that a heteroscedastic regression with errors following a Dirichlet process mixture fits the data well. They also document that the response variable is quite skewed. We consider both in-sample and out-of-sample performance of smooth mixture

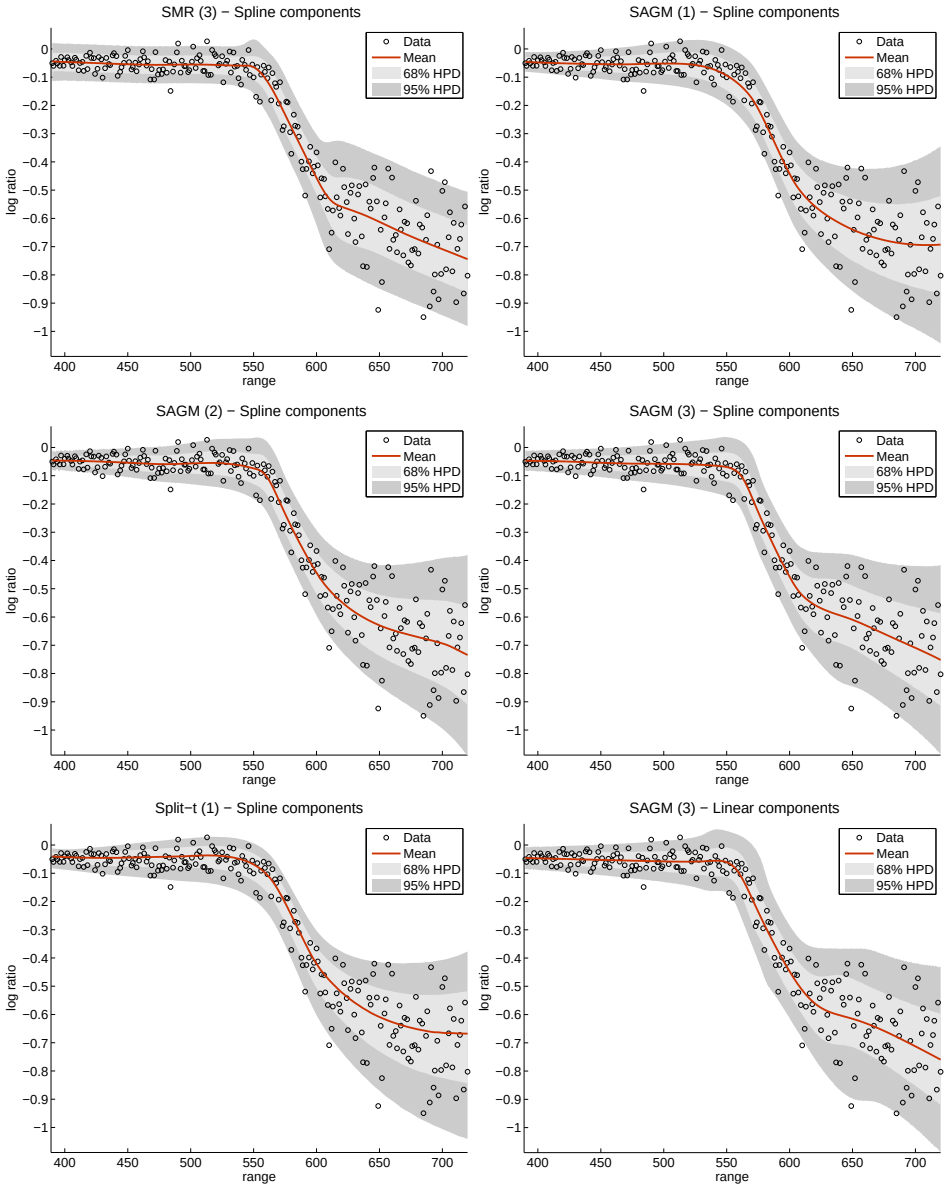

FIGURE 1. Assessing the in-sample fit of the smooth mixture models for the LIDAR data. The figure displays the actual data overlayed on HPD predictive regions. The solid line is the predictive mean.

models, using the data set in Leslie et al. (2007) without interactions. The thirteen covariates used in our application are defined in Table 4

Following Leslie et al. (2007), we mean correct the covariates, but keep their original scale.

The prior means of μ and ϕ are set equal to the median and the standard deviation of the response variable, respectively. This data snooping is innocent as we set the standard deviation of μ and ϕ to 100, so the prior is largely non-informative. The prior mean and standard deviation of the skewness parameter, λ are both set to unity. This means that we

TABLE 4. The electricity bills regressors (subsets)

Variable name	Description
log(rooms)	log of the number of rooms in the house
log(income)	log of the annual pretax household income in Australian dollars
log(people)	log of the number of usual residents in the house
mhtgel	indicator for electric main heating
sheonly	indicator for electric secondary heating only
whtgel	indicator for peak electric water heating
cookel	indicator for electric cooking only
poolfilt	indicator for pool filter
airrev	indicator for reverse cycle air conditioning
aircond	indicator for air conditioning
mwave	indicator for microwave
dish	indicator for dishwasher
dryer	indicator for dryer

are centering the prior on a symmetric model, but allowing for substantial skewness a priori. The prior mean of the degrees of freedom is set to 10 with a standard deviation of 7, which is wide enough to include both the Cauchy and essentially the Gaussian distributions. Since the data sample is fairly large, and we base model choice on the LPDS, the results are insensitive to the exact choice of priors.

TABLE 5. Log Predictive Density Score (LPDS) from five-fold cross-validation of the electricity expenditure data.

Model	$K=1$	$K=2$	$K=3$
SMR *separate*	$-8,047$	$-8,304$	$-8,703$
common	–	$-8,388$	$-8,865$
SAGM *separate*	$-8,280$	$-8,337$	$-8,703$
common	–	$-8,214$	$-8,148$
Split-normal *separate*	$-8,267$	$-8,247$	$-8,369$
common	–	$-8,192$	$-8,174$
Student's t *separate*	$-8,165$	$-8,077$	$-8,151$
common	–	$-8,186$	$-8,148$
Split-t *separate*	$-8,088$	$-8,143$	$-8,157$
common	–	$-8,274$	$-8,224$
Gamma *separate*	$-8,105$	$-8,114$	$-8,143$
common	–	$-8,333$	$-8,304$
Log-normal *separate*	$-8,168$	$-8,142$	$-8,291$
common	–	$-8,087$	$-8,090$

The numerical standard errors of the LPDS are smaller than one for all models.

We first explore the out-of-sample performance of several smooth mixture models using five-fold cross-validation of the LPDS. The five subsamples are chosen by sampling systematically

from the data set. Table 5 displays the results for a handful of models. Every model is estimated both under the assumption of separate parameters and when all parameters except the intercepts are common across components; see Section 2.2.

Looking first at the LPDS of the one-component models, it is clear that data are skewed (the skewed models are all doing better than SAGM), but the type of the skewness is clearly important (gamma is doing a lot better than split-normal and log-normal). The best one-component model is split-t, which indicates the presence of heavy-tails in additional to skewness.

The best model overall is the student-t model with two separate components, closely followed by the log-normal model also with two separate components. It seems that this particular data set has a combination of skewness and heavy-tailedness which is better modeled by a mixture than by a single skewed and heavy-tailed component.

One way to check the in-sample fit of the models on the full data set is look at the normalized residuals. We define the normalized residual as $\Phi^{-1}[F(y_i)]$, where $F(\cdot)$ is the distribution function from the model. If the model is correctly specified, the normalized residuals should be an *iid* $N(0,1)$ sample. Figure 2 displays QQ-plots for the models with one to three components. The QQ-plots should be close to the 45 degree line if the model is correctly specified. It is clear from the first row of Figure 2 that a model with one component has to be skewed in order to fit the data. As expected, most of the models provide a better fit as we add components, the main exception being the split-t which deteriorates as we go from one to two components. This may be due to the MCMC algorithm getting stuck in a local mode, but several MCMC runs gave very similar results.

Table 6 presents estimation results from the best one-component model, the split-t model. We choose to present results for this model as it is easy to interpret and requires no additional identifying restrictions. Table 6 shows that many of the covariates, including log(room) and log(people), are important in the mean function. log(income) gives a relatively low posterior inclusion probability in the mean function, but is an important covariate in the scale, ϕ. The covariate sheonly is the only important variable in the degrees of freedom function, but at least seven covariates are very important determinants of the skewness parameter.

Figure 3 depicts the conditional predictive densities $p(y|x)$ from three of the models: split-$t(1)$ (the best one-component model), student-$t(2)$ (the best model overall) and Gamma(1) (the most efficient model with a minimum number of potential parameters). The predictive densities are displayed for three different conditioning values of the most important covariates: log(rooms), log(income), sheonly and whtgel. All other covariates except the one indicated below the horizontal axis are fixed at their sample means. It is clear from Figure 3 that the predictive densities are very skewed, but also that the different models tend to produce very different types of skewness. The predictive densities from the 2-component student-t model are unimodal except for median and high values of whtgel where the two components are clearly visible.

5. Conclusions

We have presented a general model class for estimating the distribution of a continuous variable conditional on a set of covariates. The models are finite smooth mixtures of component densities where the mixture weights and all component parameters are functions of covariates. The inference methodology is a fully unified Bayesian approach based on a general and efficient MCMC algorithm. Easily specified priors are used and Bayesian variable selection is carried out to obtain model parsimony and guard against over-fitting. We use the

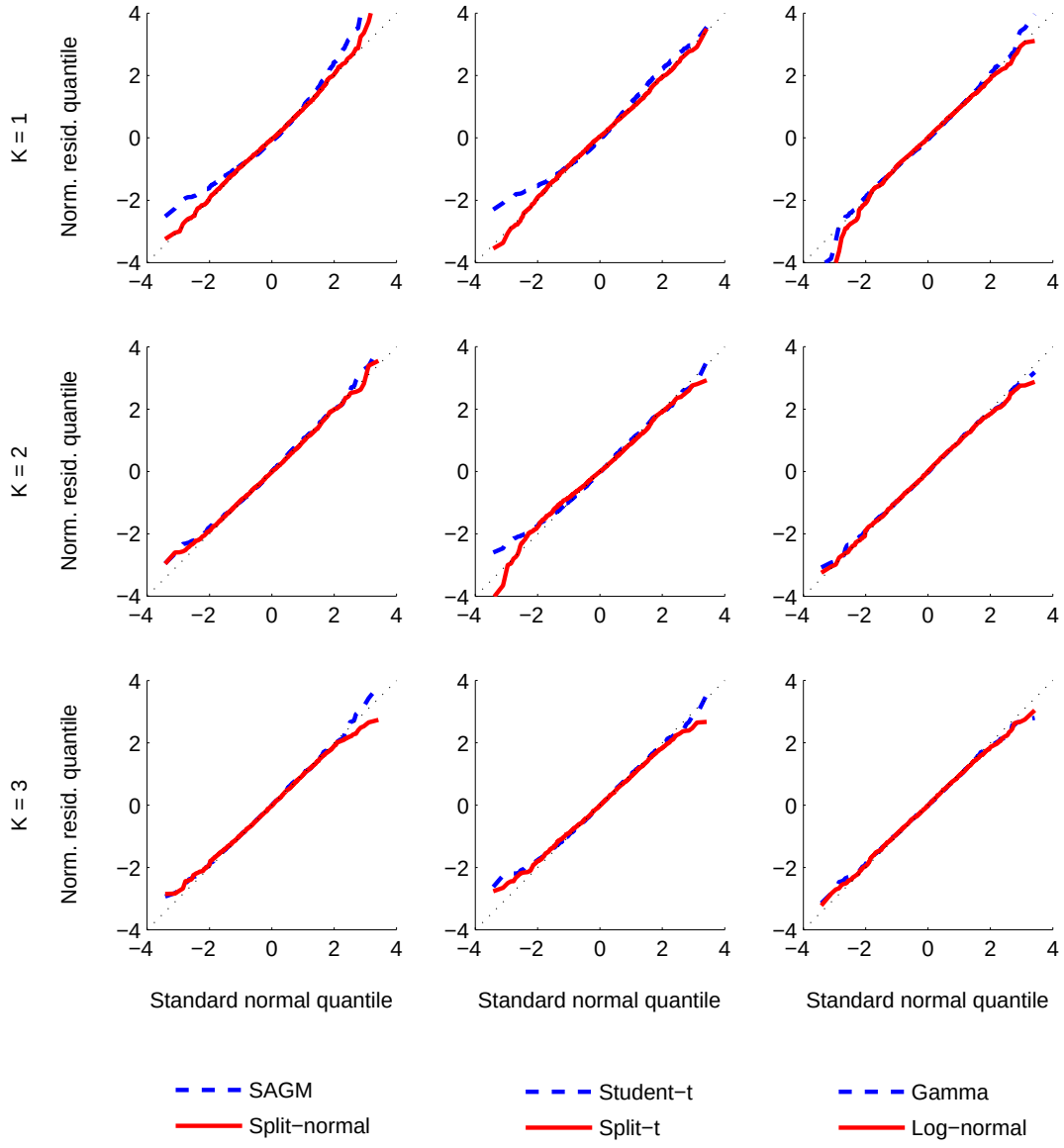

FIGURE 2. Quantiles plots of the normalized residuals resulting from SAGM and split-normal (first column); student's t and split-t (second column); gamma and log-normal (third column) with one to three separate components respectively. If the model is correct, the normalized residuals should be on the dotted reference line.

log predictive density score to determine the number of mixture components. Simulation and real examples show that using fairly complex components in the mixture is a wise strategy and that variable selection is an efficient approach to guard against over-fitting.

TABLE 6. Posterior means and inclusion probabilities in the one-component split-t model for the electricity expenditure data.

Variable	β_μ	$\mathcal{I}_\mu$	β_ϕ	$\mathcal{I}_\phi$	β_ν	$\mathcal{I}_\nu$	β_λ	$\mathcal{I}_\lambda$
Intercept	256.62	–	3.82	–	2.83	–	1.34	–
log(rooms)	**49.47**	0.90	−0.65	0.43	−0.05	0.04	**0.97**	1.00
log(income)	2.71	0.48	**−0.36**	1.00	−0.05	0.02	**0.55**	1.00
log(people)	**40.62**	1.00	−0.20	0.22	0.06	0.03	**0.34**	1.00
mhtgel	**27.28**	1.00	0.07	0.12	−0.18	0.03	0.13	0.15
sheonly	10.11	0.72	0.01	0.04	**2.10**	0.99	0.04	0.05
whtgel	17.74	0.68	−0.23	0.18	0.33	0.04	**0.82**	0.99
cookel	**27.80**	0.99	−0.19	0.14	0.01	0.04	**0.39**	1.00
poolfilt	−6.50	0.50	−0.11	0.23	1.62	0.07	0.32	0.76
airrev	**14.06**	0.91	0.06	0.07	−0.03	0.03	0.12	0.16
aircond	5.58	0.46	0.03	0.11	0.01	0.03	**0.29**	0.96
mwave	8.08	0.75	−0.38	0.49	−0.39	0.05	0.43	0.49
dish	12.96	0.66	0.08	0.05	1.16	0.04	0.11	0.07
dryer	**19.64**	0.99	0.06	0.12	−0.29	0.05	**0.20**	0.90

ACKNOWLEDGMENT

We would like to thank Denzil Fiebig for the use of the electricity expenditure data. Robert Kohn's research was partially supported by ARC Discovery grant DP0988579.

APPENDIX A. IMPLEMENTATION DETAILS FOR THE GAMMA AND LOG-NORMAL MODELS

The general MCMC algorithm documented in Section 3 only requires the gradient and Hessian matrix of the conditional posteriors for each of the parameters in the components densities. The gradient and Hessian for the split-t model is documented in Li et al. (2010). We now present the gradient and Hessian for the gamma model and log-normal model for completeness.

(1) Gradient and Hessian wrt μ and ϕ for the gamma density.

$$\frac{\partial \ln p\left(y|\mu,\phi\right)}{\partial \mu} = \frac{1}{\phi}\left(\mu + 2\mu\log\left(\frac{y\mu}{\phi}\right) - 2\mu\psi\left(\frac{\mu^2}{\phi}\right) - y\right)$$

$$\frac{\partial \ln p\left(y|\mu,\phi\right)}{\partial \phi} = \frac{\mu}{\phi^2}\left(y - \mu - \mu\log\left(\frac{y\mu}{\phi}\right) + \mu\psi\left(\frac{\mu^2}{\phi}\right)\right)$$

$$\frac{\partial^2 \ln p\left(y|\mu,\phi\right)}{\partial \mu^2} = \frac{1}{\phi}\left(3 + 2\log\left(\frac{y\mu}{\phi}\right)\right) - \frac{2}{\phi}\psi\left(\frac{\mu^2}{\phi}\right) - \frac{\mu^2}{\phi^2}\psi_1\left(\frac{\mu^2}{\phi}\right)$$

$$\frac{\partial^2 \ln p\left(y|\mu,\phi\right)}{\partial \phi^2} = -\frac{\mu}{\phi^3}\left(2y - 3\mu - 2\mu\log\left(\frac{y\mu}{\phi}\right)\right) - \frac{2\mu^2}{\phi^3}\psi\left(\frac{\mu^2}{\phi}\right) - \frac{\mu^4}{\phi^4}\psi_1\left(\frac{\mu^2}{\phi}\right)$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ are the digamma function and trigamma function respectively.

(2) Gradient and Hessian wrt μ and ϕ for the log-normal density.
It is convenient to define $h = \log\left(y/\mu\right)$ and $l = \log\left(1 + \phi^2/\mu^2\right)$.

$$\frac{\partial \ln p\left(y|\mu,\phi\right)}{\partial \mu} = \frac{\phi^2 + \left(\mu^2 + \phi^2\right)h}{\mu\left(\mu^2 + \phi^2\right)l} + \frac{\left(2\mu^2 + 3\phi^2\right)}{4\mu\left(\mu^2 + \phi^2\right)} - \frac{\phi^2 h^2}{\mu\left(\mu^2 + \phi^2\right)l^2}$$

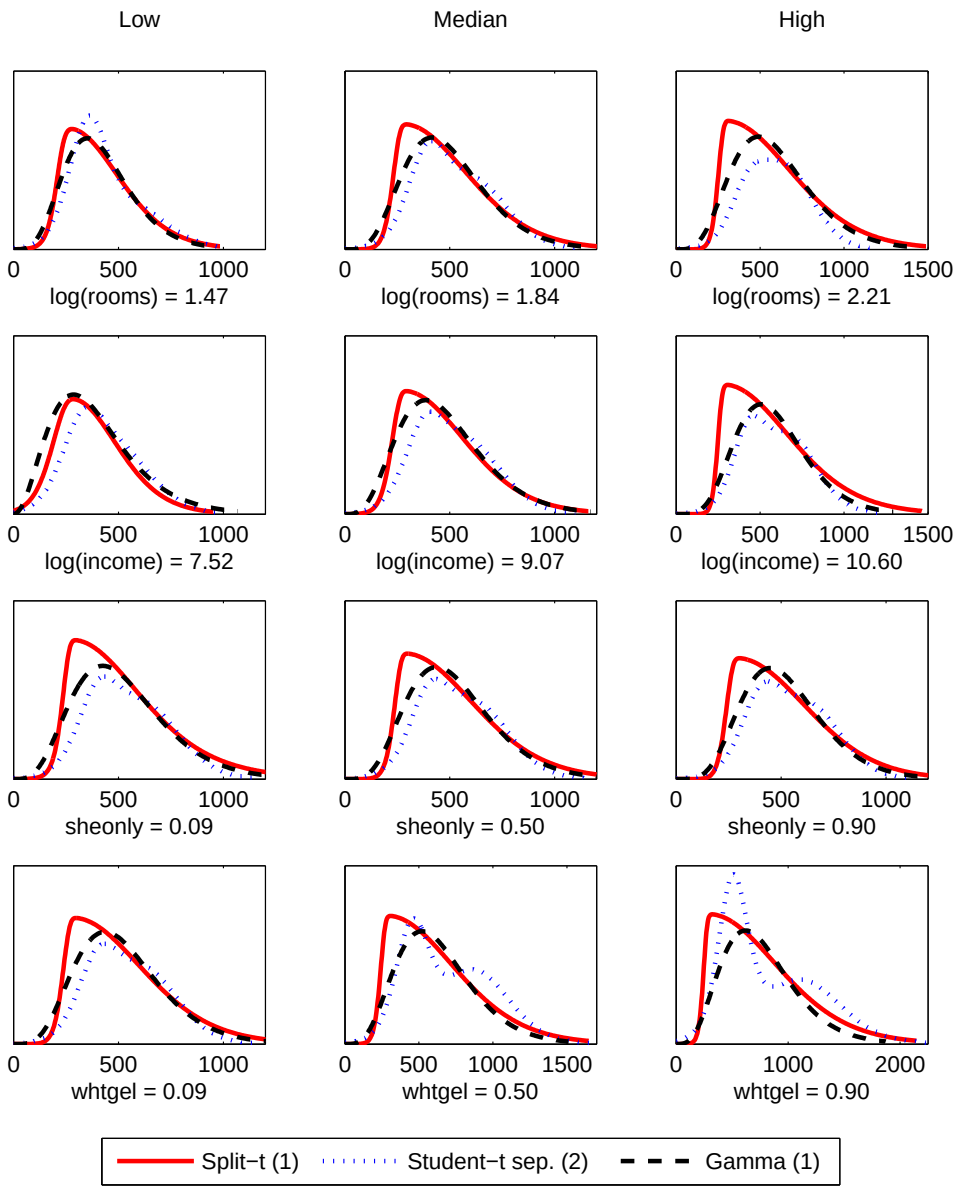

FIGURE 3. Conditional predictive densities for different values of the most important covariates. All other covariates are held fixed at their mean.

$$\frac{\partial \ln p\left(y|\mu,\phi\right)}{\partial \phi} = \frac{\phi \log^2 h}{\left(\mu^2+\phi^2\right) l^2} - \frac{\phi}{\left(\mu^2+\phi^2\right) l} - \frac{\phi}{4\left(\mu^2+\phi^2\right)}$$

$$\frac{\partial^2 \ln p\left(y|\mu,\phi\right)}{\partial \mu^2} = \frac{\left(\mu^4 + 5\mu^2\phi^2 + 2\phi^4 + \left(\mu^2+\phi^2\right)^2 h\right) l^2 + 4\phi^4 h^2}{\mu^2\left(\mu^2+\phi^2\right)^2 l^3} - \frac{\phi^2\left(\left(3h^2+4h\right)\mu^2 + \left(2+4h+h^2\right)\phi^2\right)}{\mu^2\left(\mu^2+\phi^2\right)^2 l^2}$$

$$\frac{\partial^2 \ln p\left(y|\mu,\phi\right)}{\partial \phi^2} = \frac{4\phi^2 h^2}{\left(\mu^2+\phi^2\right)^2 l^3} - \frac{2\phi^2 + \left(\mu^2-\phi^2\right)h^2}{\left(\mu^2+\phi^2\right)^2 l^2} + \frac{\mu^2-\phi^2}{\left(\mu^2+\phi^2\right)^2}\left(1+l^{-1}\right).$$

References

Bartels, R., Fiebig, D. and Plumb, M. (1996). Gas or electricity, which is cheaper? An econometric approach with application to Australian expenditure data, *The Energy Journal* **17**(4): 33–58.

Celeux, G., Hurn, M. and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions, *Journal of the American Statistical Association* **95**(451): 957–970.

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002). *Bayesian methods for nonlinear classification and regression*, Wiley, New York.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(2): 363–375.

Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures., *Journal of the American Statistical Association* **90**(430).

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Springer, New York.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**(1): 57–68.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works, *Computational Statistics and Data Analysis* **51**(7): 3529–3550.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions, *Journal of Econometrics* **138**(1): 252–290.

Gibbons, J. and Mylroie, S. (1973). Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions, *Applied Physics Letters* **22**(11): 568.

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

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of lidar measurements, *Environmetrics* **7**(4): 401–416.

Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts, *Neural Computation* **3**(1): 79–87.

Jasra, A., Holmes, C. and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling, *Statistical Science* **20**(1): 50–67.

Jiang, W. and Tanner, M. (1999a). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation, *Annals of Statistics* **27**(3): 987–1011.

Jiang, W. and Tanner, M. (1999b). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models, *Neural Computation* **11**(5): 1183–1198.

John, S. (1982). The three-parameter two-piece normal family of distributions and its fitting, *Communications in Statistics–Theory and Methods* **11**(8): 879–885.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation* **6**(2): 181–214.

Kass, R. (1993). Bayes factors in practice, *Journal of the Royal Statistical Society: Series D (The Statistician)* **42**(5): 551–560.

Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions, *Statistics and Computing* **11**(4): 313–322.

Leslie, D., Kohn, R. and Nott, D. (2007). A general approach to heteroscedastic linear regression, *Statistics and Computing* **17**(2): 131–146.

Li, F., Villani, M. and Kohn, R. (2010). Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities, *Journal of Statistical Planning and Inference* . In press, doi:10.1016/j.jspi.2010.04.031.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions, *Annals of Statistics* **38**(3): 1733–1766.

Nott, D. and Leonte, D. (2004). Sampling schemes for Bayesian variable selection in generalized linear models, *Journal of Computational and Graphical Statistics* **13**(2): 362–382.

Ntzoufras, I., Dellaportas, P. and Forster, J. (2003). Bayesian variable and link determination for generalised linear models, *Journal of statistical planning and inference* **111**(1-2): 165–180.

Peng, F., Jacobs, R. A. and Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition, *Journal of the American Statistical Association* **91**(435): 953–960.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society: Series B (Methodological)* **59**(4): 731–792.

Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric regression*, Cambridge University Press, Cambridge.

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**(1): 40–74.

Villani, M., Kohn, R. and Giordani, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures, *Journal of Econometrics* **153**(2): 155–173.

Villani, M., Kohn, R. and Nott, D. (2010). A general approach to regression density estimation for discrete and continuous data. Manuscript.

Wood, S., Jiang, W. and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression, *Biometrika* **89**(3): 513–528.

Zeevi, A. and Meir, R. (1997). Density estimation through convex combinations of densities: Approximation and estimation bounds, *Neural Networks* **10**(1): 99–109.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* **6**: 233–243.

Earlier Working Papers:

For a complete list of Working Papers published by Sveriges Riksbank, see www.riksbank.se

Sveriges Riksbank
Visiting address: Brunkebergs torg 11
Mail address: se-103 37 Stockholm

Website: www.riksbank.se
Telephone: +46 8 787 00 00, Fax: +46 8 21 05 31
E-mail: registratorn@riksbank.se

ISSN 1402-9103 E-Print, Stockholm 2010