Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities

Feng Li[*,a], Mattias Villani[b,a], Robert Kohn[c]

[a] *Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden*
[b] *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden*
[c] *Australian School of Business, University of New South Wales, UNSW, Sydney 2052, Australia*

Abstract

A general model is proposed for flexibly estimating the density of a continuous response variable conditional on a possibly high-dimensional set of covariates. The model is a finite mixture of asymmetric student-t densities with covariate-dependent mixture weights. The four parameters of the components, the mean, degrees of freedom, scale and skewness, are all modeled as functions of the covariates. Inference is Bayesian and the computation is carried out using Markov chain Monte Carlo simulation. To enable model parsimony, a variable selection prior is used in each set of covariates and among the covariates in the mixing weights. The model is used to analyze the distribution of daily stock market returns, and shown to more accurately forecast the distribution of returns than other widely used models for financial data.

Key words: Bayesian inference, Markov Chain Monte Carlo, Mixture of Experts, Variable selection, Volatility modeling.

1. Introduction

This paper is concerned with estimating the conditional predictive distribution $p(y|x)$, where y is a univariate continuous response variable and x is a possibly high-dimensional vector of covariates. Our approach is an exercise in nonparametric regression density estimation since $p(y|x)$ is modeled flexibly both for any given x but also across different covariate values.

Villani et al. (2009) propose the smooth adaptive Gaussian mixture (SAGM) model as 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 functions of the covariates x, with Bayesian variable selection in all three sets of covariates. See Frühwirth-Schnatter (2006) for a comprehensive introduction to mixture models.

Villani et al. (2009) argue in favor of a *complex-and-few* modeling philosophy where enough flexibility is used within the mixture components, so that the number of components can be kept to a minimum; see also Wood et al. (2002). This is in sharp contrast to the *simple-and-many* approach used in the machine learning literature (in particular the mixture-of-experts model introduced in Jacobs et al. (1991), and Jordan and Jacobs (1994)) where the components are often

[*]Corresponding author. Tel.: +46 816 2985; fax: +46 816 7511.
Email addresses: feng.li@stat.su.se (Feng Li), mattias.villani@riksbank.se (Mattias Villani), r.kohn@unsw.edu.au (Robert Kohn)

Forthcoming in Journal of Statistical Planning and Inference
doi:10.1016/j.jspi.2010.04.031
First version: May 28, 2009
Final version: May 10, 2010

linear homoscedastic regressions, or even constant functions. Villani et al. (2009) 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. Moreover, simulations and real applications in Villani et al. (2009) show that a simple-and-many approach can fail to fit heteroscedastic data even with a very large number of components, especially in situations with more than one or two covariates. Having heteroscedastic components in the mixture is therefore crucial for accurately modeling heteroscedastic data.

In one of their applications, Villani et al. (2009) model the distribution of daily stock market returns as a function of lagged returns and smooth measures of recent volatility. The best model uses one component to fit the strong heteroscedasticity in the data and the other two or three components to capture the additional kurtosis and/or skewness. The current paper continues the complex-and-few approach and extends the SAGM model by generalizing the Gaussian components to asymmetric student-t densities, thereby making it possible to capture skewness and excess kurtosis within the components. Each component density has four parameters: location, scale, degrees of freedom and skewness, and each of these four parameters are modeled as function of covariates. This makes it possible to have, e.g. the degrees of freedom smoothly varying over covariate space in a way dictated by the data. An efficient Markov chain Monte Carlo (MCMC) simulation method is proposed that allows for Bayesian variable selection in all four parameters of the asymmetric t density, and in the mixture weights. The variable selection makes it possible to handle a large number of covariates. Reducing the number of effective parameters by variable selection mitigates problems with over-fitting and is also beneficial for the convergence of the MCMC algorithm. The methodology is applied to model the distribution of daily returns from the S&P500 stock market index. It is shown that a smooth mixture of asymmetric student t components outperforms SAGM and other commonly used models for financial data in an out-of-sample evaluation of the predictive density during the financial turmoil in the end of year 2008 and beginning of 2009.

2. The model and prior

2.1. Smooth mixtures

Our model 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 component densities are asymmetric student t densities described in detail in the next section. 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. Jiang and Tanner (1999a,b) show that smooth mixtures with sufficiently many (generalized) linear regression components can approximate any density in the exponential family with arbitrary smooth mean functions. See also Zeevi and Meir (1997) for approximation of densities with mixture models.

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

$$\begin{aligned} \Pr(s_i &= k|x_i, \gamma) = \omega_k(x_i) \\ y_i|(s_i &= k, x_i) \sim p_k(y_i|x_i). \end{aligned}$$

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 are Gaussian with both the mean and variance functions of covariates. Our article extends this model so that the component densities belong to an asymmetric student t family. More specifically, the component models are split-t densities (Geweke, 1989; Hansen, 1994) according to the following definition.

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

$$c \cdot \kappa(\mu, \phi, \nu) I(y \leq \mu) + c \cdot \kappa(\mu, \lambda\phi, \nu) I(y > \mu),$$

where

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

is the kernel of a student t density with variance $\phi^2\nu/(\nu-2)$ and $c = 2[(1+\lambda)\phi\sqrt{\nu}Beta(\frac{\nu}{2}, \frac{1}{2})]^{-1}$ 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-t density (Figure 1, left). The skewness of split-t can approach infinity as ν approaches 3 and when ν approaches infinity, the maximum skewness approaches 1 (Figure 1, right). The split-t distribution reduces to the two-piece normal distribution in Gibbons and Mylroie (1973) and John (1982) as $\nu \to \infty$. The split-t density has the advantage that its interpretation is simple since it is equal to the well-known symmetric student t density on either side of the mode, but any other asymmetric t density can equally well be used in our MCMC methodology, see Section 3.1.

The next lemma gives the first four central moments of the split-t density. We use the following definition of skewness and excess kurtosis

$$\begin{aligned} S(y) &= \frac{E\left[y - E(y)\right]^3}{V(y)^{3/2}} \\ K(y) &= \frac{E\left[y - E(y)\right]^4}{V(y)^2} - 3, \end{aligned}$$

where $V(y)$ denotes the variance. The following lemma, which can be proved by straightforward algebra, gives some basic properties of the split-t distribution.

Lemma 1. *If* $y \sim t(\mu, \phi, \lambda, \nu)$ then

$$\begin{aligned} E(y) &= \mu + h \\ V(y) &= \frac{1+\lambda^3}{1+\lambda}\frac{\nu}{\nu-2}\phi^2 - h^2 \\ E\left[y - E(y)\right]^3 &= 2h^3 + 2h\phi^2\left(\lambda^2+1\right)\frac{\nu}{\nu-3} - 3h\phi^2\frac{\lambda^3+1}{\lambda+1}\frac{\nu}{\nu-2} \\ E\left[y - E(y)\right]^4 &= \frac{3\nu^2\phi^4\left(1+\lambda^5\right)}{(1+\lambda)(\nu-2)(\nu-4)} - 3h^4 + \frac{6h^2\left(1+\lambda^3\right)\nu\phi^2}{(1+\lambda)(\nu-2)} \\ &\quad - \frac{8h^2\left(\lambda^2+1\right)\nu\phi^2}{\nu-3}, \end{aligned}$$

where

$$h = \frac{2\sqrt{\nu}\phi(\lambda-1)}{(\nu-1)\,Beta\left(\frac{\nu}{2},\frac{1}{2}\right)},$$

and moment of order r exists exists if $\nu > r$.

The CDF of a split-t distribution is of the form

$$\frac{1}{1+\lambda} + \frac{a\cdot \mathrm{Sign}(y-\mu)}{1+\lambda}\left[1 - \frac{Beta\left(t;\frac{\nu}{2},\frac{1}{2}\right)}{Beta\left(\frac{\nu}{2},\frac{1}{2}\right)}\right]$$

where

$$t = \frac{\nu a^2\phi^2}{\nu a^2\phi^2 + (y-\mu)^2},$$

and $a = \lambda$ if $y > \mu$ and $a = 1$ otherwise, and $Beta(t;\nu/2,1/2)$ is the incomplete beta function (Abramowitz and Stegun, 1972).

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

$$\begin{aligned} \mu &= \beta_{\mu 0} + x_t'\beta_\mu \\ \ln\phi &= \beta_{\phi 0} + x_t'\beta_\phi \\ \ln\lambda &= \beta_{\lambda 0} + x_t'\beta_\lambda \\ \ln\nu &= \beta_{\nu 0} + x_t'\beta_\nu \end{aligned} \tag{3}$$

but any smooth link function can equally well be used in the 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), but this is not pursued here. A strength of our approach is that the four regression coefficient vectors: β_μ, β_ϕ, β_ν and β_λ 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 (3).

This split-t model will often be flexible enough to fit the data, but there are datasets that require a smooth mixture model, for example when the data are multimodal for some covariates values. A second example occurs when the wrong link function is used in one of the split-t parameters, where the mixture can then correct for this erroneous choice. A third example is when there are outliers in the data that cannot be accommodated by a student t density.

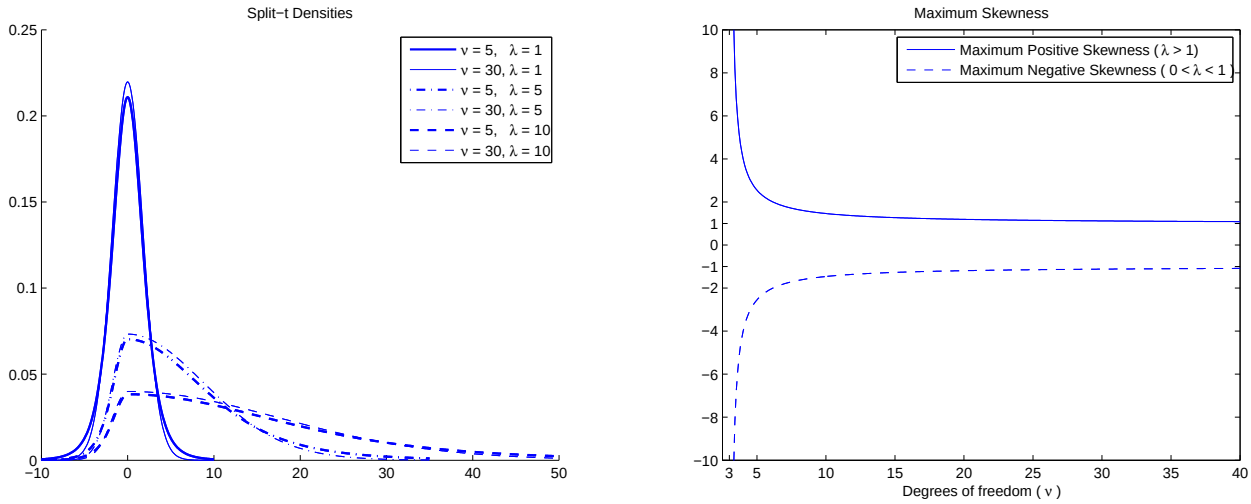

Figure 1: Graphical display of the split-t densities. The left-hand side are the split-t densities with location parameter $\mu = 0$ and skewness parameter $\lambda = 1.8$. The right-hand side is the maximum skewness of the split-t as a function of degrees of freedom.

A smooth mixture of split-t 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 in all four split-t parameters, and in the mixing function. This can lead to important simplifications of the split-t 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 Villani et al. (2007) illustrates this well). Increasing the number of components can therefore in principle even reduce the number of effective parameters in the model.

A more extreme, but often empirically relevant, simplification of the model is to assume that one or more split-t parameters are *common* to the components, that is, only the intercepts in (3) are allowed to be different across components. The unrestricted model where the regression coefficients are allowed to differ across components is said to have *separate* components.

2.3. The prior

Although the MCMC methodology (see Section 3.2) allows any prior distribution, we shall now present an easily specified prior that depends only on a few hyper-parameters. First, we standardize the covariates by subtracting the mean and dividing by the standard deviation. This allows us to assume prior independence between the intercept and the remaining regression coefficients, and the intercepts have the interpretation of being the (possibly transformed) split-t parameters at the mean of the original covariates. Since there can be a large number of covariates in the model, our strategy is to incorporate available prior information via the intercepts, and to treat the remaining regression coefficients more informally. Assuming a normal prior for μ implies a normal prior on $\beta_{\mu 0}$. The other three split-t parameters ϕ, λ and ν are assumed to follow independent log-normal priors with means m^* and s^*, where m^* and s^* are different for the different split-t parameters.

This translates into a normal prior on the intercept with mean

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

and variance

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

The regression coefficients β_μ, β_ϕ, β_ν and β_λ 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 $\mathcal{I}$. We assume the following prior for each β vector

$$\beta_\mathcal{I}|\mathcal{I} \sim N(0, \tau_\beta^2 I)$$

and $\beta_{\mathcal{I}^c}|\mathcal{I}^c$ is identically zero, where $\mathcal{I}^c$ is the complement of $\mathcal{I}$. Alternatively, one can use a g-prior (Zellner, 1986) $\beta \sim N\left[0, \tau_\beta^2 (X'X)^{-1}\right]$ and then condition on the restrictions imposed by $\mathcal{I}$; Denison et al. (2002, p. 80-81) discusses the advantages and disadvantages of these two different priors. The g-prior is less appealing in a mixture context since $(X'X)^{-1}$ may be a bad representation of the covariance between parameters in the smaller components, see Villani et al. (2009) for a discussion, and we will therefore use the identity matrix here. We use $\tau_\beta = 10$ as the default value in our application in Section 4. Given that the covariates have been standardized to zero mean and unit variance, and that the variance of y is roughly one in our empirical example, these priors are vague. We investigate the sensitivity of the posterior inferences and model comparison with respect to τ_β in Section 4.

The variable selection indicators are assumed to be independent Bernoulli with probability ω_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 ω_β as proposed in Kohn et al. (2001) with an extra Gibbs sampling step. Note that ω_β may be different for each split-t parameter. Our default prior has $\omega_\beta = 0.5$.

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 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. 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. Villani et al. (2009) experimented with several different algorithms in a related setting and the algorithm outlined below is similar to their preferred algorithm. The details of the algorithm are given in Appendix A. The method used to select the number of components is discussed in Section 3.3.

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 section. 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.

Mixture models have well-known identification problems, the most serious one being the so-called label switching problem, which means that the likelihood is invariant with respect to permutations of the components in the mixture, see e.g. Celeux et al. (2000), Jasra et al. (2005) and Frühwirth-Schnatter (2006). The aim of our article is to estimate the predictive density, so that label switching is neither a numerical nor 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).

The number of components is assumed known in our MCMC scheme. 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). However, one major drawback is that the posterior distribution of the number of components for a given data set typically depends heavily on the priors. In order to avoid that, we instead compare and select models based on the out-of-sample LPDS (see details in Section 3.3). Our *complex-and-few* approach is also helpful in this aspect as it keeps the number of components to a minimum (see Section 4).

3.2. Updating (β, $\mathcal{I}$) 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

$$p(\beta|y) \propto p(y|\beta)p(\beta) = \prod_{i=1}^{n} p(y_i|\varphi_i)p(\beta), \tag{4}$$

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 (4). 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. This proposal can be refined in many ways, using, e.g. the adaptive scheme in Nott and Kohn (2005), where the history of $\mathcal{I}$-draws is used to adaptively build up a proposal for each indicator. 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

$$\beta_{r+1} = A_r^{-1}(B_r\beta_r - s_r), \qquad (r = 0, ..., R-1), \tag{5}$$

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

$$
\begin{aligned}
s_r &= X'_{r+1} d + \frac{\partial \ln p(\beta)}{\partial \beta} \\
A_r &= X'_{r+1} D X_{r+1} + \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'} \\
B_r &= X'_{r+1} D X_r + \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'},
\end{aligned} \tag{6}
$$

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 $\partial^2 \ln p(y_i|\phi_i)/\partial \phi_i^2$.

After the first Newton iteration the parameter vector no longer changes dimension, and the generalized Newton algorithm in (5) reduces to the original Newton algorithm. Once the simultaneous update of the $(\beta, \mathcal{I})$-pair is completed, we make a final update of the non-zero parameters in β, conditional on the previously accepted $\mathcal{I}$, using the fixed dimension Newton algorithm. This additional step is needed if we choose the simple proposal of $\mathcal{I}$ where we always propose a change of (a subset of) $\mathcal{I}$. Since β and $\mathcal{I}$ are proposed jointly this means that the posterior of β would be updated very infrequently when the posterior of $\mathcal{I}$ is very precise (since most draws of $\mathcal{I}$ will then be rejected). Other ways to propose $\mathcal{I}$ may not benefit from this additional step, e.g. the adaptive scheme in Nott and Kohn (2005). The proposal density $g_1(\beta_p|\mathcal{I}_p, \beta_c)$ is again taken to be the multivariate t-density in exactly the same way as in the case without covariate selection.

When a parameter is restricted to be proportional across components (i.e. only the intercept differs between components), the common regression vector β appears in all K components. The updating step for the common β is of the same form as above, but d and D now contain the gradients and Hessians for *all* n observations, where each observation's gradient and Hessian is with respect to the component density that the observation is currently allocated to.

3.3. Model comparison

The key quantity in Bayesian model comparison is the marginal likelihood. The marginal likelihood is sensitive to the choice of prior, however, 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. 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, one can use B-fold cross-validation of the log predictive density score (LPDS):

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

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 parallel computing on widely available multi-core processors.

Cross-validation is less appealing in a time series setting, and a more natural approach is to use the most recent observations in a single test sample. Moreover, for time series data it is typically false that the observations are independent conditional on the model parameters, so that the above estimation approach cannot be used. An MCMC estimate of the LPDS of a time series can instead be based on the decomposition

$$p(y_{T+1}, .., y_{T+T^*}|y_1, .., y_T) = p(y_{T+1}|y_1, .., y_T) \cdots p(y_{T+T^*}|y_1, .., y_{T+T^*-1}),$$

with each term in the decomposition

$$p(y_t|y_1, .., y_{t-1}) = \int p(y_t|y_1, .., y_{t-1}, \theta) p(\theta|y_1, .., y_{t-1}) \mathrm{d}\theta,$$

estimated from a posterior sample of θ's based on data up to time $t-1$. The problem is that this requires $T^* - T$ complete runs with the MCMC algorithm, one for each term in the decomposition, which is typically very time-consuming (although computer parallelism can again be exploited). In situations where T is fairly large compared to T^*, we can approximate the LPDS by computing each term $p(y_t|y_1, .., y_{t-1})$ using the same posterior sample based on data up to time T. We evaluate the accuracy of this approximation in the empirical application in the next section. Villani et al. (2009) show that the Bayes factor is roughly B times more discriminatory than the LPDS. Therefore one can transform a difference in LPDS between two competing models into a Bayes factor and then use the Jeffreys rule .

Jeffreys (1961) and Kass and Raftery (1995) provide simple rules for interpreting the size of a Bayes factor between two models. A difference in LPDS between models can be seen as log Bayes factor evaluated on the observations in the test sample. Since only a subset of the data is used to evaluate the LPDS, the LPDS has less discriminatory power than the Bayes factor, but the LPDS has the advantage of being substantially less sensitive to the prior. If the scale of evidence in Kass and Raftery (1995, p. 777) is applied to the LPDS, then a difference in LPDS between two models between 3 and 5 is considered strong evidence in favor of one model, and a difference of more than five LPDS points is very strong evidence.

4. Modeling the distribution of daily stock market returns

4.1. S&P500 data and priors

Modeling the volatility/variability in financial data has been an highly active research area since the seminal paper by Engle (1982) introduced the ARCH model (see, e.g. Baillie (2006)

Table 1: The prior mean and standard deviation of the split-t parameters for the S&P500 stock return data. The prior mean of ϕ is a function of the prior mean of ν such that the variance of returns is unity as in Villani et al. (2009).

	μ	ϕ	ν	λ
m^*	0	$[(m_\nu^* - 2)/m_\nu^*]^{1/2}$	10	1
s^*	10	1	7	1

for a survey of the field), and there are large financial markets for volatility-based instruments. Financial data, such as stock market returns, are typically heavy tailed and subject to volatility clustering, i.e. a time-varying variance that evolves in a very persistent fashion. We here model the entire distribution of daily returns from the S&P500 stock market index, $p(y_t|x_t)$, where $y_t = 100\ln(p_t/p_{t-1})$ is the daily return at time t, p_t is the closing S&P500 index on day t, and x_t contains the covariate observations at time t. By focusing on the whole distribution of returns we are able to compute, e.g. the posterior distribution of the *Value-at-Risk* (VaR), i.e. the 1% quantile of the return distribution, which is of fundamental interest to financial analysts, see Villani et al. (2009) for an example based on the S&P500 datasets.

We estimate the models using data from 4646 trading days between Jan 1, 1990 and May 29, 2008. The models are then evaluated out-of-sample on the subsequent 199 trading days from May 30, 2008 to March 13, 2009. The data are plotted in the upper left sub-graph of Figure 2, with the evaluation period marked out in red. To make the results comparable to Geweke and Keane (2007) and Villani et al. (2009), we standardize the covariates to lie in the interval $[-1, 1]$, rather than making them mean zero with unit variance.

Table 1 displays the prior hyper-parameters for the split-t parameters. The prior on ν and λ are fairly vague and and the prior on μ and ϕ have been chosen to match the mean and variance in Villani et al. (2009) as closely as possible. See Section 4.3 for a sensitivity analysis with respect to these prior hyper-parameters.

4.2. Models

Geweke and Keane (2007) show that a smooth mixture of homoscedastic Gaussian regressions (the so-called smoothly mixing regression, SMR) with two covariates outperforms the typically hard-to-beat t-GARCH(1,1) model (Bollerslev, 1987) in an out-of-sample evaluation based on the LPDS (see Section 3.3). The two covariates are the return yesterday y_{t-1} (LastDay) and CloseAbs95, a geometrically decaying average of past absolute returns

$$(1-\rho)\sum_{s=0}^{\infty}\rho^s\left|y_{t-2-s}\right|,$$

where $\rho = 0.95$ is the discount factor. Following Geweke and Keane (2007) we assume the mean of each component to be constant since the level of the stock market returns are not expected to be predictable.

Villani et al. (2009) demonstrate that the SAGM model with its heteroscedastic components outperforms the SMR in Geweke and Keane (2007). Villani et al. (2009) also introduce seven additional covariates and show that they substantially improve the out-of-sample performance of

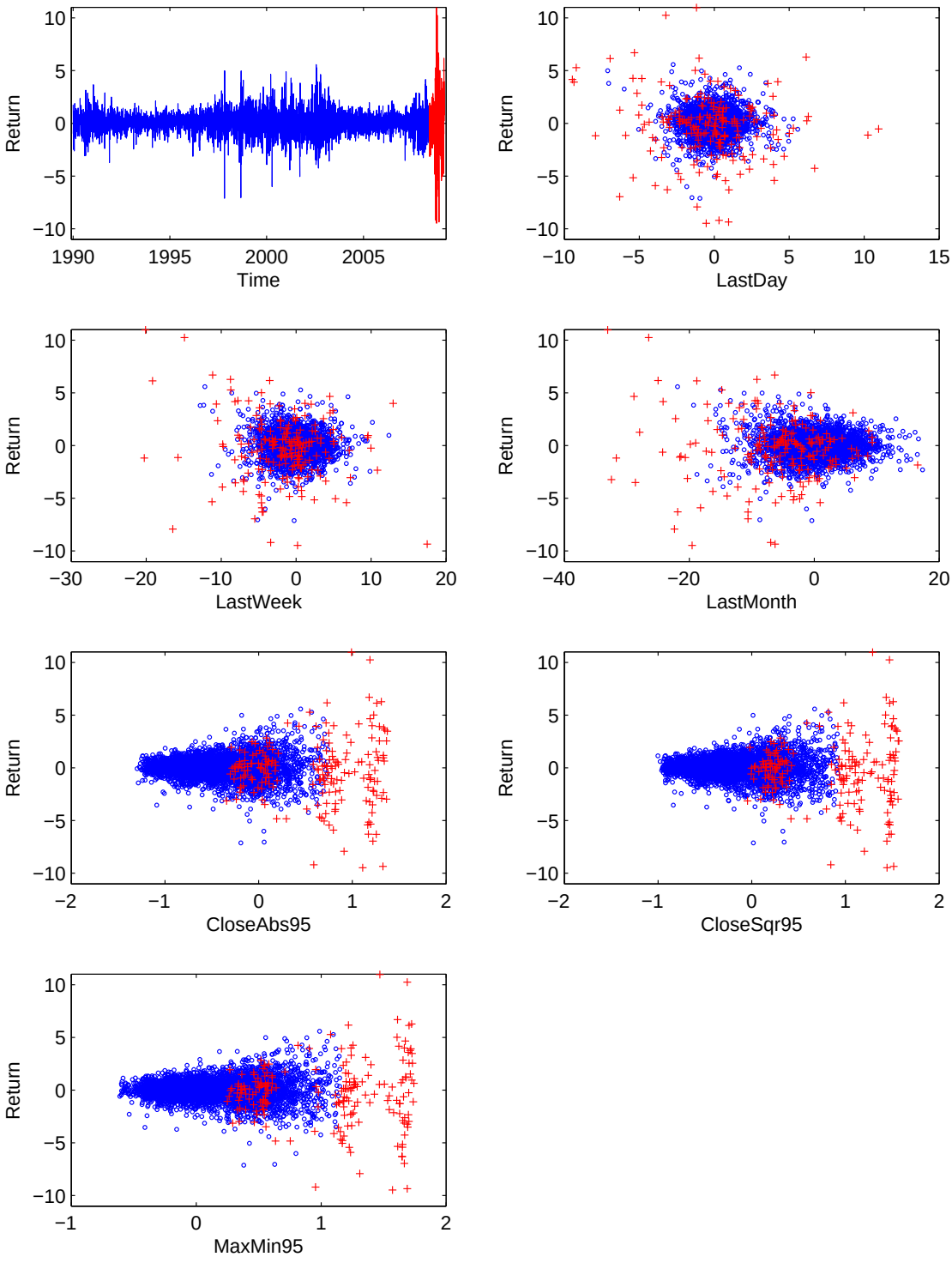

Figure 2: Graphical display of the S&P500 data from January 1, 1990 to May 29, 2008 (blue lines and circles) and May 30, 2008 to March 13, 2009 (red lines and crosses). The subgraph in the upper left position is a time series plot of Return, the other subgraphs are scatter plots of Return against a covariate.

the SAGM. We will concentrate on this nine-variable model. The seven additional covariates are: LastWeek and LastMonth, a moving average of the returns from the previous five and 20 trading days, respectively. The variable CloseAbs80, the same variable as CloseAbs95 but with $\rho = 0.80$, is also added to the covariate set, and so is the square root of $(1-\rho)\sum_{s=0}^{\infty}\rho^s y_{t-2-s}^2$, for $\rho = 0.80$ and 0.95 (CloseSqr80 and CloseSqr95). Finally, Villani et al. (2009) include a measure of volatility that has been popular in the finance literature: $(1-\rho)\sum_{s=0}^{\infty}\rho^s(\ln p_{t-1-s}^{(h)} - \ln p_{t-1-s}^{(l)})$, where $p_t^{(h)}$ and $p_t^{(l)}$ are the highest and lowest values of the S&P500 index at day t. This measure has been shown both theoretically and empirically to carry more information on the volatility than changes in closing quotes (Alizadeh et al., 2002). We consider both $\rho = 0.8$ (MaxMin80) and $\rho = 0.95$ (MaxMin95). As in Villani et al. (2009), all variables except LastDay, LastWeek and LastMonth enter the model in logarithmic form.

4.3. Results

We generated 30,000 draws from the posterior, and used the last 25,000 draws for inference. This is more than sufficient for convergence of the parameter estimates, the posterior inclusion probabilities and the LPDS; see also Villani et al. (2009) for details regarding convergence in the SAGM model. Three Newton steps were used for all parameters, but experiments with a single Newton step gave essentially the same numerical efficiency. The numerical efficiency of the algorithm is documented in some detail below.

Table 2 presents the LPDS evaluated on the 199 trading days from May 30, 2008 to March 13, 2009, a period covering the financial crisis with an unprecedented volatility. Figure 2 shows that prediction in the evaluation period is a tough test of the models because it extrapolates outside the sample used for estimation. The posterior distributions of the models are not updated during the evaluation period (see Section 3.3). With the exception of some of the more poorly fitting models, this approximation of the LPDS is quite accurate. This is documented in Villani et al. (2009) and additional evidence on this issue is provided below.

We observe from Table 2 that the SMR model does poorly, even with a large number of components, and is outperformed by the GARCH(1, 1) and t-GARCH(1, 1) models. A smooth mixture of homoscedastic components can generate some heteroscedasticity in-sample, but is likely to fail in extrapolating heteroscedastic data outside the estimation sample. The subsequent rows of Table 2 present that adding covariate-dependent skewness and/or student t components (with degrees of freedom a function of covariates) to the SMR improves the LPDS substantially when the number of mixture components is small, but the SMR performs better in its standard form with Gaussian components when K is large. This reinforces the conclusion stressed in Villani et al. (2009) that having heteroscedastic components is crucial for modeling heteroscedastic data.

Table 2 also presents that SAGM is on par with the popular t-GARCH(1, 1) already with a single component, outperforms it when $K \geq 2$, and is more than 7 LPDS units better than t-GARCH(1,1) at its maximum when $K = 4$. This is a substantial increase in LPDS since we are only using 199 observation in the evaluation sample (see Section 3.3 for a more detailed discussion).

To ensure that our shortcut of keeping the posterior distribution fixed as we go through the evaluation sample does not invalidate the conclusions from the LPDS, we re-computed the LPDS for the SMR and the SAGM with a common variance function, this time updating the posterior at every tenth observation. The results are given in Table 3. A comparison of Table 2 and 3 shows that there are fairly large differences for the most poorly fitting versions of SMR, but that the LPDS values for SAGM do not change much when the posterior is updated more frequently.

Table 2: Evaluating the out-of-sample log predictive density score (LPDS) on the 199 daily returns in the period May 30, 2008 - March 13, 2009[†].

Model	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	Max n.s.e.
SMR	−1044.78	−638.89	−505.74	−487.11	−489.19	0.98 (3)
+ Skew	−540.91	−525.07	−513.85	−506.68	−506.13	0.82 (2)
+ DF	−544.00	−518.71	−498.93	−500.14	−494.29	0.89 (1)
+ Skew + DF	−530.86	−504.63	−498.03	−498.83	−496.87	0.88 (5)
SAGM Common	−477.73	−473.10	−473.12	−470.30	−472.86	0.26 (2)
+ Skew	−474.18	−467.29	−468.75	−467.93	−467.22	0.35 (4)
+ DF	−474.74	−472.92	−470.51	−469.40	−468.87	0.34 (4)
+ Skew + DF	**−472.37**	−468.92	−469.30	−466.21	**−465.86**	0.53 (4)
SAGM Separate		−469.21	−469.50	−470.53	−471.02	0.49 (3)
+ Skew		−468.48	−466.93	−467.48	−468.02	0.58 (4)
+ DF		−469.08	−469.24	**−462.03**	−467.78	0.72 (5)
+ Skew + DF		**−466.84**	**−462.56**	−462.47	−474.58	0.74 (5)
GARCH(1,1)	−479.03					
t-GARCH(1,1)	−477.39					

[†]The posterior distribution is computed using data until May 29, 2008, and not updated thereafter, except for the two GARCH models which are based on continuously updated maximum likelihood estimates. The LPDS of the best model for a given number of components is in bold font. The last column gives the maximal numerical standard error of the LPDS for each model with the number of components for which the maximum was obtained in parenthesis. The notation for the models is such that e.g. + Skew means that covariate-dependent skewness is added to the model.

Table 3: Evaluating the out-of-sample log predictive density score (LPDS) on the 199 daily returns in the period May 30, 2008 - March 13, 2009[‡].

Model	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
SMR	−982.02	−597.47	−498.87	−484.42	−495.66
SAGM	−477.50	−472.94	−471.28	−471.53	−469.72

[‡]The posterior distribution is updated every 10th observation throughout the evaluation sample.

Table 2 presents that for the one component models, adding either covariate-dependent skewness or degrees of freedom to the SAGM model increases the LPDS by roughly three points, and adding them both increases the LPDS by a further two points. The split-t with covariate-dependent scale, skewness and degrees of freedom is the best one-component model, and its performance is close to that of the best SAGM model with four components. The one-component split-t (SAGM + Skew + DF) is similar to the ARCD model of Hansen (1994) which he uses to model the conditional density of the U.S. Dollar / Swiss Franc exchange rate.

If we restrict the scale, skewness and degrees of freedom to be common across components (up to a proportionality constant) we see that adding components to the split-t model improves its forecasting performance. However, we can get an even better LPDS by using separate components. Note that adding components in this case introduces as much as 41 new parameters to the model for every newly added component, and still we do not seem to over-fit even when the number of components is fairly large. This is because of the self-adjustment mechanism emphasized in Villani et al. (2009): when an additional component is added to the mixture, the variable selection simplifies not only the new component but also the already existing components. The number of effective parameter can therefore even decrease as components are added. But there is a limit to what variable selection can do (see also Figure 4 below), and there are clear signs of over-fitting when $K = 5$. Also, the MCMC algorithm struggles when we use $K \geq 4$ separate components in the split-t model, with lower acceptable probabilities and higher risk of getting stuck in a local mode. Moreover, the split-t model with separate components has one dominant component which is very similar to the one-component model, except for the five-component model which seems to pick up a more complicated structure. We will describe the estimation results for the one-component model in detail below.

Our way to assess the quality of the predictive densities in an absolute sense is to investigate the normalized residuals from the model. A normalized residual is defined as $\Phi^{-1}\left[F(y_t)\right]$, where $F(\cdot)$ is the cumulative predictive distribution, where the parameter have been integrated out with respect to the posterior distribution based on the estimation sample, so the residuals in Figure 3 are therefore out-of-sample. If the model is correct, the normalized residuals should be *iid* $N(0,1)$, see e.g. Berkowitz (2001). It is clear from Figure 3 that even the SMR with largest LPDS produces much to large residuals during the most volatile period, and so does the GARCH(1,1) and t-GARCH(1, 1). As indicated in the graph, 19.5% of the normalized residuals from the SMR(4) lie outside a 95% probability interval according to the $N(0,1)$ reference distribution. The SAGM(1) does better than the SMR, but this model also generates to many outliers: 3.5% of the residuals are outside the 99% reference interval. The remaining four models in Figure 3 have rather similar seemingly homoscedastic and independent residuals, and they all have close to the right coverage. The one-component split-t model is doing remarkably well during this very difficult time period.

We now take a more detailed look at the inferences from the one-component split-t model. Table 4 presents summaries of the posterior distribution. The results from the variable selection among the covariates in the scale parameter is very similar to the results for the variance function in Villani et al. (2009): the covariates MaxMin95, LastWeek and LastMonth have a posterior inclusion probability close to one, and all other covariates are essentially excluded. There is support for some small skewness in the model, but no covariates enter λ. The degrees of freedom at the posterior mean is $\exp(2.482) = 11.96$, (assuming all other covariates at their mean) which is not very heavy tailed, but LastWeek enters the model with probability 0.638 and with a large negative coefficient, so the degrees of freedom is very small for the largest values of LastWeek (recall that

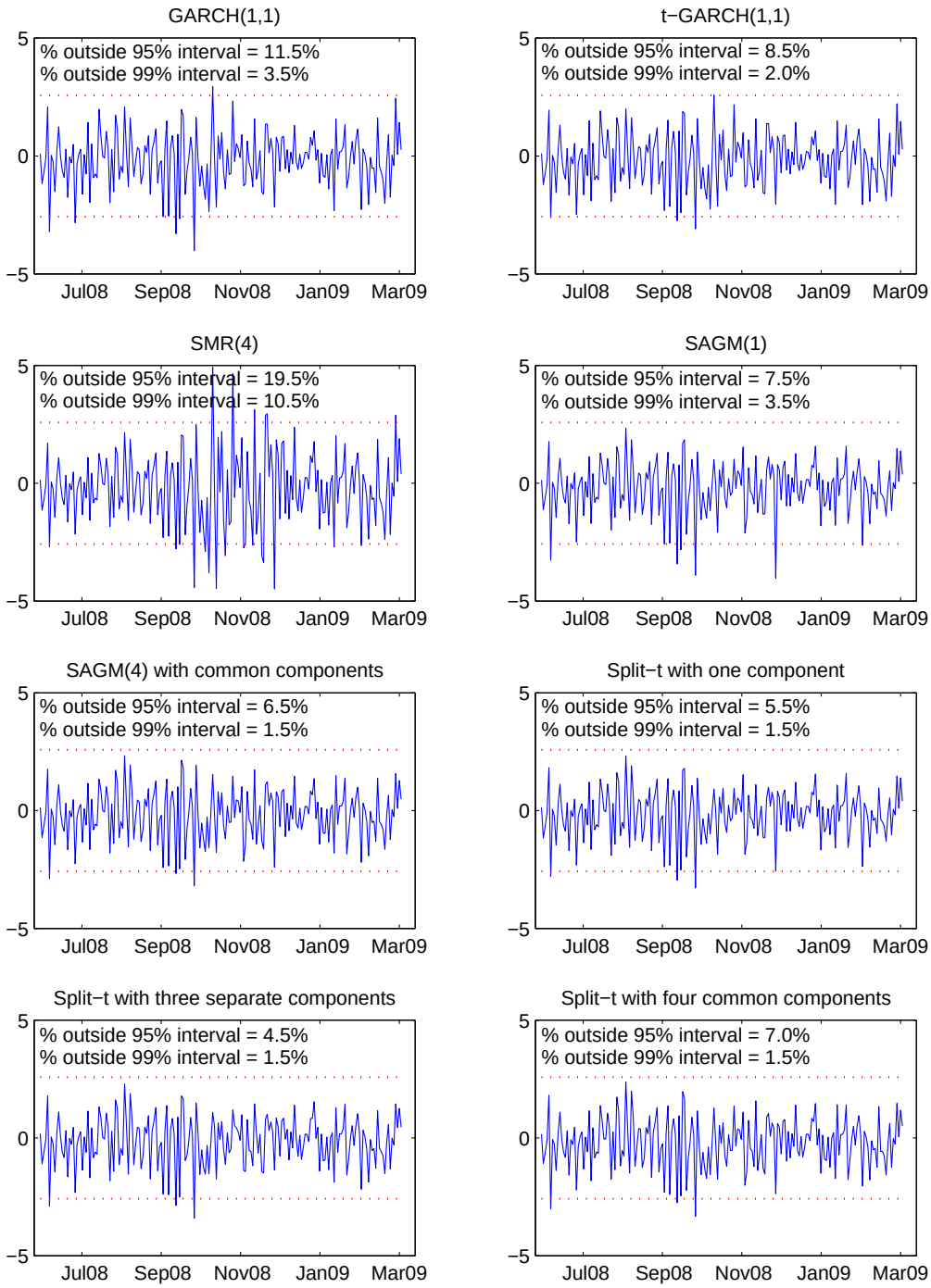

Figure 3: Plot of the 199 normalized residuals in the evaluation sample over time (solid lines). The dotted lines are the 99% probability intervals under the $N(0,1)$ reference distribution. Each sub-graph displays the percentage of normalized residuals outside the 95% and 99% probability intervals of the $N(0,1)$ reference distribution.

Table 4: Posterior summary of the one-component split-t model[8].

Parameters	Mean	Stdev	Post.Incl.	IF
		Location μ		
Const	0.084	0.019	–	9.919
		Scale ϕ		
Const	0.402	0.035	–	7.125
LastDay	-0.190	0.120	0.036	0.903
LastWeek	**-0.738**	**0.193**	**0.985**	**18.519**
LastMonth	**-0.444**	**0.086**	**0.999**	**4.133**
CloseAbs95	0.194	0.233	0.035	1.445
CloseSqr95	0.107	0.226	0.023	2.715
MaxMin95	**1.124**	**0.086**	**1.000**	**6.012**
CloseAbs80	0.097	0.153	0.013	–
CloseSqr80	0.143	0.143	0.021	–
MaxMin80	-0.022	0.200	0.017	–
		Degrees of freedom ν		
Const	2.482	0.238	–	5.708
LastDay	0.504	0.997	0.112	2.899
LastWeek	**-2.158**	**0.926**	**0.638**	**5.463**
LastMonth	0.307	0.833	0.089	5.560
CloseAbs95	0.718	1.437	0.229	3.020
CloseSqr95	1.350	1.280	0.279	2.758
MaxMin95	1.130	1.488	0.222	6.564
CloseAbs80	0.035	1.205	0.101	2.789
CloseSqr80	0.363	1.211	0.112	3.330
MaxMin80	-1.672	1.172	0.254	4.178
		Skewness λ		
Const	-0.104	0.033	–	10.423
LastDay	-0.159	0.140	0.027	1.170
LastWeek	-0.341	0.170	0.135	8.909
LastMonth	-0.076	0.112	0.016	–
CloseAbs95	-0.021	0.096	0.008	–
CloseSqr95	-0.003	0.108	0.006	–
MaxMin95	0.016	0.075	0.008	–
CloseAbs80	0.060	0.115	0.009	–
CloseSqr80	0.059	0.111	0.010	–
MaxMin80	0.093	0.096	0.013	–

[8]The posterior mean, standard deviation and inefficiency factors (IF) are computed conditional on a covariate being in the model. The IFs are not computed for parameters with posterior probabilities smaller than 0.02.

LastWeek$\in [-1, 1]$). The last column of Table 4 gives the inefficiency factor (IF) for all parameters with inclusion probabilities larger than 0.02. It is clear that the MCMC algorithm is very efficient, almost all parameters have IFs smaller than 10. The MH acceptance probabilities for the regression coefficients in μ, ϕ, ν and λ are as high as 95%, 81%, 75% and 94%, respectively.

To explore the sensitivity to variations in the rather arbitrarily set prior parameter τ_β^2 (see Section 2.3), we compute the LPDS for the one-component split-t model using $\tau_\beta^2 = 1$, 10 and 100 (the default), obtaining an LPDS of -472.89, -472.61 and -472.37, respectively. Since the LPDS is based on the posterior distribution from a large sample (unlike the marginal likelihood which is based on the prior), this insensitivity to the prior is reassuring but not surprising. We also compare the posterior inference on the regression coefficients for the same three values of τ_β^2. The posterior means and standard deviations are very insensitive to changes in τ_β^2 while the posterior inclusion probabilities generally decrease with τ_β^2, but not to the extent of overturning the results about the importance of individual covariates. The effect of the prior on the inclusion probabilities is smaller for the covariates that almost certainly enter the model. As an example, the posterior inclusion probabilities for LastDay in ϕ is 0.290, 0.110 and 0.036 for $\tau_\beta^2 = 1$, 10 and 100, respectively, while for MaxMin95 they are 1.000, 0.999 and 1.000 for the same three priors. Interestingly, the only significant covariate in the degrees of freedom function, LastWeek, has posterior inclusion probabilities of 0.66, 0.76 and 0.64 in ν for the three different values of τ_β^2.

The LPDS is also fairly insensitive to the prior on the intercepts in Table 1. As an example the LPDS for the split-t model with two separate components changes from -466.84 to -466.86, -466.63 and -468.40 when we double the prior standard deviation of the intercept in ϕ, ν and λ, respectively.

Figure 4 presents box plots of the posterior distribution of the number of included parameters, i.e. $p(\sum_{k=1}^{K}(\sum_{q=1}^{Q}\sum_{p=1}^{P}\mathcal{I}_{kqp} + \sum\mathcal{I}_{mix}))$, where $\mathcal{I}_{kqp}$ is the Bayesian variable selection indicator for the pth variable in the qth parameter in the kth component density and $\sum\mathcal{I}_{mix}$ is the sum of variable selection indicators in the mixing function. Figure 4 shows that the SMR(4) has 26 effective parameters on average, while SAGM(1), which performs better than any SMR model, has only five effective parameters on average. Moreover, the one-component split-t model contains only four more effective parameters than SAGM(1), but the split-t model has much high LPDS. Figure 4 (bottom right) also shows that the proportion of posterior included parameters to potential parameters is close to 0.5 in the SAGM and split-t models with a large number of components. This result is in part a reflection of our choice of a Bernoulli(0.5) prior for the variable selection indicators. This prior implies that the prior on the number of effective parameters is a binomial distribution with mean $N/2$ and standard deviation $\sqrt{N/4}$, where N is the number of potential parameters in the model. For models with large N the prior is therefore fairly tightly centered on a large number of effective parameters. Other priors on the variable selection indicators are straightforward to implement, however, e.g. the uniform prior in Denison et al. (2002) or the hierarchical prior in Kohn et al. (2001).

To investigate the stability of the predictive densities for different sets of sample sizes we estimate the one-component split-t model using five samples with an increasing number of observations. The samples consist of the first 1000, 2000, 3000, 4000 trading days and then finally using the full sample between Jan 1, 1990 and Mar 13, 2009. Figure 5 displays the conditional predictive densities for the three sets of covariates values present on the 4648th, 4725th, and 4753th trading day where MaxMin95 is 0.2503, 0.9043, and 1.737, respectively (hence representing states of low, medium and high volatility). Figure 5 shows that the 1% quantiles (VaR) of the return distribution

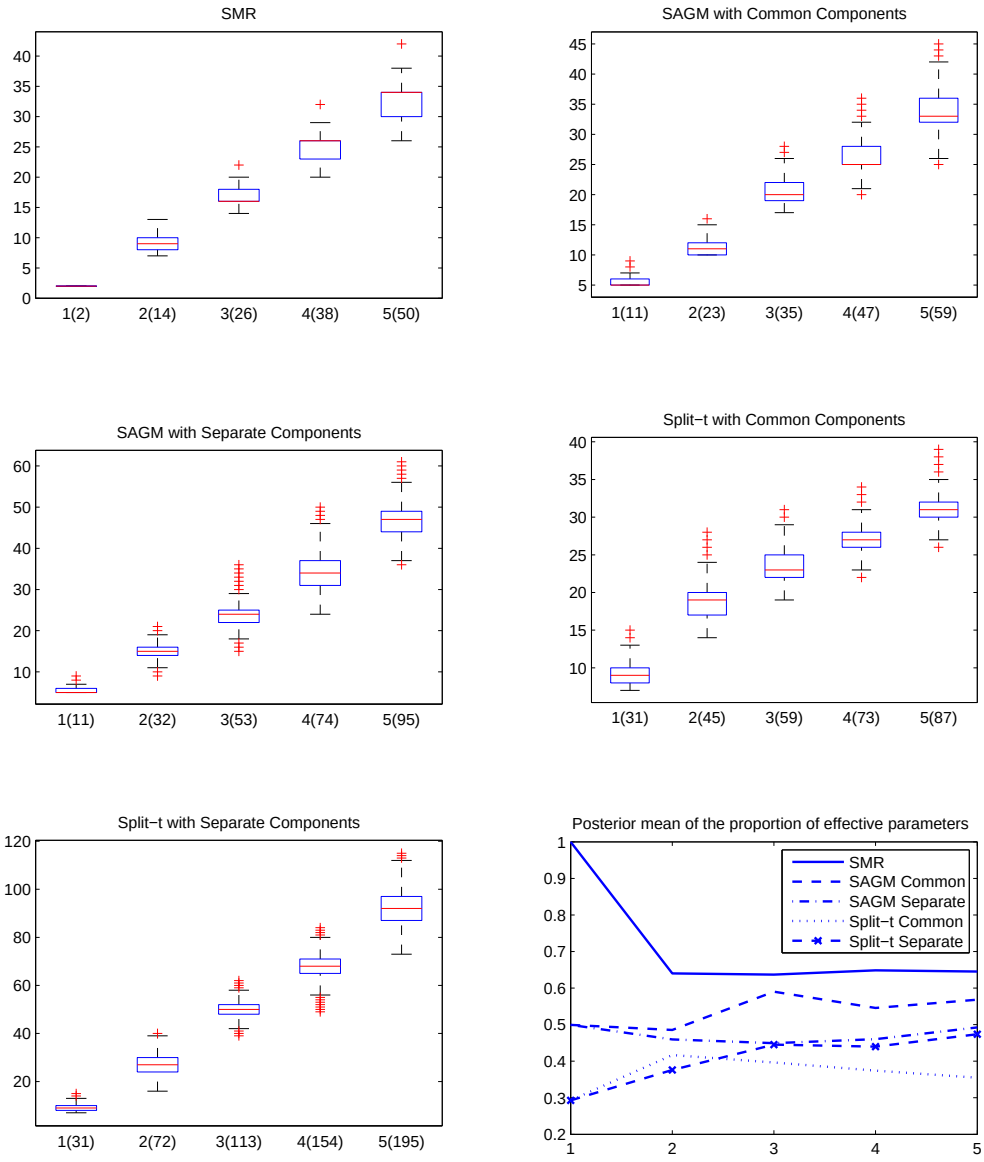

Figure 4: The posterior distribution of the number of included parameters. On first five subplots, the horizontal axis measures the number of components (with the number of potential parameters in parentheses) and the vertical axis is the total number of effective parameters in the model. All models are estimated using the S&P 500 data up to Mar 13, 2009. The right-bottom subplot is the posterior mean of the proportion of effective parameters in each model.

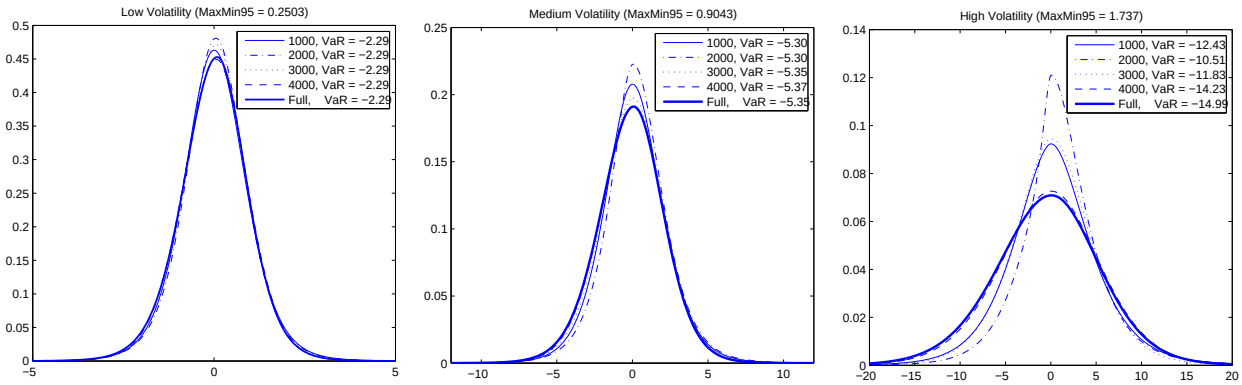

Figure 5: Investigating estimation stability over different subsamples. The subgraphs show predictive densities for different sets values on the covariates (low, medium and high volatility). The model is estimated on the first 1000, 2000, 3000, 4000 trading days starting from Jan 1, 1990 and the full sample between Jan 1, 1990 and Mar 13, 2009 using the one-component split-t model.

do not change significantly over five estimation samples.

Finally, Figure 6 presents some posterior moments, such as the standard deviation and skewness, for the one-component split-t model over the latter part of the sample (including the evaluation sample). The model is estimated on all available data up to March 13, 2009. Figure 6 shows that the median of the degrees of freedom actually increased during the most volatile part of the financial crisis (but at the same time the scale parameter rose dramatically to bring about a very large boost in standard deviation of returns), but, during some spells, the posterior distribution of ν also has a long left tail with substantial probability mass on very small values of ν.

5. Conclusions

A general model is presented for estimating the distribution of a continuous variable conditional on a set of covariates. The model is a mixture of asymmetric student t densities with the mixture weights and all four component parameters, location, scale, degrees of freedom and skewness, being functions of covariates. We take a Bayesian approach to inference and estimate the model by an efficient MCMC simulation method. Bayesian variable selection is carried out to obtain model parsimony and guard against over-fitting. The model is applied to analyze the distribution of daily stock market returns conditional on nine covariates and outperforms widely used GARCH models and other recently proposed mixture models in an out-of-sample evaluation of returns during the recent financial crisis.

Acknowledgments

We thank the editor and two anonymous referees for the helpful comments that improved the content and presentation of the paper. The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank. Kohn was partially supported by ARC Grant DP0667069.

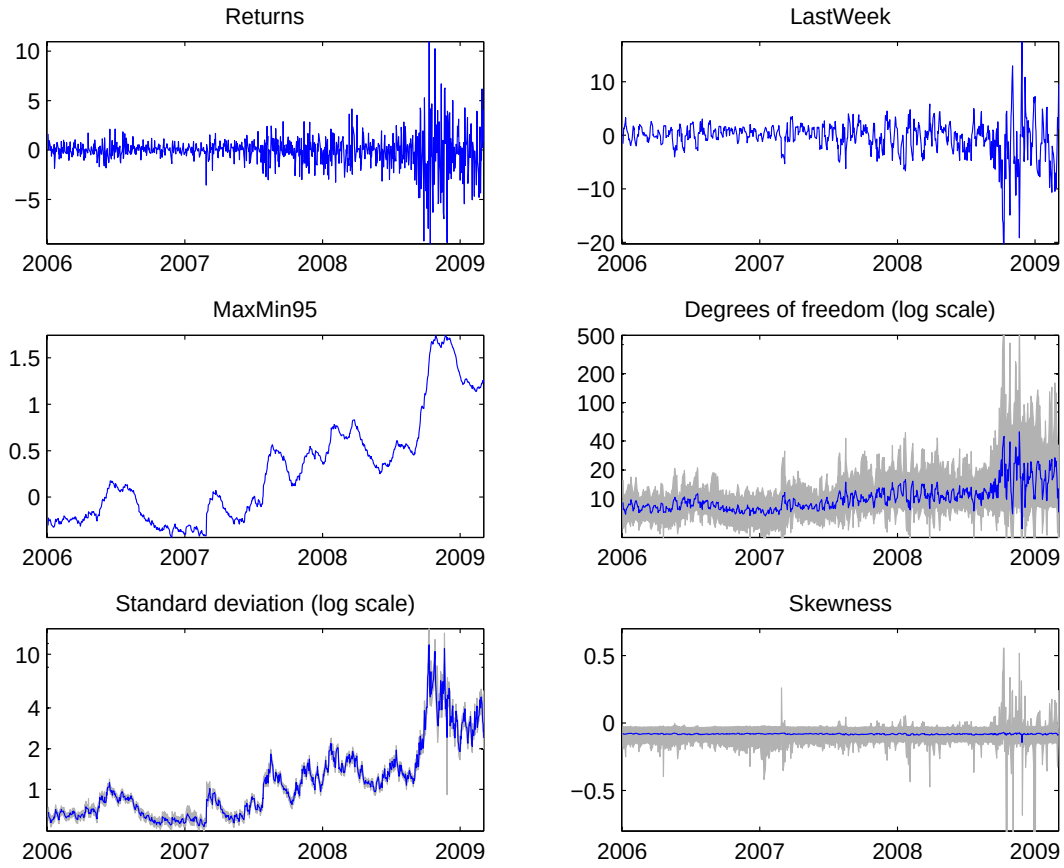

Figure 6: Time series plot of the posterior median and 95% probability intervals for some moments of the return distribution. The time series of returns and two of the key covariates are also plotted. The posterior distribution is based on the full sample up to March 13, 2009. The distribution of the standard deviation and the skewness are conditioned on $\nu > 2$ and $\nu > 3$, respectively.

References

Abramowitz, M., Stegun, I. (Eds.), 1972. Handbook of mathematical functions with formulas, graphs, and mathematical table. Courier Dover Publications, New York.

Alizadeh, S., Brandt, M., Diebold, F., 2002. Range-based estimation of stochastic volatility models. Journal of Finance 57 (3), 1047–1091.

Baillie, R. T., 2006. Handbook of Econometrics. Vol. 1. Palgrave Macmillan, New York, Ch. Modeling volatility, pp. 737–764.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business & Economic Statistics 19 (4), 465–474.

Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. The Review of Economics and Statistics 69 (3), 542–547.

Celeux, G., Hurn, M., 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., Smith, A., 2002. Bayesian methods for nonlinear classification and regression. Wiley, New York.

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

Engle, R., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society 50 (4), 987–1007.

Escobar, M., 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., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57 (6), 1317–1339.

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

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

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

Hansen, B., 1994. Autoregressive conditional density estimation. International Economic Review 35 (3), 705–730.

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

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

Jeffreys, H., 1961. Theory of probability, 3rd ed. Oxford, New York.

Jiang, W., 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., 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., 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.

Kass, R., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

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

Nott, D., Kohn, R., 2005. Adaptive sampling for Bayesian variable selection. Biometrika 92 (4), 747–763.

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

Richardson, S., 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., 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., Giordani, P., 2007. Nonparametric regression density estimation using smoothly varying normal mixtures. Sveriges Riksbank Working Paper Series, no. 211, Available at www.riksbank.com.
Villani, M., Kohn, R., Giordani, P., 2009. Regression density estimation using smooth adaptive Gaussian mixtures. Journal of Econometrics 153 (2), 155–173.
Wood, S., Jiang, W., Tanner, M., 2002. Bayesian mixture of splines for spatially adaptive nonparametric regression. Biometrika 89 (3), 513–528.
Zeevi, A., 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.

A. MCMC implementation

To implement the MCMC algorithm we need the gradient and Hessian matrix of the conditional posteriors for each of the four split-t parameters. Since the priors on the regression coefficients in each split-t parameter is a multivariate normal density, the prior gradient and Hessian matrix are

$$\frac{\partial \ln p(\beta)}{\partial \beta} = -\Sigma_\beta^{-1}(\beta - \mu_\beta) \text{ and } \frac{\partial^2 \ln p(\beta)}{\partial \beta \partial \beta'} = -\Sigma_\beta^{-1}.$$

To derive the gradient and Hessian matrix with respect to the likelihood, we write the likelihood as

$$p(y|x, \mu, \phi, \nu, \lambda) = \prod_{\mathcal{S}_1} t(y|\mu, \phi, \nu) \prod_{\mathcal{S}_2} t(y|\mu, \lambda\phi, \nu),$$

where $t(y|\mu, \phi, \nu)$ denotes the student-t density

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[\frac{\nu}{\nu + \frac{(y-\mu)^2}{\phi^2}}\right]^{(\nu+1)/2}.$$

$\mathcal{S}_1$ is the set of observations such that $y \leq \mu$ and $\mathcal{S}_2$ denotes the observations $y > \mu$. It is convenient to define the indicator function

$$I_\mu = \begin{cases} 1 \text{ if } y > \mu \\ 0 \text{ if } y \leq \mu \end{cases},$$

and $a = \lambda^{I_\mu}$.

The following subsections present the gradient and the Hessian for each split-t parameter.

Gradient and Hessian wrt μ

$$\frac{\partial}{\partial \mu} \ln p\left(y|\mu, \nu, \phi, \lambda\right) = \frac{(1+\nu)(y-\mu)}{\nu a^2 \phi^2 + (y-\mu)^2}$$
$$\frac{\partial^2}{\partial \mu^2} \ln p\left(y|\mu, \nu, \phi, \lambda\right) = \frac{(1+\nu)\left[(y-\mu)^2 - a^2\phi^2\nu\right]}{\left[(y-\mu)^2 + a^2\phi^2\nu\right]^2}.$$

Gradient and Hessian wrt ϕ

$$\frac{\partial}{\partial\phi}\ln p\left(y|\mu,\nu,\phi,\lambda\right)=\frac{\nu\left[\left(y-\mu\right)^2-a^2\phi^2\right]}{\phi\left[\left(y-\mu\right)^2+\nu a^2\phi^2\right]}$$
$$\frac{\partial^2}{\partial\phi^2}\ln p\left(y|\mu,\nu,\phi,\lambda\right)=\frac{\nu\left[\phi^4a^4\nu-\left(y-\mu\right)^4-\left(1+3\nu\right)\left(y-\mu\right)^2\phi^2a^2\right]}{\left[\phi\left(y-\mu\right)^2+\phi^3a^2\nu\right]^2}.$$

Gradient and Hessian wrt ν

$$\begin{aligned}\frac{\partial}{\partial\nu}\ln p\left(y|\mu,\nu,\phi,\lambda\right)=&\frac{\left(y-\mu\right)^2-\phi^2a^2}{2\left[\left(y-\mu\right)^2+\nu\phi^2a^2\right]}+\frac{1}{2}\ln\left(\frac{\nu}{\nu+\frac{\left(y-\mu\right)^2}{\phi^2a^2}}\right)\\&+\frac{1}{2}\left[\psi\left(\frac{\nu+1}{2}\right)-\psi\left(\frac{\nu}{2}\right)\right]\\\frac{\partial^2}{\partial\nu^2}\ln p(y|\mu,\nu,\phi,\lambda)=&\frac{\left(y-\mu\right)^4+\nu\phi^4a^4}{2\nu\left(\left(y-\mu\right)^2+\nu\phi^2a^2\right)^2}+\frac{1}{4}\left[\psi_1\left(\frac{\nu+1}{2}\right)-\psi_1\left(\frac{\nu}{2}\right)\right]\end{aligned}$$

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

Gradient and Hessian wrt λ

$$\frac{\partial}{\partial\lambda}\ln p\left(y|\mu,\nu,\phi,\lambda\right)=-\frac{1}{1+\lambda}+\frac{\left(1+\nu\right)\left(y-\mu\right)^2I_\mu}{\left(y-\mu\right)^2\lambda+\nu\phi^2\lambda^3}$$
$$\frac{\partial^2}{\partial\lambda^2}\ln p\left(y|\mu,\nu,\phi,\lambda\right)=\frac{1}{\left(1+\lambda\right)^2}-\frac{\left(1+\nu\right)\left(y-\mu\right)^2\left[\left(y-\mu\right)^2+3\nu\phi^2\lambda^2\right]I_\mu}{\left[\left(y-\mu\right)^2\lambda+\nu\phi^2\lambda^3\right]^2}.$$

Let $l(\cdot)$ denote a link function of any parameter in the split-t model, e.g. the function linking the degrees of freedom with the covariates as $l(\nu)=x'\beta_\nu$, so $\nu=l^{-1}(x'\beta_\nu)$. Using gradient, Hessian and (4), it is straightforward to link the derivatives of posterior density β with any of the split-t parameters ($l^{-1}(x'\beta)$) by applying the chain rule

$$\frac{\partial\ln(y|\mu,\nu,\phi,\lambda)}{\partial\beta}=\frac{\partial\ln(y|\mu,\nu,\phi,\lambda)}{\partial l^{-1}(x'\beta)}\frac{\partial l^{-1}(x'\beta)}{\partial\beta}$$
$$\frac{\partial^2\ln(y|\mu,\nu,\phi,\lambda)}{\partial\beta\partial\beta'}=\frac{\partial\ln(y|\mu,\nu,\phi,\lambda)}{\partial l^{-1}(x'\beta)}\frac{\partial^2l^{-1}(x'\beta)}{\partial\beta\partial\beta'}+\frac{\partial^2\ln(y|\mu,\nu,\phi,\lambda)}{\partial^2l^{-1}(x'\beta)}\frac{\partial l^{-1}(x'\beta)}{\partial\beta}\frac{\partial l^{-1}(x'\beta)}{\partial\beta'}.$$