HAL open science

Regression density estimation using smooth adaptive Gaussian mixtures

Mattias Villani, Robert Kohn, Paolo Giordani

To cite this version:

Mattias Villani, Robert Kohn, Paolo Giordani. Regression density estimation using smooth adaptive Gaussian mixtures. Econometrics, 2009, 153 (2), pp.155. 10.1016/j.jeconom.2009.05.004 . hal-00599456

HAL Id: hal-00599456

https://hal.science/hal-00599456v1

Submitted on 10 Jun 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript

Regression density estimation using smooth adaptive Gaussian mixtures

Mattias Villani, Robert Kohn, Paolo Giordani

PII: S0304-4076(09)00141-9
DOI: 10.1016/j.jeconom.2009.05.004
Reference: ECONOM 3207

To appear in: *Journal of Econometrics*

Received date: 5 November 2007
Revised date: 11 May 2009
Accepted date: 20 May 2009

Please cite this article as: Villani, M., Kohn, R., Giordani, P., Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* (2009), doi:10.1016/j.jeconom.2009.05.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

REGRESSION DENSITY ESTIMATION USING SMOOTH ADAPTIVE GAUSSIAN MIXTURES

MATTIAS VILLANI, ROBERT KOHN, AND PAOLO GIORDANI

Abstract. We model a regression density flexibly so that at each value of the covariates the density is a mixture of normals with the means, variances and mixture probabilities of the components changing smoothly as a function of the covariates. The model extends existing models in two important ways. First, the components are allowed to be heteroscedastic regressions as the standard model with homoscedastic regressions can give a poor fit to heteroscedastic data, especially when the number of covariates is large. Furthermore, we typically need fewer components, which makes it easier to interpret the model and speeds up the computation. The second main extension is to introduce a novel variable selection prior into all the components of the model. The variable selection prior acts as a self-adjusting mechanism that prevents overfitting and makes it feasible to fit flexible high-dimensional surfaces. We use Bayesian inference and Markov Chain Monte Carlo methods to estimate the model. Simulated and real examples are used to show that the full generality of our model is required to fit a large class of densities, but also that special cases of the general model are interesting models for economic data.

Keywords: Bayesian inference, Markov Chain Monte Carlo, Mixture of Experts, Nonparametric estimation, Splines, Value-at-Risk, Variable selection.
JEL: C11, C50.

1. Introduction

Nonlinear and nonparametric regression models are widely used in statistics (Ruppert, Wand and Carroll, 2003), and are increasingly used in econometrics (Li and Racine, 2007). Our article considers the general problem of *nonparametric regression density estimation*, i.e. estimating the predictive density of the response variable at all points in the covariate space, while making relatively few assumptions about its functional form and how that functional form changes across the space of covariates. This is an important problem in empirical

Villani: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden* and *Department of Statistics, Stockholm University. E-mail: mattias.villani@riksbank.se.* Kohn: *Australian School of Business, University of New South Wales, UNSW, Sydney 2052, Australia.* Giordani: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden.* We thank the editor, associate editor and two referees for extensive comments that helped improve 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. Villani was partially supported by a grant from the Swedish Research Council (Vetenskapsrådet, grant no. 412-2002-1007) and Kohn was partially supported by ARC grant DP0667069.

ACCEPTED MANUSCRIPT

economics, *e.g.* in the analysis of financial data where accurate estimation of the left tail probability is often the final goal of the analysis (Geweke and Keane, 2007), but also in many other areas, such as machine learning, where the predictive density is typically highly nonlinear and multimodal (Bishop, 2006).

Our approach generalizes the popular finite mixture of Gaussians model (McLachlan and Peel, 2000; Früewirth-Schnatter, 2006) to the regression density case. The model extends the *Mixture-of-Experts* (ME) model (Jacobs, Jordan, Nowlan and Hinton (1991); Jordan and Jacobs (1994)), which has been frequently used in the machine learning literature to flexibly model the mean regression. The ME model is a mixture of regressions where the mixing probabilities are functions of the covariates, leading to a partitioned covariate space with stochastic (soft) boundaries. A generalization of the ME model, the *Smoothly Mixing Regression* (SMR), was recently introduced in econometrics and used for regression density estimation by Geweke and Keane (2007).

The early machine learning literature used SMRs with many simple component regressions (constant or linear). Some recent statistical/econometric literature takes the opposite approach of using a small number of more complex component regressions. The most common approach is to use basis expansion methods (polynomials, splines) to allow for nonparametric component regressions, see *e.g.* Wood, Jiang and Tanner (2002) and Geweke and Keane (2007). One motivation of the few-but-complex approach comes from a growing awareness that mixture models can be quite challenging to estimate and interpret, especially when the number of mixture components is large (Celeux, Hurn and Robert (2000), Geweke (2007)). It is then sensible to make each of the components very flexible and to use extra components only when they are required.

Jiang and Tanner (1999a,b) prove that a smooth mixture of sufficiently many (generalized) linear regressions can approximate essentially any function or a single density in the exponential family with a constant dispersion parameter. Similarly, it is expected that the SMR should in principle be able to fit heteroscedastic data if the number of component regressions is large enough, but it is unlikely to be the most efficient model for that situation. Simulations in Section 4 show that the SMR model can have difficulties in modelling heteroscedastic data, and that its predictive performance quickly deteriorates as the number of covariates grows. If the component regressions themselves are heteroscedastic, we clearly need fewer of them.

Our article generalizes the SMR model by using Gaussian *heteroscedastic* regression components with the three parts of each component, i.e. the means, variances and the mixing probabilities, functions of the covariates. In the most general form of our model each of these three parts is modelled flexibly using spline basis function expansions. We take a Bayesian approach to inference with a prior that allows for variable selection among the covariates in

ACCEPTED MANUSCRIPT

the mean, variance and mixing probabilities. When using splines, the centering of the spline basis functions (knots) are therefore determined automatically from the data as in Smith and Kohn (1996), Denison, Mallick and Smith (1998) and Dimatteo, Genovese and Kass (2001). This is particularly important in soft partition models as it allows the estimation method to automatically downweight or remove basis functions from a regression in the region where the component regression has small probability. Such basis functions are otherwise poorly identified and may cause instability in the estimation. Moreover, since variable selection typically reduces the effective number of parameters at each iteration, it helps to make the Metropolis-Hastings (MH) steps computationally tractable. The variable selection prior we use for the component means and variances is novel because it takes into account the mixing probability of a component regression when deciding whether to include a basis function in that component. The variable selection prior is very effective at simplifying the model and in particular allows us to reach the linear homoscedastic model if such a model is warranted. The empirical illustrations on real and simulated data in Section 4 show that each aspect of our model may be necessary to obtain a satisfactory and interpretable fit of the predictive distribution. We use the cross-validated log of the predictive density for model comparison and for selecting the number of components in the model to reduce sensitivity to the prior.

Early Bayesian analyses of finite Gaussian mixtures are given in Diebolt and Robert (1994) and Escobar and West (1995). The first Bayesian paper on smooth mixtures is Peng, Jacobs and Tanner (1996) who used the random walk Metropolis algorithm to sample from the posterior. Wood et al. (2002) and Geweke and Keane (2007) propose more elaborate extensions of this model and devise more efficient inferential algorithms. Leslie, Kohn and Nott (2007) propose a model of the conditional regression density using a Dirichlet Process (DP) mixture prior whose components do not depend on the covariates. Green and Richardson (2001) discuss the close relationship between finite mixture models and DP mixtures. A more detailed discussion of these estimators is given in Section 2. An alternative approach to regression density estimation is given by De Iorio, Muller, Rosner and MacEarchen (2004), Dunson, Pillai and Park (2007) and Griffin and Steel (2007) who use a dependent DP prior. An attractive feature of this prior is that different partitions of the data can have differing numbers of components. To carry out the inference we develop efficient MCMC samplers that compare favorably to existing MCMC samplers for the special case of smooth homoscedastic mixtures.

2. Smooth Adaptive Gaussian Mixtures

2.1. **The model.** Regression density estimation entails estimating a sequence of densities, one for each covariate value, x. A single density can usually be modelled adequately by a

ACCEPTED MANUSCRIPT

finite mixture of Gaussians. For example, the simulations in Roeder and Wasserman (1997) suggest that mixtures with less than 10 components can model even highly complex univariate densities. To extend the basic mixture of Gaussians model to the regression density case we need to make the transition between densities smooth in x. We propose that the means, variances and the mixing probabilities of the mixture components vary smoothly across the covariate space according to the *Smooth Adaptive Gaussian Mixture (SAGM)* model

$$y_i|(s_i = j, v_i, w_i) \sim N[\alpha_j' v_i, \sigma_j^2 \exp(\delta_j' w_i)], \quad (i = 1, ..., n,\ j = 1, ..., m), \tag{2.1}$$

where $s_i \in \{1, ..., m\}$ is an indicator of component membership for the ith observation, v_i is a p-dimensional vector function of covariates for the conditional mean of observation i with coefficients α_j that vary across the m components, and w_i is an r-dimensional vector of covariates for the conditional variance of observation i. The responsibility of component j for the ith observation is modelled by a multinomial logit *mixing function*

$$\Pr(s_i = j|z_i) = \pi_j(z_i; \gamma) = \frac{\exp(\gamma_j' z_i)}{\sum_{k=1}^m \exp(\gamma_k' z_i)}, \tag{2.2}$$

where z_i is a q-dimensional vector function of covariates for observation i, and $\gamma_1 = 0$ for identification. The three sets of terms, v_i, w_i, and z_i can be (high-dimensional) basis expansions (polynomials, splines etc.) of other predictors. For example, basis expansion in the mixing function gives us the flexibility to vary the number of effective mixture components quite dramatically across the covariate space. In the case of splines, let κ_k^v, κ_k^w and κ_k^z denote the position of the kth knot in the mean, variance and mixing functions, respectively. We denote the original vector of covariate observations from which the terms (v_i, w_i, z_i) were constructed by x_i.

Bayesian inference for the parameter in the mixture components and the parameters in the mixing function are discussed in Section 3. We determine the number of mixture components, m, using an out-of-sample equivalent to the marginal likelihood, see Section 3.4 for details.

Many of the models in the nonparametric literature are special cases of the SAGM model in (2.1) and (2.2). For $m = 1$, SAGM reduces to the heteroscedastic spline regression in Ruppert et al. (2003). The model in Wood, Jiang and Tanner (2002) is the special case with $\delta_j = 0$ and $\sigma_j = \sigma$, for $j = 1, ..., m$. The model in Geweke and Keane (2007) is obtained if we set $\delta_j = 0$ for all j, and use polynomial expansions of the covariates. Both Wood et al. and Geweke and Keane (2007) use a multinomial probit mixing function with an identity covariance matrix for the random utilities. This means that the component probabilities must be computed by very time-consuming numerical integration. Both these articles use cleverly designed MCMC schemes where the π_j need not be evaluated in the posterior sampling, but evaluating predictive densities/likelihoods is still computationally

ACCEPTED MANUSCRIPT

very demanding (Geweke and Keane, 2007). This is clearly a drawback compared to the multinomial logit mixing function used in the SAGM, where the component probabilities are available in closed form. The model in Leslie et al. (2007) is a heteroscedastic regression with a nonparametric modelling of the disturbances using a Dirichlet process mixture prior. This can be viewed as a special case of the SAGM model with $\delta_j = \delta$ for all j, mixing probabilities that do not depend on x, and means and (log) variances of the component that differ by constants for all x. Bishop's (2006) mixture density network is a related model in the neural network field. The mixture density network model is more restrictive than the SAGM, see Bishop (2006) for details.

We will also allow for automatic variable selection in all three sets of covariates. Let $\mathcal{V}$ denote a $p \times m$ matrix of zero-one indicators for the mean covariates in v. If the element in row k, column j of $\mathcal{V}$ is zero, then the coefficient on the kth v-covariate in the jth component is zero ($\alpha_{kj} = 0$); if the indicator is one, then α_{kj} is unrestricted. This is best viewed as a two-component mixture prior for α_{kj} with one of the components degenerate at $\alpha_{kj} = 0$. Similarly, let $\mathcal{W}$ $(r \times m)$ and $\mathcal{Z}$ $(q \times m)$ denote the variable selection indicators for the variance and mixing functions, respectively.

We would like to emphasize that many problems encountered in economics will not require the full flexibility of the SAGM model. Linear components or a single nonparametric component may be sufficient, as in the US stock returns example in Section 4.4. Variable selection in principle simplifies the SAGM model to the required flexibility, but it may be a good strategy to remove unnecessary model features from the outset, at least to simplify posterior sampling and interpretation. There are two additional restrictions on the model that we have found very useful in practice. First, we may restrict the heteroscedasticity to be the same across components: $\delta_1 = \cdots = \delta_m = \delta$. The model will often be flexible enough even under this restriction, especially when the variance and/or the mixing function are nonparametric. Second, we may restrict the covariate selection indicators to be the same across components. That is, either a covariate has a non-zero coefficient in all of the components or its coefficient is zero for all components. Our posterior sampling algorithms handle all these restrictions. These two restrictions also allow us to interpret parts of the model without additional identifying assumptions (see the next paragraph and the US stock return example in Section 4.4).

Mixture models have well known identification issues, e.g. the likelihood is invariant with respect to permutations of the components in the mixture (label switching), see e.g. Celeux et al. (2000), Jasra, Holmes and Stephens (2005) and Früewirth-Schnatter (2006). We are mainly interested in the predictive density $p(y|x)$ for which label switching is neither a conceptual or a numerical problem (Geweke, 2007). On the few occasions that we interpret the

ACCEPTED MANUSCRIPT

mixture components, we identify the model with order restrictions on a subset of the components' parameters or by other restrictions (e.g. common variance function). Order restrictions should be used on parameters that are expected to differ appreciably between components, so this choice is problem specific. The inefficiencies in the MCMC due to order restrictions described in Celeux et al. (2000) were not manifested in our empirical applications in Section 4. Jasra, Holmes and Stephens (2005) surveys other ways of solving the identification problem.

In many applications interest centers on the first derivative of the mean function $E(y|x)$ with respect to the covariates. Ruppert et al. (2003, Sec. 6.8) give several examples, including the question of whether or not labor income eventually declines at older ages (the derivative of the mean function becomes negative). It is easy to show that the first derivative of the SAGM mean function, $E(y|x) = \sum_{j=1}^{m} \pi_j(z)\alpha_j' v$, is of the form

$$\frac{\partial}{\partial x}E(y|x) = \sum_{j=1}^{m} \pi_j(z) \left[\left(\frac{\partial z}{\partial x}\right)' \left[\gamma_j - \sum_{g=1}^{m} \pi_g(z)\gamma_g\right] \alpha_j' v + \left(\frac{\partial v}{\partial x}\right)' \alpha_j \right]. \tag{2.3}$$

The form of the matrices $\partial z/\partial x$ and $\partial v/\partial x$ is typically simple, see Ruppert et al. (2003, p. 153-154) for explicit matrix expressions for some commonly used spline functions. With linear components, $\partial z/\partial x$ and $\partial v/\partial x$ are simply selection matrices that extract subsets of covariates from x. The MCMC draws can be used in the usual way to obtain the posterior distribution of the first derivative. We return to the first derivative in Section 4.2, where it is used to define the persistence in a nonlinear time series model.

We use the following notation. Let $Y = (y_1, ..., y_n)'$ be the n-vector of responses, and $X = (x_1, ..., x_n)'$ the $n \times p_x$ dimensional covariate matrix. The covariates are standardized to have zero mean and unit variance to simplify the prior elicitation. Let $V = (v_1, ..., v_n)', W = (w_1, ..., w_n)'$ and $Z = (z_1, ..., z_n)'$ be the $n \times p$, $n \times r$ and $n \times q$ dimensional matrices of covariates expanded from X. Let $s = (s_1, ..., s_n)'$ denote the n-vector of component indicators for the full sample. Furthermore, define the $p \times m$ matrix of mean coefficients, $\alpha = (\alpha_1, ..., \alpha_m)$, and similarly the $r \times m$ matrix $\delta = (\delta_1, ..., \delta_m)$ with heteroscedasticity parameters. The corresponding disturbance variances are collected in $\sigma^2 = (\sigma_1^2, ..., \sigma_m^2)'$. Define $\gamma = (\gamma_2', ..., \gamma_m')'$ to be the $q(m-1)$ vector of multinomial logit coefficients.

2.2. **The prior distribution and variable selection.** We adopt a Bayesian approach to inference with a prior that decomposes as

$$p(\alpha, \sigma^2, \delta, \gamma, s, \mathcal{V}, \mathcal{W}, \mathcal{Z}) = p(\alpha, \sigma^2, \mathcal{V} \mid \gamma)p(\delta, \mathcal{W}|\gamma)p(\gamma, \mathcal{Z}, s).$$

The conditioning on γ in the priors for $(\alpha, \sigma^2, \mathcal{V})$ and $(\delta, \mathcal{W})$ comes from our specific variable selection prior described below. Consider first $p(\alpha, \sigma^2, \mathcal{V} \mid \gamma)$. We assume *a priori* that the coefficients are independent between components. Let $\mathcal{V} = (\mathcal{V}_1, ..., \mathcal{V}_m)$, where $\mathcal{V}_j$ contains the

ACCEPTED MANUSCRIPT

variable selection indicators for the jth component. Let $\alpha_{\mathcal{V}_j}$ and $\alpha_{\mathcal{V}_j^c}$ denote the subvectors of α_j with non-zero coefficients and zero coefficients, respectively. The prior for component j is

$$\begin{aligned} \sigma_j^2 &\sim IG(\psi_1, \psi_2) \\ \alpha_{\mathcal{V}_j}|\mathcal{V}_j, \sigma_j^2 &\sim N(0, \tau_\alpha^2 \sigma_j^2 I) \end{aligned} \tag{2.4}$$

where IG denotes the inverse Gamma distribution with shape parameter ψ_1 and scale parameter ψ_2, and $\alpha_{\mathcal{V}_j^c}|\mathcal{V}_j$ is identically zero. The exact choice of prior hyperparameters is discussed below. The prior in (2.4) assumes that the elements of $\alpha_{\mathcal{V}_j}$ are *a priori* independent. An alternative is to use a g-prior with covariance matrix $\tau_{\alpha_j}^2 \sigma_j^2 (V'V)^{-1}$. Whether an independent prior or a g-prior is preferred for nonparametric regression models is still an open question, see e.g. the discussion in Denison, Holmes, Mallick and Smith (2002, p. 80-81), and clearly the answer depends also on the choice of basis for the spline. An additional complication with using the g-prior for regression mixtures is that $V'V$ is a global measure of precision while regression components are typically local in covariate space. The matrix $V'V$ may therefore be a poor representation of the precision for an individual regression component. A better choice would be $V_j'V_j$, where V_j is the covariate matrix for the observations allocated to component j. This prior is obviously conditional on the component allocation, which in turn would make the elements of s dependent in the full conditional posterior, even if we condition on all other model parameters. This greatly complicates the posterior sampling of s, so this prior will not be used here. Our preferred approach is to standardize and demean the covariates, and assume prior independence between regression coefficients. This compromise is likely to work well across a large variety of data sets. We obtained slightly better performance (in terms of fit and MCMC convergence) when the covariates are standardized to the interval $[-1, 1]$, rather than to a unit standard deviation.

The prior for variable inclusion/exclusion has a novel form to deal with a problem that has gone unnoticed in the literature on smooth mixtures. An *a priori* positioning of a knot at location κ in covariate space runs the risk that one of the components may have very low probability in the neighborhood of that point ($\pi_j(\kappa; \gamma) \approx 0$ for at least some j). The coefficient for the knot of the component is then poorly estimated, or may even be unidentified. This may not necessarily have a large impact on the fit (since π_j is small), but keeping the coefficients of these knots unrestricted makes the MCMC algorithms a lot less efficient and complicates interpretation. To deal with this problem, we use the prior

$$\mathcal{V}_{kj}|\gamma \sim Bern[\omega_\alpha \pi_j(\kappa_k^v; \gamma)], \qquad (k = 1, ..., p; \;\; j = 1, ..., m), \tag{2.5}$$

ACCEPTED MANUSCRIPT

where $Bern()$ denotes the Bernoulli distribution, $0 \leq \omega_\alpha \leq 1$, and $\mathcal{V}_{kj}$ are assumed to be *a priori* independent conditional on γ. Note how the prior inclusion probability decreases as the components's responsibility for the knot decreases. In the limit where the jth component has zero responsibility for κ_k^v, that knot is automatically excluded from component j with probability one. The variable indicators for covariates other than those generated by the knots have prior $Bern(\omega_\alpha)$. It is possible to estimate ω_α as in Kohn, Smith and Chan (2001) with an extra MH step.

The prior on the variance function is essentially of the same form as the prior on the mean function:

$$\begin{aligned} \delta_{\mathcal{W}_j}|\mathcal{W}_j &\sim N(0, \tau_\delta^2 I) \\ \mathcal{W}_{kj}|\gamma &\sim Bern[\omega_\delta \pi_j(\kappa_k^w; \gamma)], \quad (k = 1, ..., r,\ j = 1, ..., m). \end{aligned}$$

The variance function has so far been parametrized as $\sigma_j^2 \exp(\delta_j' w_i)$. In the important special case when $\delta_1 = \cdots = \delta_m = \delta$, we draw the $(\sigma_1^2, ..., \sigma_m^2)$ and δ in separate blocks in the posterior sampling algorithm in Section 3. But when either $m = 1$ or when the δ_j are not equal across components, it is more efficient for the posterior sampling to parametrize the variance function as $\exp(\bar{\delta}_j' \bar{w}_i)$, where $\bar{w} = (1, w)$, $\bar{\delta}_j = (\delta_0, \delta_j)$, and $\delta_0 = \ln \sigma_j^2$, because then only $\bar{\delta}_j$ is sampled and there is no additional $(\sigma_1^2, ..., \sigma_m^2)$ block. We will use the prior $\delta_0 \sim N(\mu_{\delta_0}, \tau_{\delta_0}^2)$, with μ_{δ_0} and $\tau_{\delta_0}^2$ chosen so that $\sigma_j^2 = \exp(\delta_0)$ has the same mean and variance as in the $IG(\psi_1, \psi_2)$ prior above.[1] It is not crucial for the posterior sampling algorithms to have a prior for $\bar{\delta}_j$ in multivariate normal form, but it is convenient and the above formulation has the additional advantage that the prior is always specified by eliciting the mean and degrees of freedom of σ_j^2.

The prior on the mixing function decomposes as

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

The variable indicators in $\mathcal{Z}$ are assumed to be *iid* $Bern(\omega_\gamma)$. The prior on γ is assumed to be of the form

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

and $\gamma_{\mathcal{Z}^c} = 0$ with probability one. Finally, $p(s|\gamma, \mathcal{Z})$ is given by the multinomial logit model in (2.2).

The user needs to specify the prior hyperparameters ψ_1, ψ_2, τ_α, τ_δ, τ_γ, ω_α, ω_δ and ω_γ. It is clearly impossible to have a prior that can handle every conceivable data set. Nevertheless, it is our experience that the following prior is reasonable for a wide range of problems in

[1]Some algebra shows that $\tau_{\delta_0}^2 = \ln[(\psi_1 - 1)/(\psi_1 - 2)]$ and $\mu_{\delta_0} = \ln[\psi_2/(\psi_1 - 1)] - \tau_{\delta_0}^2/2$ imply the same mean and variance for σ^2 as in the $IG(\psi_1, \psi_2)$ distribution.

ACCEPTED MANUSCRIPT

economics, at least as a good starting point for more carefully elicited problem-specific priors. Our default prior choice for the prior inclusion probabilities is $\omega_\alpha = \omega_\delta = \omega_\gamma = \omega = 0.5$ for a linear variable and $\omega = 0.2$ for knots. It should be noted that while ω is important for the absolute level of the posterior inclusion probability of a particular variable or knot, it has no effect on the relative importance of variables/knots and typically does not matter very much for the predictive density. The smoothing parameters τ_α, τ_δ and τ_γ are all set to 10. The original variables are standardized to the interval $[-1, 1]$, so these choices give diffuse, but proper, priors. Again, this choice clearly affects the absolute posterior inclusion probabilities, but typically has little bearing on the model's predictive density. Finally, our default prior for the σ_j^2 is mildly data-based, with the degrees of freedom ψ_1 set to 3, and ψ_2 is implicitly set so that the prior expected values of σ_j^2 is μ, a pre-specified constant. We use three different ways to compute μ depending on our prior beliefs about the relationship between y and the v-covariates: i) Weak or no relationship - μ is set equal to the variance of y in the training sample, ii) Strong linear relationship - μ is set equal to the residual variance from a linear regression on the v-covariates, and iii) Strongly non-linear relationship - Divide the data into m clusters (by k-means clustering), fit a linear regression in each cluster and set μ equal to the smallest residual variance in all clusters. The smallest residual variance is chosen because it is much worse to have μ substantially larger than any of the σ_j^2 than it is to set μ much smaller than all σ_j^2. The reason is that the left tail of the inverse gamma density dies off very quickly so that setting μ too high is likely to lead to severe overestimation of the smallest σ_j. For most problems in economics, options i) or ii) are sufficient, and the difference between the two is minor.

3. MCMC sampling for the SAGM model

3.1. **Algorithms.** We use MCMC methods to sample from the joint posterior distribution of the model parameters. With a model as elaborate as the SAGM, it is crucial to use a very efficient posterior sampling algorithm. The algorithms presented here draw the model coefficients and do variable selection in tandem. We have experimented with several algorithms, and we now outline two efficient algorithms, leaving the details to Appendix A and B.

The first algorithm is a Metropolis-within-Gibbs sampler that draws parameters using the following four blocks: i) α, σ^2 and $\mathcal{V}$, ii) δ and $\mathcal{W}$, iii) γ and $\mathcal{Z}$, and iv) $s = (s_1, ..., s_n)$. Sampling the first block is straightforward since, conditional on s and δ, the model reduces to m independent linear (spline) regressions with known heteroscedasticity. Since δ is known we can easily transform the regressions to be homoscedastic and apply the sampling method of Smith and Kohn (1996) to draw α, σ^2 and $\mathcal{V}$. The full conditional posteriors of $(\delta, \mathcal{W})$ and $(\gamma, \mathcal{Z})$ are both non-standard and cannot be sampled directly. The next section describes

ACCEPTED MANUSCRIPT

a general method that generates highly efficient MH proposals for these parameter blocks. Finally, the elements of s are independent conditional on the model parameters and can therefore be drawn simultaneously.

The second algorithm simulates from the joint posterior using the following marginal-conditional decomposition

$$p(\alpha, \sigma^2, \gamma, s, \delta|Y, X) = p(\alpha, \sigma^2|Y, X, \gamma, s, \delta)p(\gamma, s, \delta|Y, X).$$

This is possible since α and σ^2 can be integrated out once we condition on s, and hence $p(\gamma, s, \delta|Y, X)$ is available in closed form. One can then sample from $p(\gamma, s, \delta|Y, X)$ by a three-block Metropolis-Hastings algorithm and subsequently use these draws to generate from $p(\alpha, \sigma^2|Y, X, \gamma, s, \delta)$ by direct simulation. This simulation is straightforward and details of sampling from $p(\gamma, s, \delta|\mathcal{D})$ are given in Appendix B. We present the algorithm for a fixed set of covariates, but the extension to covariate selection is exactly as for the Gibbs sampler if $\mathcal{V}, \mathcal{W}$ and $\mathcal{Z}$ are simulated in the (γ, s, δ)-block. Following Liu, Wong and Kong (1995) we refer to this approach as a collapsed algorithm. Collapsed algorithms, where some parameters are integrated out, are typically more efficient than pure Gibbs sampling (Liu, Wong and Kong, 1995). The main drawback is that the n elements of s are no longer conditionally independent once α, σ^2 are integrated out, and therefore cannot be drawn simultaneously. This means that a single update of s requires computing marginal likelihoods for nm Gaussian regressions and makes the collapsed scheme very time consuming for moderate and large n. Fortunately, the change from one marginal likelihood computation to the next consists of a simple re-allocation of a single observation from one component to another and we use rank-one Choleski updates to speed up the computations as described in Appendix B. This is typically too slow anyway and we consider using a MH step for drawing s to further speed up computing time of the collapsed sampler. A MH step reduces computations from an $O(nm)$ operation to a $O(2n_s)$ operation, where n_s is the number of observation for which we propose a change of s_i, as it is unnecessary to compute marginal likelihoods for observations whose allocations are unchanged. We consider two different kinds of proposals: i) proposing s from the mixing function, and ii) adaptive proposals from the empirical distribution of s. Nott and Kohn (2005) prove that this type of adaptation produces draws that converge in distribution to the target distribution.

3.2. **Variable dimension Newton proposals.** We now describe a general method for constructing tailored proposal densities for the Metropolis-Hasting algorithm with variable selection. The method was introduced by Gamerman (1997) for generalized linear models within the exponential family, and extended by Nott and Leonte (2004) to handle variable selection. Their algorithm is presented here in a more general setting which is not restricted to the

ACCEPTED MANUSCRIPT

exponential family. We first briefly sketch the algorithm without variable selection. Let θ be a vector of parameters with a non-standard posterior density $p(\theta|y)$ of the form

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

where $\varphi_i = X_i\theta$ and X_i is a covariate matrix for the ith observation. As an example, if $\theta = \delta$ is the vector of parameters in the SAGM variance function, then $\varphi_i = w_i'\delta$. Note that $p(\theta|y)$ may be a conditional posterior density and the algorithm can then be used as a step in a Metropolis-within-Gibbs algorithm. Assume also that the gradient and Hessian of the log posterior are available in closed form, or that numerical derivatives are computationally practical. We can use Newton's method to iterate K steps from the current point θ_c toward the mode of $p(\theta|y)$, to obtain $\hat{\theta}$ and the Hessian at $\hat{\theta}$. Note that $\hat{\theta}$ may not be the mode but is typically close to it already after a few Newton iterations, so setting $K = 1, 2$ or 3 is usually sufficient. This makes the algorithm very fast. Moreover, we can speed up the algorithm by computing the gradient and Hessian on a (random) subset of the data in each iteration. The Hessian can also be replaced with its expected value

$$E\left[\frac{\partial^2 \ln p(\theta|y)}{\partial\theta\partial\theta'}\right]$$

in the Newton iterations. This typically improves numerical stability, with only a slightly worse approximation of $p(\theta|y)$. The proposal is now drawn from the multivariate t-distribution with $\zeta > 2$ degrees of freedom:

$$\theta_p|\theta_c \sim t\left[\hat{\theta}, -\left(\frac{\partial^2 \ln p(\theta|y)}{\partial\theta\partial\theta'}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}, \zeta\right],$$

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

Consider now the variable selection case. The p-dimensional parameter vector θ is then augmented by a vector of binary covariate selection indicators $\mathcal{J} = (j_1, ..., j_p)$, and we propose θ and $\mathcal{J}$ simultaneously using the following decomposition

$$g(\theta_p, \mathcal{J}_p|\theta_c, \mathcal{J}_c) = g_1(\theta_p|\mathcal{J}_p, \theta_c)g_2(\mathcal{J}_p|\theta_c, \mathcal{J}_c),$$

where θ_p and $\mathcal{J}_p$ are the proposed iterates, $\theta_c, \mathcal{J}_c$ are the current iterates, g_2 is the proposal distribution for $\mathcal{J}$ and g_1 is the proposal density for θ conditional on $\mathcal{J}_p$. The Metropolis-Hasting acceptance probability is

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

The proposal density at the current point $g_1(\theta_c|\mathcal{J}_c, \theta_p)$ is a multivariate t-density with mode $\hat{\theta}_R$ and covariance matrix equal to the negative inverse Hessian evaluated at $\hat{\theta}_R$, where $\hat{\theta}_R$ is

ACCEPTED MANUSCRIPT

the point obtained by iterating K steps with the Newton algorithm, this time starting from θ_p. A simple way to propose $\mathcal{J}_p$ is to randomly pick a small subset of $\mathcal{J}_p$ and then always propose a change of the selected indicators (Metropolized move). This proposal can be refined in many ways, using e.g. the adaptive scheme in Nott and Kohn (2005), where the history of $\mathcal{J}$-draws is used to adaptively build up a proposal for each indicator. It is important to note that θ_c and θ_p may now be of different dimensions, so the original Newton iterations no longer apply. We will instead generate θ_p using the following generalization of Newton's method. Let X_{ic} denote the matrix of included covariates at the current draw (i.e. selected by $\mathcal{J}_c$), and let $\varphi_{ic} = X_{ic}\theta_c$ denote the corresponding functional. Also, let $\varphi_{ip} = X_{ip}\theta_p$ denote the same functional for the proposed draw, where X_{ip} is the matrix of covariates in the proposal draw. We exploit the idea that when the parameter vector θ changes dimensions, the dimensions of the functionals $\varphi_{ic} = X_{ic}\theta_c$ and $\varphi_{ip} = X_{ip}\theta_p$ stay the same, and the two functionals are expected to be quite close. A generalized Newton update is

$$\theta_{k+1} = A_k^{-1}(B_k\theta_k - g_k), \qquad (k = 0, ..., K-1), \tag{3.1}$$

where $\theta_0 = \theta_c$, and

$$\begin{aligned} g_k &= \sum_{i=1}^{n} X_{ip}' \frac{\partial \ln p(y_i|\varphi_i)}{\partial \varphi_i} + \frac{\partial \ln p(\theta)}{\partial \theta} \\ A_k &= \sum_{i=1}^{n} X_{ip}' \frac{\partial^2 \ln p(y_i|\varphi_i)}{\partial \varphi_i \partial \varphi_i'} X_{ip} + \frac{\partial^2 \ln p(\theta)}{\partial \theta \partial \theta'} \\ B_k &= \sum_{i=1}^{n} X_{ip}' \frac{\partial^2 \ln p(y_i|\varphi_i)}{\partial \varphi_i \partial \varphi_i'} X_{ic} + \frac{\partial^2 \ln p(\theta)}{\partial \theta \partial \theta'}, \end{aligned}$$

all evaluated at $\theta = \theta_k$. For the prior gradient this means that $\partial \ln p(\theta)/\partial\theta$ is evaluated at θ_k, including all zero parameters, and that the subvector conformable with θ_{k+1} is extracted from the result. The same applies to the prior Hessian (which does not depend on θ however, if the prior is Gaussian). After the first Newton iteration the parameter vector no longer changes in dimension, and the generalized Newton algorithm in (3.1) reduces to the original Newton algorithm. The proposal density $g_1(\theta_p|\mathcal{J}_p, \theta_c)$ is again taken to be the multivariate t-density in exactly the same way as in the case without covariate selection. Once the simultaneous update of the $(\theta, \mathcal{J})$-pair is completed, we make a final update of the non-zero parameters in θ, conditional on the previously accepted $\mathcal{J}$, using the fixed dimension Newton algorithm.

3.3. **Comparison of MCMC algorithms.** We now compare five MCMC algorithms for the SAGM: i) the Gibbs sampler in Appendix A (*Gibbs*), ii) *Collapsed-Full*, the collapsed sampler where s_i is drawn directly from its posterior conditional on γ, s_{-i}, δ, but with α and σ^2 integrated out, iii) *Collapsed-Mixing*, where each s_i is proposed from the mixing function, iv)

ACCEPTED MANUSCRIPT

Collapse-Adaptive, and v) *Auxiliary*, where the parameters in the mixing function are sampled conditional on auxiliary unobserved utilities as in Geweke and Keane (2007). The Geweke-Keane approach is very elegant and fast (roughly 4-5 times faster than the Gibbs sampler in Appendix A), but introducing auxiliary variables comes at the cost of inferior mixing (Holmes and Knorr-Held, 2006; Del Moral, Doucet and Jasra, 2007). It should be noted that the Geweke-Keane approach is based on the probit mixing function, but it typically gives a nearly identical predictive density as the one obtained with the logit mixing function. Villani, Kohn and Giordani (2007) computes the inefficiency factors for all the above mentioned algorithms for the LIDAR data set, which is often used for evaluating non-parametric fitting methods (see e.g. Ruppert et al., 2003). They find that the introduction of auxiliary variables inflates the inefficiency factors for the parameters in the mixing function, and that some of that inefficiency spills over to other model parameters.

Mixture models can have several local maxima in the likelihood function, even excluding label switching issues. A slowly mixing posterior sampling algorithm would then not only sample the posterior inefficiently, but also be more likely to get stuck in a local maximum. To learn more about this, we simulated SAGM data and recorded how fast the different sampling algorithms converged to a neighborhood in the vicinity of the global posterior mode. We generated 25 data sets of size $n = 200$ from a SAGM(3) model with two covariates (sampled uniformly in the unit plane). For each data set we ran all five posterior samplers five times each, every time with a new seed for the random number generator. The parameters in the data generating model are $\alpha_1 = (0, 0, 0)'$, $\alpha_2 = (0, 1, 1)'$, $\alpha_3 = (0, -1, -1)'$, $\sigma = (0.05, 0.05, 0.05)'$, $\gamma_2 = (10, -10)'$ and $\gamma_3 = (0, 10)'$. The mixing function parameters are intentionally chosen to quite sharply separate the three regression components. We define the *posterior mode region* as the region in parameter space where the difference in log posterior from the log posterior at the mode is smaller than 10. The posterior mode is taken to be the parameter draw with highest posterior density over all 25 runs of the algorithms for a given data set. As mentioned above, we need to resort to numerical integration to compute the probit mixing probabilities in the Geweke-Keane model. This is extremely time-consuming in a simulation study, so we take a short-cut and use the logit mixing function when evaluating the predictive density of the Geweke-Keane model. We verified that the results are very similar to using a probit mixing function, and the generous definition of the posterior mode region (difference in the log posterior is smaller than 10 units) makes sure that the Geweke-Keane sampler is not falsely classified as non-converging as a result of this approximation, and vice versa.

Columns 2-5 in Table 1 (under the heading 'separated components') report the percentage of posterior sampling runs that had not yet visited the posterior mode region after a certain

ACCEPTED MANUSCRIPT

number of MCMC iterations. No burn-in is used here. It is clear that the first four samplers converge quickly to the right region, but that the auxiliary sampler gets stuck in a minor mode for a long time, and in 3.52% of the runs it does not reach the posterior mode region in 10,000 draws. The two last columns in Table 1 report the percentage of mis-classified observations after 10,000 draws (using the posterior mode classification rule) and the computing time of the algorithm. The auxiliary sampler is almost five times faster than the Gibbs sampler.

We repeated the simulations above, this time with $\alpha_j = (0, 0, 0)'$ for all j, $\sigma = (5, 1, 0.1)'$ and $\gamma_2 = (-1, 0, 0)'$ and $\gamma_3 = (1, 0, 0)'$. This model is a scale mixture of three normals that generates heavy tailed data. Here the components are not at all separated. The lower portion of Table 1 gives the results. In this setting, all five algorithms perform well, almost all runs reached the posterior mode region already after 500 draws. The Gibbs sampler failed to converge in one of the runs. Collapse-Adapt has an impeccable performance: all runs reach the posterior mode region in less than 100 draws. The classification problem was naturally much more difficult with completely overlapping components, and misclassification rate are indeed larger.

Our preferred algorithm is the Gibbs sampler which is used in the rest of the article. It provides the best balance of efficiency and computing time, especially for large data sets and when the time to evaluate the predictive density is also taken into consideration.

3.4. **Model Comparison.** The marginal likelihood is typically used in Bayesian model comparisons, see e.g. Früewirth-Schnatter (2006) for a discussion of estimators in mixture models. It is well known however that the marginal likelihood is very sensitive to the choice of prior, especially 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, we 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 $\tilde{y}_b = (y_{n+1}, ..., y_{n+n_b})$ contains the n_b observations in the bth test sample and $\tilde{y}_{-b} = (y_1, ..., y_n)$ denotes the remaining observations. If we assume that the observations are independent conditional on θ, then

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

ACCEPTED MANUSCRIPT

Separated components						
Sampler	> 500	> 1000	> 2500	> 10000	% Misclassified	CPU time sec.
Gibbs	2.35	0.00	0.00	0.00	6.01	22.26
Collapse-Full	3.52	1.17	0.00	0.00	5.48	252.43
Collapse-Mixing	4.70	0.00	0.00	0.00	5.56	31.20
Collapse-Adapt	1.17	1.17	0.00	0.00	5.51	58.12
Auxiliary	92.94	63.52	23.52	3.52	6.77	4.56
Scale mixture						
Sampler	> 100	> 500	> 1000	> 10000	% Misclassified	CPU time sec.
Gibbs	4.80	2.40	1.60	0.80	15.56	22.26
Collapse-Full	0.80	0.80	0.80	0.00	15.71	252.43
Collapse-Mixing	4.00	0.00	0.00	0.00	15.84	31.20
Collapse-Adapt	0.00	0.00	0.00	0.00	15.77	58.12
Auxiliary	39.20	1.60	0.80	0.00	16.20	4.56

TABLE 1. Simulations from a two-covariate SAGM(3) with two different parameter settings: a) separated components (upper half of the table) and b) scale mixture of normals (lower half of the table). The table displays the proportion of full simulation runs that needed more than a certain number of MCMC iterations to reach the posterior mode region. The last two columns give the percentage of misclassified observations from a posterior mode classification rule and the computing time. The Collapse-Adapt sampler starts adapting after 500 initial draws with the Collapse-Full algorithm.

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 $p(\theta|\tilde{y}_{-b})$.

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

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

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. In the situation where T is fairly

ACCEPTED MANUSCRIPT

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 US inflation and US stock returns examples in Sections 4.2 and 4.4.

One way to calibrate the LPDS is to transform a difference in LPDS between two competing models into a Bayes factor. One can then use the well-known rule-of-thumb for Bayes factors of Jeffreys (1961) to assess the strength of evidence. We note however that the original Bayes factor evaluates all the observations, whereas the cross-validated LPDS is an average over the B test samples. The Bayes factor is therefore roughly B times more discriminatory than the LPDS; this is the price paid by the LPDS for using most of the data to train the prior. Other authors have proposed summing the log predictive density over the B test samples (see Geisser and Eddy (1979) for the case with $B = n$, and Kuo and Peng (2000) for $B < n$), which would multiply any LPDS difference by a factor B. We choose not to do so as the LPDS can then no longer be calibrated by Jeffreys scale of evidence.

4. Empirical illustrations

In this section we analyze three real data sets, and study the performance of the SAGM in a simulation study. Unless otherwise stated, the reported results were generated from $10,000$ draws after discarding $2,000$ burn-in draws. We use ten degrees of freedom in multivariate-t Newton-based proposal densities for δ and γ. The proposals for δ were generated with $K = 1$ Newton step with a Hessian equal to its expected value, whereas $K = 3$ Newton steps seemed to be a better default value for γ. At every iteration of the algorithm, the probability of updating a given variable selection indicator was set to 0.2. The component allocation is initialized with the k-means clustering algorithm with m clusters. The remaining parameters of the model were initialized with GLS estimates conditional on the initial component allocation. All computations were performed using uncompiled Matlab 7.6.0 code on a HP6910 laptop with an Intel 2 GHz processor and 3GB of RAM memory.

4.1. **Inverse problem.** Our first example is based on an a inverse problem in robotics discussed by Bishop (2006). Suppose that for a given y, $x = y + 0.3\sin(2\pi y) + u$, where u is $U(0, 1)$. We generate 1000 values of x_i by taking the y_i to be equally spaced on $[0, 1]$ and the u_i independent and uniform. The resulting data set is plotted in the left column of Figure 1. From an econometrician's point of view, the data are highly non-standard, and are used here to demonstrate some features of the SAGM model. We wish to estimate the density of $p(y|x)$. This is a challenging regression density estimation problem as the density $p(y|x)$ is heteroscedastic and multimodal for some x.

ACCEPTED MANUSCRIPT

The extreme features of this data set makes it necessary to depart from the default prior in Section 2.2 along one dimension. The rapidly changing shape of Bishop's data over the covariate space requires very large mixing function coefficients, and the default prior with $\tau_\gamma = 10$ shrinks too much. We therefore set $\tau_\gamma = 1000$, but keep all other aspects of the default prior (with the third option for specifying the prior mean of σ_j^2). We used truncated quadratic splines (see e.g. Ruppert et al., 2003) with 20 equally spaced knots in the mean, variance and mixing function. This means that the SAGM model is very richly parametrized even if we assume a common variance function. As an example, the SAGM(3) model has $3 \cdot 22 = 66$ mean parameters, $3 + 21 = 24$ variance parameters and $22 \cdot 2 = 44$ parameters in the mixing function, summing up to a total of 134 parameters, plus all the variable selection indicators. The variable selection indicators in the mixing function are restricted so that the coefficient of a knot is either unrestricted in all components, or zero in all of them. It takes roughly 24 minutes with the Gibbs sampler to generate a posterior sample with 2,000 burn-in draws followed by an additional 10,000 draws. The average MH acceptance probability is 73% for δ and 34% for γ. The reason for the relatively low acceptance probability for the multinomial logit parameters is the extremely sharp separation of the components in this example, giving an asymmetric posterior for γ. The sampling is very efficient: the mean inefficiency factors (IF) for the α, σ^2, δ and γ, were 2.54, 11.75, 4.77 and 17.05, respectively (the maximal IFs for each group of parameters were 9.44, 14.96, 12.64, 75.73).

Figure 1 displays the simulated Bishop data, the estimated 95% Highest Posterior Density (HPD) intervals in the predictive distribution, the mixing function and the predictive standard deviation (dashed line is the truth) as a function of x for four different models. The seemingly odd behavior of the intervals at points in covariate space where the number of modes of the density is changing (e.g. at $x \approx 0.27$) is an artifact of the HPD interval construction, the actual predictive densities are well behaved. The first row displays the results for the nonparametric SAGM with a single component, which clearly is not flexible enough. The SAGM(3) model in the second row of Figure 1 does a very good job in capturing the data. Note its highly nonlinear mixing function. The SAGM(3) model is again fitted in the third row of Figure 1, but with the knots excluded in the mixing function (the mean and variance are still nonparametric). The terrible fit of this model clearly demonstrates the importance of a flexible mixing function. Finally, the last row of Figure 1 again analyzes the SAGM(3) with nonparametric mean, variance and mixing function, but this time without knot selection. As expected, this model is very adaptive, but the fit is too wiggly, as is most clearly seen in the variance function. Note also that a smaller smoothing parameter (τ_γ) is not a solution here as that would not give us enough flexibility in the regions where it is needed. Estimating τ_γ will not help either.

ACCEPTED MANUSCRIPT

4.2. **US Inflation.** Our second application is a nonlinear time series model for quarterly US inflation during 1952Q1-2004Q4. It has been documented that both the volatility and the persistence of US inflation seem to increase with the level of inflation (see *e.g.* Christiano and Fitzgerald, 2003), and there is some economic theory to support these findings (Akerlof et al., 2000). Here we show that a SAGM model of inflation with lags of inflation as covariates is able to generate these features. A SAGM generalization of the AR(k) process is of the form

$$\begin{aligned} y_t|(s_t &= j, y_t^H) = c^{(j)} + \textstyle\sum_{i=1}^k \rho_i^{(j)} y_{t-i} + \varepsilon_t \\ var(\varepsilon_t|s_t &= j, y_t^H) = \sigma_j^2 \exp(\textstyle\sum_{i=1}^k \delta_i^{(j)} y_{t-i}), \end{aligned} \tag{4.1}$$

where $y_t^H = (y_{t-1}, ..., y_{t-k})'$. The latent allocation variables s_t follow the multinomial logit model in (2.2) with y_t^H as covariates. The mean function is similar to the SETAR and STAR-type models in the nonlinear time series literature, see e.g. Teräsvirta (2006) for a recent overview. Our methodology allows the errors to be heteroscedastic, and we jointly select the subset of variables that define the thresholds (variable selection in the mixing function) and estimate the locations of the smooth thresholds.

This is an example where we are interested in interpreting the components, so we identify the model using an order restriction on the mean coefficient for x_{t-1}. A very similar predictive density was also obtained without identifying restrictions. We estimate a 2-component SAGM with $k = 4$ lags. The model with separate δ's in the two components had only a marginally better LPDS than the model with a common variance function on a test sample with the last 10 years of data, so we present results for the common variance model. We used separate variable selection in the mean, variance and mixing function with the prior

$$p(k) = \begin{cases} 0.5^k & \text{if lag } 1, .., k-1 \text{ are in the model} \\ 0 & \text{otherwise.} \end{cases}$$

We expect a fairly strong linear relation between inflation and at least some of its lags, so we use the second type of default prior for the σ_j^2, see Section 2.2. Computing time with the Gibbs sampler was 176.54 seconds for $2,000$ burn-in draws followed by $10,000$ additional draws used for inference. All variable selection indicators were updated in each iteration of the Gibbs sampler.

Table 2 reports the results from the $SAGM(2)$ with four lags.[2] First, the posterior sampling is very efficient with only 4 out of 21 parameters having inefficiency factors above 10.

[2]The (out-of-sample) LPDS on the last 10 years of data for the SAGM(2) model is -65.663 (numerical standard error 0.071. The LPDS estimate is -65.451 if we update the posterior at every observation). The same model with $m = 1$ gives -65.737 (n.s.e. 0.053. The LPDS is -64.803 with sequential updating of the posterior), and the LPDS of the usual homoscedastic $AR(4)$ is -67.583 (n.s.e. 0.084. The LPDS is -66.448 with sequential updating of the posterior). This suggest that heteroscedasticity is a more important feature of the model than the change in mean dynamics, at least in the latter part of the sample.

ACCEPTED MANUSCRIPT

The inefficiency factors for the parameters in the variance function are all small. Figure 2 displays the cumulative estimates of the posterior inclusion probabilities over the MCMC iterations. Convergence is rapid given the fairly small number of observations in the data set. The first lag in the mean of the first component ($\rho_1^{(1)}$) is highly significant with a large coefficient (remember that the data are standardized, see below for a more interpretable persistence measure), whereas all lags in the mean of the second component have small posterior inclusion probabilities. There is strong support for heteroscedasticity in the data; the posterior inclusion probability for x_{t-1} in the variance function is 0.945. The estimation results in Table 2 also suggest that lag 2, 3 and 4 are essentially redundant in the mean, variance and mixing function[3]. We will therefore for simplicity continue the graphical analysis using the model with $k = 1$. Note that the SAGM model is not additive for $m > 1$ (because of the non-linear logit mixing function), so having only a single lag in the model simplifies the graphical analysis. We can then, for example, graph the predictive density of x_t as a function of x_{t-1} without making the difficult choice of conditioning values for the other lags.

The upper left subgraph of Figure 3 displays the fit of a $SAGM(2)$ with one lag. The estimated model is clearly heteroscedastic (the posterior inclusion probability of x_{t-1} in the variance function is 0.936). The predictive mean has an interesting kink just above zero inflation, suggesting that inflation persistence varies with the level of inflation (see also the mixing functions in the upper right part of Figure 3). A more formal measure of the persistence is given by the first derivative of the mean function with respect to y_{t-1} (Kapetanios, 2007). Using (2.3), this persistence measure is

$$\pi_1(y_{t-1})\rho^{(1)} + \pi_2(y_{t-1})\rho^{(2)} + \pi_1(y_{t-1})\pi_2(y_{t-1})\gamma^{(2)}\left[E(y_t|y_{t-1}, s_t = 2) - E(y_t|y_{t-1}, s_t = 1)\right],$$

where $\gamma^{(2)}$ is the mixing function coefficient on y_{t-1} for the second component. The posterior distribution of this persistence measure is shown in Figure 3. The mean persistence is roughly zero when inflation is negative or near zero (posterior inclusion probability of x_{t-1} in the low persistence component is 0.277), it then increases quite rapidly in the region $0\% - 3\%$ inflation to finally settle down around 0.9 when inflation is above 4%. In models with more than one lag, persistence can be defined as the modulus of the largest eigenvalue of the companion matrix with the usual AR coefficients replaced by the corresponding derivatives of the mean function (Kapetanios, 2007).

[3]The (out-of-sample) LPDS on the last 10 years of data with $k = 1$ is -66.755 (n.s.e. 0.042. The LPDS is -66.843 when the posterior is updated sequentially). This is close to the LPDS from the model with $k = 4$.

ACCEPTED MANUSCRIPT

Parameter	Mean	Stdev.	Post. Incl.	IF
σ_1	1.498	0.515	–	2.800
$c^{(1)}$	3.809	0.569	–	1.463
$\rho_1^{(1)}$	7.045	1.512	1.000	2.165
$\rho_2^{(1)}$	0.240	0.966	0.356	2.060
$\rho_3^{(1)}$	0.614	1.098	0.284	1.287
$\rho_4^{(1)}$	−0.000	0.099	0.001	1.075
σ_2	1.226	0.349	–	17.71
$c^{(2)}$	2.164	1.603	–	15.398
$\rho_1^{(2)}$	0.297	2.814	0.345	19.149
$\rho_2^{(2)}$	0.118	1.169	0.056	1.474
$\rho_3^{(2)}$	0.058	0.526	0.015	2.802
$\rho_4^{(2)}$	−0.000	0.042	0.000	0.996
δ_1	0.866	0.759	0.945	7.283
δ_2	0.409	0.682	0.309	2.363
δ_3	0.004	0.076	0.006	0.984
δ_4	0.000	0.016	0.000	0.988
γ_0	−4.488	2.636	–	68.902
γ_1	−7.518	6.416	0.797	36.448
γ_2	0.315	2.078	0.122	7.567
γ_3	0.007	0.500	0.007	2.173
γ_4	0.001	0.115	0.000	1.947
MH acc. prob. δ	91.26%			
MH. acc. prob. γ	81.26%			

TABLE 2. US Inflation data. Summaries of the posterior distribution for the SAGM(2) with $k = 4$ lags.

	1	5	100
τ_α	−64.464	−65.277	−66.532
τ_δ	−65.572	−65.698	−65.161
τ_γ	−65.334	−65.395	−65.954

TABLE 3. US Inflation data. Exploring the sensitivity of the LPDS to changes in the prior hyperparameters. Each number in the table is the LPDS for a prior with τ_α, τ_γ or τ_δ changed one at a time from the default prior setting $\tau_\alpha = \tau_\gamma = \tau_\delta = 10$.

Finally, Table 3 reports the sensitivity of the LPDS on the last 10 years of the data with respect to the prior hyperparameters τ_α, τ_γ and τ_δ. The LPDS is not sensitive to these prior hyperparameters, except possibly for τ_α.

ACCEPTED MANUSCRIPT

4.3. **Simulated heteroscedastic data.** Jiang and Tanner (1999b) prove that the class of hierarchical smooth mixtures of generalized linear models, which includes the SMR as a special case, can approximate any density in the exponential family with (Sobolev) smooth mean function, but constant dispersion parameter. The approximation rate is $O(m^{-4/s})$ in Kullback-Leibler divergence, where m is the number of components in the mixture and s is the number of covariates. The rate of approximation thus deteriorates quite rapidly as the number of covariates grows. This is a very interesting theoretical result, but it does not take into account that the estimation uncertainty increases with m. The approximation rate of an *estimated* SMR could therefore be dramatically lower than the Jiang-Tanner bound, or it could even fail to correctly approximate the density for any m. Moreover, as pointed out by Jiang and Tanner (1999b), the theoretical rate is not necessarily optimal. Finally, Jiang and Tanner's target density class does not include the Gaussian heteroscedastic regression.

We therefore investigate by simulation how well an estimated SMR can capture heteroscedastic data in finite samples, and in particular how this ability depends on the number of covariates. Data were generated from a single zero mean linear heteroscedastic component with $1, 2, 3$ and 5 covariates. The covariates were generated uniformly in the hypercube $[-1, 1]^p$. The heteroscedasticity parameters were set to $\delta = (-2, -1, 0, 1, 2)$ in the model with 5 covariates, $\delta = (-2, -1, 0)$ in the model with 3 covariates, $\delta = (1, -1)$ in the model with two covariates and $\delta = 1$ in the model with a single covariate. We used $\sigma = 0.1$ in all simulations. For each model we generated 25 data sets, each with a 1000 observations, and then fitted SMR and SAGM models with linear components. To simplify the comparisons of strength of evidence with the real data examples later in this section we use cross-validation (see Section 3) here even if we know the true DGP. The prior with $\tau_\alpha = \tau_\delta = \tau_\gamma = 10$ and $\psi_1 = \psi_2 = 0.01$ was used for all models. Variable selection was not used for simplicity. Both the SMR and SAGM models were fitted with one to five components.

Figure 4 displays box plots of the difference in 5-fold cross-validated LPDS between the SMR models with a given number of components and the estimated SAGM(1) model. The test samples thus contain 200 observations. With a single covariate the predictive performance of the SMR models with $m \geq 3$ is fairly close to that of SAGM(1). As the number of covariates grows, the SMR model has increasing difficulty in fitting the data, relative to the SAGM(1) model, and it seems that its predictive performance cannot be improved by adding more than five components. There are already some signs of overfitting with five components. Even with two covariates the evidence is decisively in favor of the SAGM(1) model (Jeffreys, 1961). We also simulated data from a model with 10 covariates (not shown), and the results followed the same trend: the performance of the SMR relative to the SAGM(1) was much inferior to the case with five covariates.

ACCEPTED MANUSCRIPT

We also investigated the consequences of fitting a SAGM model when the true DGP is an SMR model. Two hundred and fifty datasets were simulated from a five-covariate SMR(2) model with the coefficients in α generated independently from the $N(0,1)$ distribution (i.e. a new α for each data set). The gating coefficients in the DGP were fixed to $\gamma = (1, 1, -1, 2, 0, 0)$. We then fitted the SMR(2) and SAGM(2) models using 5-fold cross-validation exactly as above. The SMR(2) had a higher LPDS than the SAGM(2) in 91.6% of the generated data sets, but the differences in LPDS were typically very small. A 95% interval for the difference in LPDS between the two models ($LPDS_{SAGM}$-$LPDS_{SMR}$) ranged from -1.368 to 0.366, with a median of -0.640. Note also that variable selection could have been used to exclude covariates in the variance function of SAGM, which should have further reduced the gap.

4.4. **US stock returns.** Our final example revisits the analysis of the distribution of returns to the S&P500 stock market index in Geweke and Keane (2007). Our data set contains 4646 daily returns from January 1, 1990 to May 29, 2008 (the sample in Geweke and Keane (2007) ends in the last trading day of the 1990's). The response variable is Return: $y_t = 100 \ln(p_t/p_{t-1})$, where p_t is the closing S&P500 index on day t. A time plot of the variable Return is given in the upper left subgraph of Figure 5.

Geweke and Keane (2007) conduct an out-of-sample evaluation of the conditional distribution of Return where the SMR model outperforms the popular t-GARCH(1,1) and several other widely used models for volatility in stock return data. One of our aims is to see if the SAGM can improve on the SMR by more effectively capturing the heteroscedasticity in Return using the heteroscedastic component so that the mixture can concentrate more heavily on modelling the fat tails and skewness.

We begin with the two predictors used by Geweke and Keane (2007): RLastDay y_{t-1} and CloseAbs95, a geometrically declining average of past absolute returns $(1-\varphi)\sum_{s=0}^{\infty}\varphi^s\,|y_{t-2-s}|$, with $\varphi = 0.95$.[4] We later add seven additional covariates to the model. Our model comparison criterion is an out-of-sample LPDS evaluation of the data in the period between January 1, 2000 and May 29, 2008, giving us 2528 observations for estimation and 2118 observations for predictive evaluation. We evaluate the LPDS using the posterior distribution of the model parameters available just before the start of the evaluation sample, with no additional posterior updating as we go through the evaluation sample. Evidence to support the accuracy of this approximation is given below. We report results from the model where the heteroscedasticity is common to all components as it outperformed the model with separate δ in terms of the LPDS. We do not expect a mean relation between the returns and the predictors, so the

[4]As in Geweke and Keane (2007), we use 10 years of data before the start of our sample to initialize the geometric averages.

ACCEPTED MANUSCRIPT

mean of each component is restricted to be constant, as in Geweke and Keane (2007) and much of the literature on stock market data, and we use the first type of prior for the σ_j^2. We generated 30,000 draws from the posterior, and used the last 25,000 draws for inference. This was sufficient for convergence of the estimates of the model parameters, the posterior inclusion probabilities and the LPDS.

Preliminary analysis suggests that the log variance of Return is strongly related to CloseAbs95, but that this relationship is nonlinear. A logarithmic transformation of CloseAbs95 makes the relationship linear, however. This suggests that it may be sufficient to model the mixture components as simple linear (heteroscedatic) regressions once CloseAbs95 enters the model in logarithmic form. To investigate this more formally we compared the predictive performance of the following three models: i) a SAGM(1) with linear components in the original variables, ii) a SAGM(1) with linear components in RLastDay and the logarithm of CloseAbs95 and iii) a SAGM(1) in the original variables with the components modelled very flexibly as two-dimensional thin plate spline surfaces. Separate spline surfaces, each with 20 knots in $\mathbb{R}^2$, were used in the variance and gating functions in the spline model. The locations of the knots were chosen by the algorithm in Appendix C. The LPDS is -2997.67 for the first model, -2983.86 for the second model, and finally -2984.92 for the third model, suggesting that linear components do as well as fully nonparametric components, but only if CloseAbs95 enters in logarithmic form. The results are also a testimony to the strength of the spline surface model since it is apparently able to automatically find the correct transformation without any model specification search prior to the estimation. Based on these results, we will continue the analysis with linear components and CloseAbs95 entering in logarithmic form.

The upper part of Table 4 reports the LPDS for the SMR and SAGM models with 1-5 linear components in the two covariates RLastDay and (log) CloseAbs95. The SMR improves its predictive performance quite rapidly as we go from one to three components, where it seems to level off so that more than three or four components do not seem to improve the model. It is clearly possible to improve on the SAGM(1) by adding more components, and the maximal LPDS is obtained with four components. There is more than a 12 unit difference in LPDS between the best SAGM and the best SMR.

The results are insensitive to the exact choice of prior hyperparameters. As an example, the LPDS of the SAGM(3) model for the prior with $\tau_\gamma = \tau_\delta = 1$ is -2954.075, and for the prior $\tau_\gamma = \tau_\delta = 100$ it is -2952.33, which are very close to the LPDS of -2953.64 for the default prior. It is clear that the large estimation sample with 2528 observations has reconciled these three very different priors.

ACCEPTED MANUSCRIPT

Two covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-3360.41	-2990.44	-2962.45	-2959.39	-2960.95
	(0.46)	(0.98)	(0.61)	(0.68)	(0.83)
$SAGM$	-2984.31	-2957.03	-2953.64	-2946.68	-2949.39
	(0.17)	(0.31)	(0.64)	(0.74)	(0.73)
Nine covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-3360.41	-2966.74	-2930.07	-2937.21	-2941.28
	(0.46)	(0.77)	(0.76)	(0.78)	(0.94)
$SAGM$	-2921.99	-2909.09	-2892.35	-2894.69	-2903.31
	(0.45)	(0.82)	(0.56)	(0.77)	(0.943)

TABLE 4. SP500 data. Evaluating the LPDS of the models on the observations from January 1, 1990 to May 29, 2008. The posterior distribution of the model parameters are updated only once at the end of the training sample. All models have linear components. Numerical standard errors are given in parenthesis.

Two covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-3327.24	-2992.42	-2962.02	-2955.12	-2958.62
$SAGM$	-2984.26	-2957.54	-2948.49	-2948.81	-2947.72
Nine covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-3327.24	-2964.56	-2924.81	-2921.45	-2925.62
$SAGM$	-2920.84	-2907.53	-2891.33	-2888.28	-2888.52

TABLE 5. SP500 data. Evaluating the LPDS of the models on the observations from January 1, 1990 to May 29, 2008. The posterior distribution of the model parameters is updated every 100th trading day. All models have linear components.

The LPDS reported so far is computed with no additional posterior updating after the end of the estimation sample. As argued in Section 3.4, this practice results in an approximation when applied to time series data, and we now report the accuracy of this approximation in some detail. Table 5 recomputes the LPDS in Table 4, but this time updating the posterior every 100th trading day. A comparison of Tables 4 and 5 shows that the LPDS with sequentially updated posterior are remarkably similar to the ones obtained from a single posterior update. The largest discrepancy is observed for the $SMR(1)$ model, here the LPDS changes

ACCEPTED MANUSCRIPT

quite substantially when the posterior is sequentially updated. This reflects the instability in fitting a fixed variance model to a time series with a clear volatility clustering. We also computed the LPDS for the $SMR(1)$ and the $SAGM(1)$ models when the posterior was updated every 10th trading day, yielding -3325.08 and -2984.33, respectively. These are essentially the same LPDS estimates as for the posterior updated every 100th trading day. In summary, the LPDS for this data set is accurately computed without sequential posterior updating, with the exception of the poorly fitting $SMR(1)$ model.

Figure 6 displays contour plots of the posterior mean of the predictive standard deviation (SD) as a function of the two covariates. The full data set was used in the estimation. The estimated SD changes a lot as more components are added to the SMR model, whereas in the SAGM model the SD is stabler as more components are added. This suggests that the SD can be captured quite well with a single heteroscedastic component. It takes five homoscedastic components to come close to the SD function of the SAGM(1) model.

The difference in interpretation between the SMR and SAGM models is clearly revealed in Figure 7, which depicts the posterior mean of the mixing function in the SMR(3) (left column) and the SAGM(3) (middle column) are plotted. A scatter of the covariates are overlaid in the upper right subplot. An order restriction on the σ_j was used for identification. The SMR components in Figure 7 have been ordered by their variances in descending order from top to bottom, and it is clear that the SMR is using the components to capture the heteroscedasticity in the data (compare with Figure 6). The interpretation of the SAGM model is quite different, with a global component (component no. 1) focusing on capturing the heteroscedasticity. The other two components are more local and take care of the heavy tails. Note also that the mixture weights for the SAGM components are mainly determined by RLastDay. The third column of Figure 7 displays the mixing function from the SAGM model with splines surfaces fitted in the original (untransformed) covariates, from which it is clear that this model picks up something very similar in spirit to the SAGM with linear components. The LPDS of the SAGM(3) with spline surface components is approximately 6 LPDS units lower than the model with linear components.

The different modelling of the tails in the SMR and SAGM models has important consequences for the stock trader which we now explore through Value-at-Risk (VaR) analysis. Figure 8 displays contour plots of the 1% quantile of the predictive distribution. As for the predictive SD, the VaR varies a lot more across the number of components in the SMR than it does for the SAGM. But there are larger differences between the SAGM models in Figure 8 than between the SAGM models in Figure 6. This suggests that while one heteroscedastic component is enough to capture the variance of the S&P500, additional components are needed to model the heavy tails.

ACCEPTED MANUSCRIPT

We now consider the effect of adding seven additional covariates to the model: RLastWeek and RLastMonth, 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 $\varphi = 0.80$, is also added to the covariate set. We also included the square root of $(1-\varphi)\sum_{s=0}^{\infty}\varphi^s y_{t-2-s}^2$, for $\varphi = 0.80$ and 0.95 (CloseSqr80 and CloseSqr95). Finally, we included a measure of volatility that has been popular in the finance literature: $(1-\varphi)\sum_{s=0}^{\infty}\varphi^s(\ln p_{t-1-s}^{(h)} - \ln p_{t-1-s}^{(l)})$, where $p_t^{(h)}$ and $p_t^{(l)}$ are the highest and lowest values of the S&P500 index at day t. Intuitively, this measure should carry more information on the volatility than changes in closing quotes, and indeed this has been shown theoretically and empirically (Alizadeh, Brandt and Diebold, 2002). We consider both $\varphi = 0.8$ (MaxMin80) and $\varphi = 0.95$ (MaxMin95). Following our analysis of the two-covariate model, all variables except RLastDay, RLastWeek and RLastMonth enter the linear components in logarithmic form. The lower part of Table 4 presents the LPDS from this extended model. The first thing to note is the substantial boost in predictive power resulting from the additional covariates, for both the SMR and the SAGM. Second, the best SAGM(3) is now as much as 37 LPDS units better than the best SMR. This follows the same pattern as the simulations in Section 4.3: SAGM's performance on heteroscedastic data relative to the SMR improves with the dimension of the covariate space. It is also interesting to note that SAGM(1) is now 9 LPDS points better than the best SMR, and the computing time of the SAGM(1) is less than a third of the computing time for the SMR(3). We also note that there are now some signs of over-fitting for large m with the larger covariate set, especially for the SMR. Finally, a comparison of the lower parts of Tables 4 and 5 shows again that our conclusions are not altered when the posterior is sequentially updated in the estimation of the LPDS.

The results from the variable selection for the best fitting nine-variable SAGM(3) are quite clear.[5] The following three variables were the only ones to attain a posterior inclusion probability larger than 0.5 in the variance function: RLastMonth (1.000), MaxMin95 (1.000), RLastDay (0.947), where the actual posterior inclusion probabilities are in parenthesis. In the mixing function, only the two covariates RLastDay (1.000) and RLastWeek (0.999) exceed the 0.5 threshold. It is interesting to note that MaxMin95 is crowding out the otherwise very important CloseAbs95 from the variance function, thereby supporting its theoretically superior information content. Figure 9 displays the cumulative MCMC estimates of the posterior inclusion probabilities as a function of the number of iterations. These estimates begin to settle down already after $10,000$ draws, and are very stable after $30,000$ draws (this is even truer if the first $5,000$ draws are discarded as burn-in). Perhaps more importantly,

[5]Since we are using a common variance function for all components and common variable inclusion indicators in the mixing function, we can directly interpret the results from the variable selection without any additional identifying assumptions.

ACCEPTED MANUSCRIPT

multiple runs with different seeds to the random number generator produced very similar results.

To give an idea of the numerical precision in the estimates of the nine-variable SAGM(3) model, we report that the mean inefficiency factor (IF) for the parameter blocks α, σ^2, δ and γ, were 18.87, 27.49, 21.40 and 35.66, respectively (the maximal IFs for each groups of parameters were 44.34, 39.96, 69.85 and 61.53). We experimented also with different numbers of Newton steps in the proposal for the mixing function parameters. With one, two and three Newton steps computing time was roughly 19, 21 and 24 minutes, respectively (for 12,000 iterations), and the corresponding mean acceptance probability in the γ-step was 61.2, 71.4 and 72.5 percent. The IFs were not notably larger in the one-step algorithm compared to the three-step algorithm. A single Newton step was sufficient for the δ-proposal, as the acceptance probability for the nine-variable SAGM(3) was as high as 90%.

Finally, we completed the revision of this article in the midst of the financial crisis in the fall of 2008, and decided to evaluate the predictive performance of the SMR and SAGM models during this turbulent period. The data are specifically marked out in Figure 5, where it is seen that predicting the distribution of Return during this period is a matter of extrapolating outside the estimation sample. Table 6 reports the LPDS of Return from May 30, 2008 until October 28, 2008, a total of 106 observations. The models were estimated using data up to May 29, 2008, and the posterior was not updated sequentially. From Table 6 we see that the best SAGM model is roughly 11 (two covariate case) and 15 (nine covariate case) LPDS points better than the best SMR model, a huge difference considering that we only used 106 observations to evaluate the LPDS. The relatively poor performance of the SMR in this extrapolation exercise comes from the fact that the SMR essentially models the variance by a step function, see Figure 6, but the variance in the data continues to grow out-of-sample, see e.g. the scatter plot of Return vs MaxMin95 in Figure 5. Note also that SAGM(1) does better than the best SMR model, especially in the nine-variable model.

5. Conclusions

A general model is presented, with accompanying Bayesian MCMC methodology, that can be used to flexibly and accurately model a wide range of regression densities $p(y|x)$, with limited user input. Analysis of real data and evidence from simulation experiments showed that our extension with heteroscedastic mixture components can be crucial when the data are heteroscedastic. This was shown to be especially true when the model included more than a couple of covariates. We proposed a Bayesian variable selection procedure with a novel prior that allows us to automatically reduce the model's complexity, to determine the

ACCEPTED MANUSCRIPT

Two covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-535.35	-337.00	-282.78	-255.03	-256.01
$SAGM$	-254.65	-243.82	-247.39	-246.77	-244.30
Nine covariates					
Model	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
SMR	-535.35	-333.11	-267.13	-253.05	-248.47
$SAGM$	-238.30	-235.05	-236.78	-234.59	-233.34

TABLE 6. SP500 data. Evaluating the LPDS of the models on the observations from May 30, 2008 to October 28, 2008. The posterior distribution of the model parameters is based on the data up to May 29, 2008, and not updated thereafter. All models have linear components. The n.s.e. of the SAGM models ranges from 0.05 to 0.18, the n.s.e. for the SMR models are between 0.37 and 0.96.

optimal location of the spline knots, and to investigate the importance of covariates in the mean, variance and mixing functions.

Like any data augmentation approach to MCMC sampling, our approach is time-consuming for very large data sets. We showed that data sets with at least 5,000 observations are certainly manageable, and we have successfully applied the SAGM model to a data set with more than 20,000 observations, but data sets with more than $50,000$ observations or so may be too onerous. The algorithms may be speeded up by computing the gradient and Hessian matrix in the Newton proposals on a subset of the data, but additional algorithmic development in this area will be useful.

APPENDIX A. THE GIBBS SAMPLER

We now describe the updating steps of the sampling schemes in detail. We make use of the following transformation from a heteroscedastic regression to a homoscedastic one:

$$(Y, V) \to (G_\delta Y, G_\delta V) = (\tilde{Y}, \tilde{V}),$$

where $G_\delta = \text{diag}[\exp(-\delta' w_1/2), ..., \exp(-\delta' w_n/2)]$. The Jacobian of this transformation is $|G_\delta| = \exp(-\delta' \sum w_i/2)$. The extension to the case where δ is different for each component is immediate. We use the following notation. Let n_j denote the number of observations allocated to the jth component for a given s, and V_j the $n_j \times p$ submatrix containing the rows of V corresponding to those observations. Z_j, W_j and Y_j are analogously defined.

Updating α, σ^2 and $\mathcal{V}$

Conditional on s and δ, we can integrate out α and σ^2 to show that the $\mathcal{V}_j$ are independently

ACCEPTED MANUSCRIPT

distributed, and that

$$p(\mathcal{V}_{kj}=1|\mathcal{V}_{-k,j},Y,X,s,\delta) \propto \left|\tilde{V}_j'\tilde{V}_j+\tau_{\alpha_j}^{-2}H_\alpha\right|^{-1/2}\left(\frac{d_j}{2}+\psi_{2j}\right)^{-(n_j+2\psi_{1j})/2}, \tag{A.1}$$

where $\tilde{V}_j$ is the covariate matrix for the jth component assuming the presence of the kth covariate, $\mathcal{V}_{-k,j}$ is $\mathcal{V}_j$ with $\mathcal{V}_{kj}$ excluded, $d_j=\tilde{Y}_j'\tilde{Y}_j-\tilde{Y}_j'\tilde{V}_j(\tilde{V}_j'\tilde{V}_j+\tau_{\alpha_j}^{-2}H_{\alpha_j})^{-1}\tilde{V}_j'\tilde{Y}_j$ is the residual sum of squares of the regression of $\tilde{Y}_j$ on $\tilde{V}_j$.

The non-zero elements of α and the elements in σ^2 can now be generated conditional on $\mathcal{V}$ from

$$\begin{aligned}\sigma_j^2|\mathcal{V}_j,s,\delta,Y,X &\sim IG\left(\frac{n_j+p_j+2\psi_{1j}-1}{2},\frac{d_j+2\psi_{2j}}{2}\right)\\ \alpha_{\mathcal{V}_j}|\sigma_j^2,\mathcal{V}_j,s,\delta,Y,X &\sim N(\mu_{\alpha_j},\Omega_{\alpha_j}),\end{aligned}$$

where $\alpha_{\mathcal{V}_j}$ contains the p_j non-zero coefficients in α_j, $\Omega_{\alpha_j}^{-1}=\sigma_j^{-2}(\tilde{V}_j'\tilde{V}_j+\tau_{\alpha_j}^{-2}H_\alpha)$, $\mu_{\alpha_j}=\sigma_j^{-2}\Omega_{\alpha_j}\tilde{V}_j'\tilde{Y}_j$. Note that $\tilde{V}_j$ and $\tilde{Y}_j$, and H_α are now assumed to be conformable with the current draw of $\mathcal{V}$, so that for example $\tilde{V}_j$ contains only the covariates with non-zero coefficients.

Updating δ and $\mathcal{W}$

We describe the case with a fixed set of covariates, the extension to variable selection is immediate from Section 3.2. The full conditional posterior of the variance function parameters is of the form

$$\begin{aligned}p(\delta|\sigma^2,\alpha,Y,X) &\propto p(Y|\delta,\sigma^2,\alpha,X)p(\delta)=|G_\delta|\,p(\tilde{Y}|\delta,\sigma^2,\alpha,X)p(\delta)\\ &\propto \exp(-\delta'\textstyle\sum w_i/2)\displaystyle\prod_{i=1}^n\exp\left[-\frac{1}{2\sigma_{s_i}^2}(\tilde{y}_i-\alpha_{s_i}'\tilde{v}_i)^2\right]\exp\left(-\frac{\tau_\delta^{-2}}{2}\delta'H_\delta\delta\right).\end{aligned}$$

The full conditional posterior of δ is of non-standard form, and we use the K-step Newton proposal to generate from it. The gradient and Hessian are given by

$$\begin{aligned}\frac{\partial\ln p(\delta|\cdot)}{\partial\delta} &= \frac{1}{2}\sum_{j=1}^m W_j'(\eta_j-\iota_{n_j})-H_\delta\delta\\ \frac{\partial^2\ln p(\delta|\cdot)}{\partial\delta\partial\delta'} &= -\frac{1}{2}\sum_{j=1}^m W_j\,\mathrm{diag}(\eta_j)W_j'-H_\delta,\end{aligned}$$

where $\eta_j=\sigma_{s_i}^{-2}(\tilde{Y}_j-\tilde{V}_j\alpha_j)^2$. It is also possible to replace the Hessian with its expected value

$$E\left[\frac{\partial^2\ln p(\delta|\cdot)}{\partial\delta\partial\delta'}\right]=-\frac{1}{2}W'W$$

in the Newton iterations. The case where the δ's differ across components is handled in exactly the same way since the δ_j are independent conditional on s.

ACCEPTED MANUSCRIPT

Updating γ and $\mathcal{Z}$

γ and $\mathcal{Z}$ are updated using the K-step Newton method. We first describe the case without variable selection. The full conditional posterior of the multinomial logit parameters $\gamma = (\gamma_2', ..., \gamma_m')'$ is of the form

$$p(\gamma|s, X) \propto p(s|X, \gamma)p(\gamma) = \left(\prod_{i=1}^{n} \frac{\exp(\gamma_{s_i}' z_i)}{\sum_{k=1}^{m} \exp(\gamma_k' z_i)}\right) \exp\left(-\frac{\tau_\gamma^{-2}}{2} \sum_{j=1}^{m} \gamma_j' H_\gamma \gamma_j\right), \tag{A.2}$$

which is a non-standard density, so we use a MH step with Newton proposals to draw γ. The gradient is of the form

$$\frac{\partial \ln p(\gamma|\cdot)}{\partial \operatorname{vec} \gamma} = \operatorname{vec}[Z'(D - P) - H_\gamma \gamma],$$

where D is an $n \times m$ matrix where the ith row is zero in all positions except in position s_i where it is unity, and P is the $n \times m$ matrix of component probabilities $\Pr(s_i = j|z_i, \gamma)$. The Hessian consists of $(m-1)^2$ blocks of $q \times q$ matrices of the form

$$\frac{\partial^2 \ln p(\gamma|\cdot)}{\partial \gamma_j \partial \gamma_u'} = \begin{cases} Z'[I_q \otimes P_j(P_u - \iota_n)]Z - H_\gamma\ , & \text{if } j = u \\ Z'[I_q \otimes P_j P_u]Z, & \text{if } j \neq u \end{cases}$$

where P_j is the jth column of P. The matrix P is evaluated at the value of γ at the kth iteration of the Newton algorithm. To handle covariate selection in the gating function we can apply the generalized K-step Newton algorithm in Section 3.2. The matrix A_k in (3.1) is block-diagonal with blocks of the form

$$A_{k,ju} = \begin{cases} Z_j'[I_q \otimes P_j(P_u - \iota_n)]Z_u - H_{\gamma,ju}\ , & \text{if } j = u \\ Z_j'[I_q \otimes P_j P_u]Z_j, & \text{if } j \neq u \end{cases},$$

where Z_j contains the selected covariates for γ_j in the kth iteration of the Newton algorithm, and Z_u contains the selected covariates for γ_u. The matrix P is evaluated at the value of γ at the kth iteration of the Newton algorithm. The matrix B_k and the vector g_k in (3.1) are defined analogously. Note that when the prior for $\mathcal{V}$ depends on the value of the mixing function at the knots (see Section 2.2), then the conditional posterior of γ equals the expression in (A.2) multiplied by

$$\prod_{j=1}^{m} \prod_{k=p_v+1}^{p} Bern[\mathcal{V}_{kj}|\omega_\alpha \pi_j(\kappa_k; \gamma)].$$

A similar factor should be used for $\mathcal{W}$ when the δ's differ across components.

Updating s

The component indicator, s_i $(i = 1, ..., n)$ are independent conditional on the other model parameters, and can therefore be drawn simultaneously. The full conditional posterior of s_i

ACCEPTED MANUSCRIPT

is

$$\begin{aligned} p(s_i &= j|Y,X,\sigma^2,\alpha,\gamma,\delta) \propto p(Y|X,\sigma^2,\alpha,\delta,\gamma,s_i=j)p(s_i=j|Z,\gamma) \\ &\propto \sigma_j^{-1}\exp\left[-\frac{1}{2\sigma_j^2}(\tilde{y}_i-\alpha_j'\tilde{v}_i)^2\right]\exp(\gamma_j'z_i), \qquad (i=1,...,n,\ j=1,...,m). \end{aligned}$$

Appendix B. The collapsed sampler

We present the algorithm for a fixed set of covariates, but the extension to covariate selection is the same as for the Gibbs sampler if $\mathcal{V},\mathcal{W}$ and $\mathcal{Z}$ are simulated in the (γ,s,δ)-block.

Updating γ

This step is the same as the γ-step in the Gibbs sampler.

Updating δ

This MH step is similar to the corresponding step in the Gibbs sampler. The proposal is now obtained by taking K Newton steps toward the mode of $p(\delta|\alpha=\hat{\alpha},\sigma^2=\hat{\sigma}^2,Y,X,s,\gamma)$, where $\hat{\alpha}$ and $\hat{\sigma}^2$ are the posterior mean of α and σ^2 conditional on the current values of s and δ. Conditional on $\alpha=\hat{\alpha},\sigma^2=\hat{\sigma}^2$, this step is directly analogous to the δ-step in the Gibbs sampling algorithm, except that the posterior density function in the MH acceptance ratio is now a product of m *marginal* likelihoods (since we have integrated out α and σ^2), one for each expert.

Updating s

When we integrate out α and σ^2, the component indicators, s_i $(i=1,...,n)$ are no longer independent. It is straightforward to show that the conditional posterior of s_i is of the form

$$\begin{aligned} p(s_i &= j|Y,X,s_{-i},\gamma,\delta) \propto \left(\prod_{j=1}^m p(Y_j|X_j,s,\gamma,\delta)\right)p(s_i=j|X,\gamma) \\ &\propto \exp(\gamma_{s_i}'z_i)\prod_{j=1}^m\left|\tilde{V}_j'\tilde{V}_j+\tau_{\alpha_j}^{-2}H_\alpha\right|^{-1/2}\left(\frac{d_j}{2}+\psi_{2j}\right)^{-(n_j+2\psi_{1j})/2}, \end{aligned} \tag{B.1}$$

where s_{-i} denotes s with the ith element deleted, and $d_j=\tilde{Y}_j'\tilde{Y}_j-\tilde{Y}_j'\tilde{V}_j(\tilde{V}_j'\tilde{V}_j+\tau_\alpha^{-2}H_{\alpha_j})^{-1}\tilde{V}_j'\tilde{Y}_j$ is the residual sum of squares of the regression of $\tilde{Y}_j$ on $\tilde{V}_j$. Note how the marginal likelihood $p(Y|X,s,\gamma,\delta)$ factors as a product of m marginal likelihoods, one for each component. We refer to $p(Y|X,s,\gamma,\delta)$ as the marginal likelihood and $p(Y_j|X_j,s,\gamma,\delta)$ as component j's marginal likelihood. $p(Y_j|X_j,s,\gamma,\delta)$ can be efficiently computed as follows. Let R_j be the upper

ACCEPTED MANUSCRIPT

triangular Choleski factor of $\tilde{V}_j'\tilde{V}_j + \tau_{\alpha_j}^{-2} H_\alpha$. Then

$$\left|\tilde{V}_j'\tilde{V}_j + \tau_{\alpha_j}^{-2} H_\alpha\right|^{-1/2} = \left|R_j' R_j\right|^{-1/2} = \left(\prod_{i=1}^{p} r_{ii}^{(j)}\right)^{-1},$$

where $r_{ii}^{(j)}$ is the ith diagonal element of R_j. Moreover, $d_j = \tilde{Y}_j'\tilde{Y}_j - a_j' a_j$, where $a_j = R_j'^{-1}\tilde{V}_j'\tilde{Y}_j$. a_j is thus efficiently solved from the system of equations $R_j a_j = \tilde{V}_j'\tilde{Y}_j$ by back-substitution.

Note, however, that we need to compute the marginal likelihood $p(Y|X, s, \gamma, \delta)$ in (B.1) nm times for a single update of all component allocations. Fortunately, the change from one computation to the next consists of a simple re-allocation of a single observation from one component to another. For example, computing $p(s_i = j|Y, X, s_{-i}, \gamma, \delta)$ requires that we move the ith observation from its current allocation with component j^* to component j. This requires modifying the Choleski factors from $\tilde{V}_{j^*}'\tilde{V}_{j^*} + \tau_{\alpha_{j^*}}^{-2} H_\alpha$ to $\tilde{V}_{j^*}'\tilde{V}_{j^*} + \tau_{\alpha_{j^*}}^{-2} H_\alpha - \tilde{v}_i\tilde{v}_i'$ (i.e. removing observation i from component j^*, which is called a *downdate* of the Choleski with $\tilde{v}_i$) and from $\tilde{V}_j'\tilde{V}_j + \tau_{\alpha_j}^{-2} H_\alpha$ to $\tilde{V}_j'\tilde{V}_j + \tau_{\alpha_j}^{-2} H_\alpha + \tilde{v}_i\tilde{v}_i'$ (i.e. adding observation i to component j^*, which is called an *update* of the Choleski with $\tilde{v}_i$).

Even with sequential Choleski updating, the updating of s can be slow when m and n are large. One way to improve the speed of the algorithm is to sample s using the Metropolis-Hasting algorithm. There are two important advantages to this approach: i) we only need to evaluate $p(s_i = j|Y, X, s_{-i}, \gamma, \delta)$ for the observations where we propose a change (i.e. if observation i is proposed to stay with the same component as before, then the acceptance probability is unity), and ii) whenever a change of component allocation is proposed we only need to evaluate $p(s_i = j|Y, X, s_{-i}, \gamma, \delta)$ at the current and proposed allocations. If n_c denotes the number of observations where a change is proposed, then a draw of the vector s has been reduced from an $O(nm)$ operation to an $O(2n_c)$ operation, which is typically a substantial reduction since in the typical case $n_c << n$. There are many ways to propose s. Among them is to propose from the mixing function $p(s_i = j|z_i, \gamma)$, where γ is the most recently accepted draw of the gating function coefficients. Another option is to use an adaptive scheme where s_i is proposed from the empirical distribution of past allocations (after generating a suitable number of draws to build up the empirical distribution). Nott and Kohn (2005) prove that this type of adaptation produces draws that converge in distribution to the target distribution. It is also possible to combine the different updating schemes in a hybrid sampler where the schemes are selected at random with fixed selection probabilities. For example, with a (small) probability θ we go through all observations and sample s directly from $p(s_i = j|Y, X, s_{-i}, \gamma, \delta)$, and with probability $1 - \theta$ we sample s using an MH step. This combined strategy reduces the possibility of getting stuck in a local mode because of a poorly chosen MH proposal kernel.

ACCEPTED MANUSCRIPT

Appendix C. An algorithm for knot placement

Let x_i denote the p-dimensional covariate vector for the ith unit in the sample. Let $d_A(x_i, x_j) = [(x_i - x_j)'A^{-1}(x_i - x_j)]^{1/2}$ denote the Mahalanobis distance in p-space, where A is a positive definite matrix. A *Mahalanobis ϵ-ball* around $\tilde{x}$ in $\mathbb{R}^p$ is defined to be the set $\{x \in \mathbb{R}^p\colon\ d_A(x_i, x_j) \leq \epsilon\}$. The following algorithm determines the knot locations for a given *global radius* $\epsilon > 0$ and *local radius shrinkage* factor α.

Algorithm C.1.

0. *Form $X = (x_1', ...x_n')'$. Compute $S = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})'$, where $\bar{x}$ is the sample mean.*
1. *Compute the mean $\bar{x}$ of X.*
2. *Find the observation x_c in X that is closest to $\bar{x}$ according to the Mahalanobis distance $d_S(\cdot, \cdot)$.*
3. *Form a Mahalanobis ϵ-ball around x_c. Let n_c denote the number of observations in X that belong to this ϵ-ball .*
4. *Locally adapt the radius to $\epsilon_c = \epsilon/(n_c)^\alpha$.*
5. *Place a knot at the observation that is closest to the mean of the observations in the ϵ_c-ball in step 4.*
6. *Remove the observations that belong to the ϵ_c-ball in step 4 from X.*
7. *Repeat steps 1-6 until X is empty.*

The radius shrinkage factor α determines the extent to which regions of high density are given more knots compared to lower density regions; $\alpha = 1/p$ is a good choice. A root-finding algorithm can be used to search for the global radius ϵ that gives exactly a pre-specified number of knots.

References

[1] Akerlof, G., Dickens, W. T., and Perry, G. L. (2000). Near rational wage and price setting and the optimal rates of inflation and unemployment, *Brookings Papers on Economic Activity*, 5, 1-60.

[2] Alizadeh, S., Brandt, M., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models, *Journal of Finance*, **57**, 1047-1092.

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[4] Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture distributions, *Journal of the American Statistical Association*, **95**, 957-970.

[5] Christiano, L. J. and Fitzgerald, T. J. (2003). Inflation and monetary policy in the twentieth century, *Chicago Fed Economic Perspectives*, **1**, 1-24.

[6] De Iorio, M., Muller, P., Rosner, G. L., and MacEarchen, S.N. (2004). An ANOVA model for dependent random measures, *Journal of the American Statistical Association*, **99**, 205-215.

ACCEPTED MANUSCRIPT

[7] Del Moral, P., Doucet, A. and Jasra, A. (2007). Sequential Monte Carlo for Bayesian computation, in *Bayesian Statistics 8* (eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M. and West, M.). Oxford University Press, Oxford.

[8] Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society*, **60**, 330-350.

[9] Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Regression and Classification*, Wiley, Chicester.

[10] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society B*, **56**, 163-175.

[11] Dimatteo, I, Genovese, C. R., and Kass, R. E. (2001). Bayesian curve fitting with free-knot splines, *Biometrika*, **88**, 1055-1071.

[12] Dunson, D. B., Pillai, N., and Park, J. H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society* B, **69**, 163-183.

[13] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577-588.

[14] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York.

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

[16] Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153-160.

[17] Geweke, J. (2007). Interpretation and inference in mixture models: simple MCMC works, *Computational Statistics and Data Analysis,* **51***, 3529-3550.*

[18] Geweke, J, and Keane, M. (2007). Smoothly mixing regressions, *Journal of Econometrics,* **138***,* 252-290.

[19] Green, P. J. and Richardson, S. (2001). Modeling heterogeneity with and without the Dirichlet Process, *Scandinavian Journal of Statistics*, **28**, 355-375.

[20] Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

[21] Griffin, J. E., and Steel, M. F. J. (2007). Bayesian nonparametric modelling with the dirichlet process regression smoother, unpublished manuscript. http://www.kent.ac.uk/ims/personal/jeg28/BayNPsm.pdf.

[22] Holmes, C. C., and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis*, **1**, 145-168.

[23] Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts, *Neural Computation*, **3**, 79-87.

[24] Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling, *Statistical Science*, **20**, 50-67.

[25] Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.

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

[27] Jiang W., and Tanner, M. A. (1999b). On the approximation rate of hierarchical mixture-of-experts for generalized linear models, *Neural Computation*, **11**, 1183-1198.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[28] Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, **6**, 181-214.

[29] Kapetanios, G. (2007). Measuring conditional persistence in nonlinear time series, *Oxford Bulletin of Economics and Statistics*, **69**, 363-386.

[30] Kass, R. E. (1993). Bayes factors in practice, *The Statistician*, **42**, 551-560.

[31] Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions, *Statistics and Computing*, 313-322.

[32] Kuo, L., and Peng, F. (2000). A mixture-model approach to the analysis of survival data. In *Generalized Linear Models: A Bayesian Perspective*, Dey, D., Ghosh, S., and Mallick, B. (eds)., Marcel Dekker, New York, 255-270.

[33] Leslie, D. S., Kohn, R., and Nott, D. J. (2007). A general approach to heteroscedastic linear regression, *Statistics and Computing*, 17, 131-146.

[34] Li, Q., and Racine, J. S. (2007). *Nonparametric Econometrics*, Princeton University Press, Princeton.

[35] McLachlan, G., and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

[36] Nott, D. J., and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection, *Biometrika*, **92**, 747-763.

[37] Nott, D. J., and Leonte, D. (2004). Sampling schemes for Bayesian variables selection in generalized linear models, *Journal of Computational and Graphical Statistics*, **13**, 362-382.

[38] Peng, F., Jacobs, R. A. and Tanner, M. A. (1996). Bayesian inference in mixture-of-experts and hierarchical mixtures-of-experts models, *Journal of the American Statistical Association*, **91**, 953-960.

[39] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society*, B, **59**, 731-792.

[40] Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association*, **92**, 894-902.

[41] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

[42] Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, 75, 317-344.

[43] Teräsvirta, T. (2006). Univariate nonlinear time series models, in *Palgrave Handbook of Econometrics, Vol. 1 Econometric Theory* (eds. Mills, T. C. and Patterson, K.).

[44] Villani, M., Kohn, R., and Giordani, P. (2007). Nonparametric regression density estimation using smoothly varying normal mixtures, Sveriges Riksbank Working Paper Series no. 211. Available at www.riksbank.com.

[45] Wood, S., Jiang, W. and Tanner, M. A. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression, *Biometrika*, **89**, 513-528.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

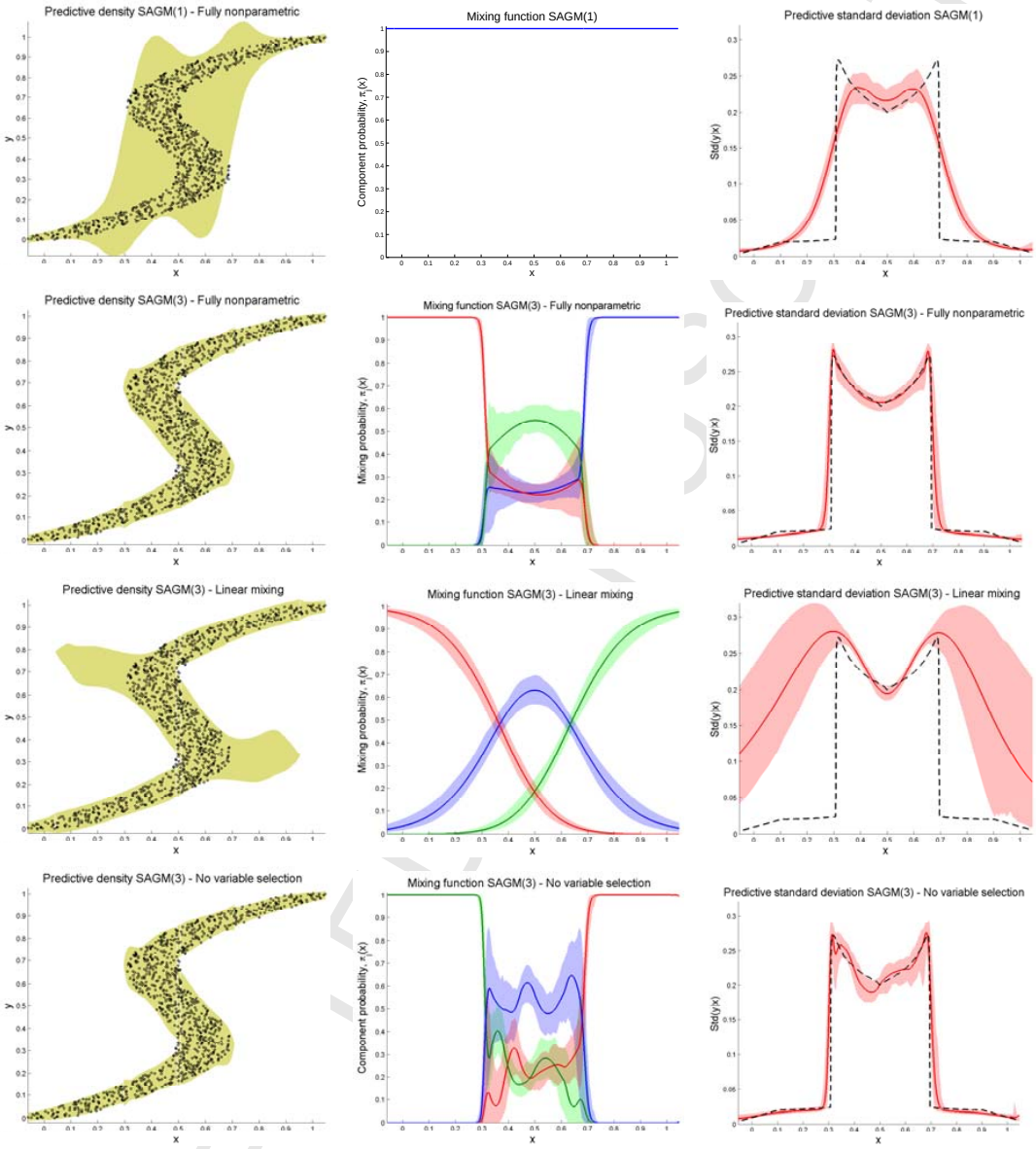

FIGURE 1. Inverse problem data. The first column displays the data and the 95 percent HPD intervals in the predictive density for four different SAGM models. The second and third columns present the posterior mean and 95 percent probability intervals for the mixing and predictive standard deviation function, respectively, for the same four SAGM models.

ACCEPTED MANUSCRIPT

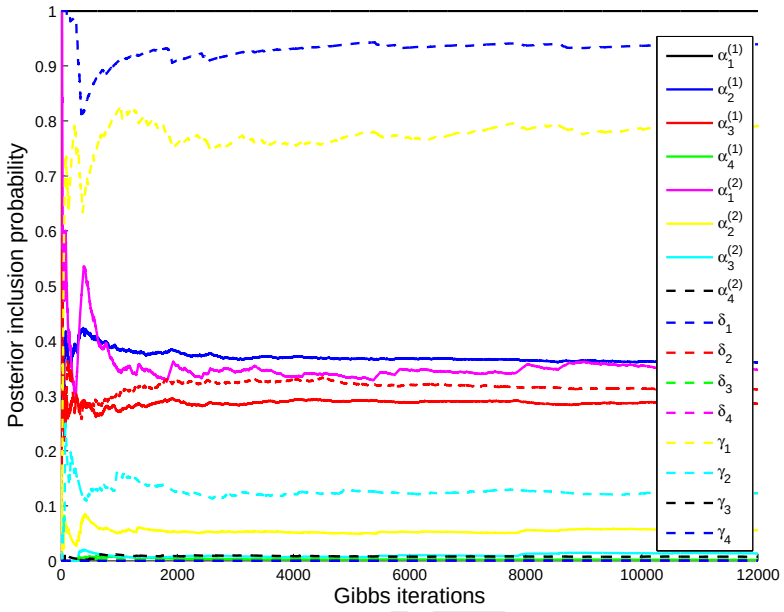

FIGURE 2. US inflation data. Cumulative estimates of the posterior inclusion probabilities as a function of the number of MCMC iterations. The burn-in iterates are included in the figure.

ACCEPTED MANUSCRIPT

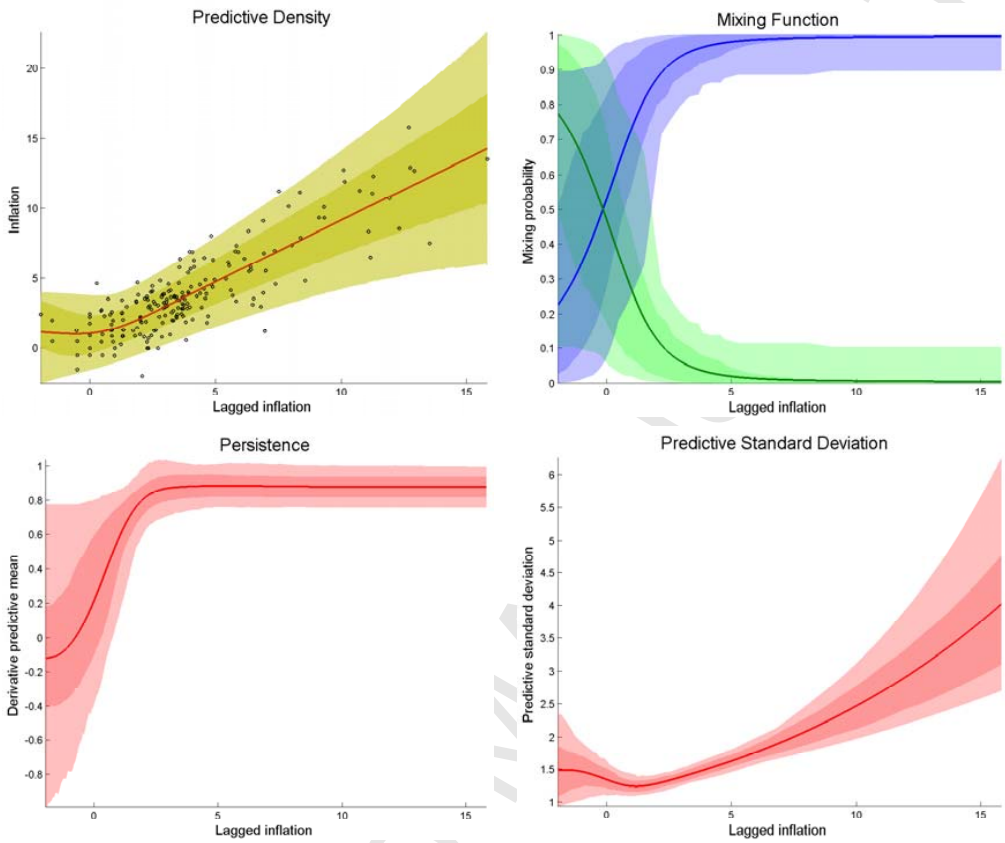

FIGURE 3. US inflation data. The upper left graph displays the data with the 68 and 95 percent HPD intervals in the predictive density of the SAGM(2) model with one lag. The other graphs depict the posterior distribution of the mixing probabilities, the persistence and the predictive standard deviation.

ACCEPTED MANUSCRIPT

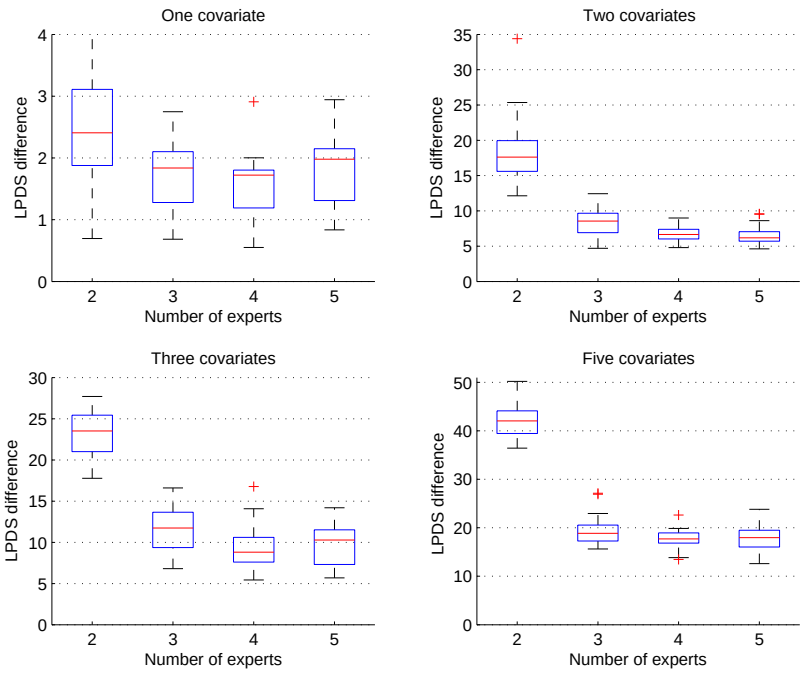

FIGURE 4. Simulated heteroscedastic data. Box plots of the difference in 5-fold cross-validated log predictive density score (LPDS) between the estimated SAGM(1) model and the SMR model as a function of the number of components in the SMR model. Each test sample contains 200 observations.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

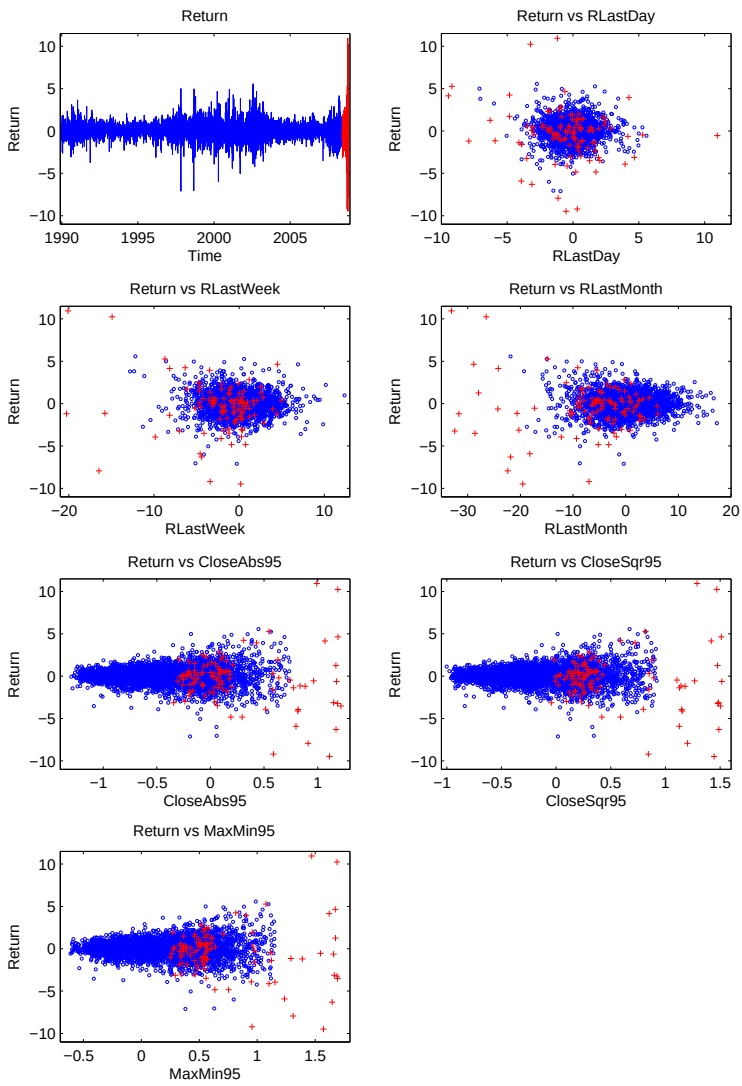

FIGURE 5. Graphical display of the S&P500 data from January 1, 1990 to May 29, 2008 (blue lines and circles) and May 30, 2008 to October 28, 2008 (red lines and crosses). The subgraph in the upper left position is a time series plot of Return, the remaining graphs are scatter plots of Return against a covariate.

ACCEPTED MANUSCRIPT

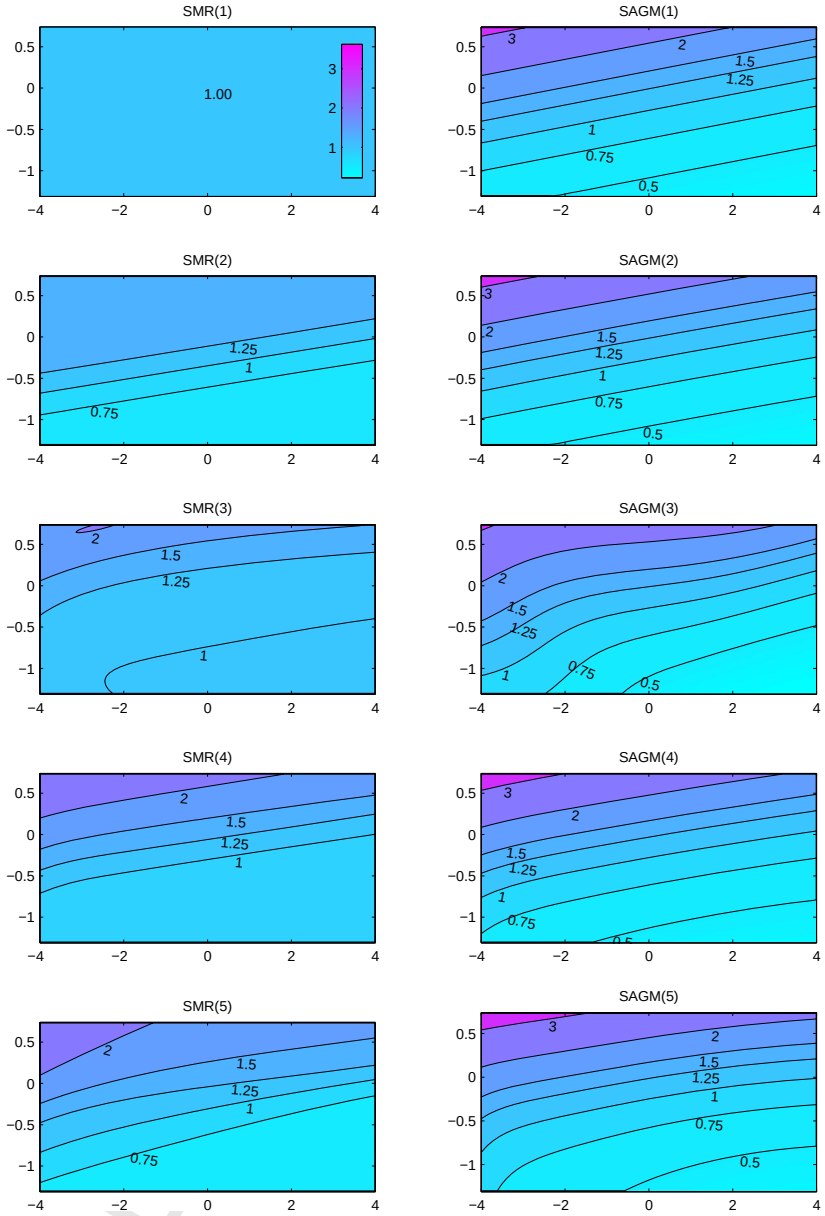

FIGURE 6. U.S. stock return data. Contour plots of the posterior mean of the predictive standard deviation for the SMR (left) and SAGM (right) as a function of the two covariates RLastDay (horizontal axis) and CloseAbs95 (vertical axis) for different number of components in the mixture.

ACCEPTED MANUSCRIPT

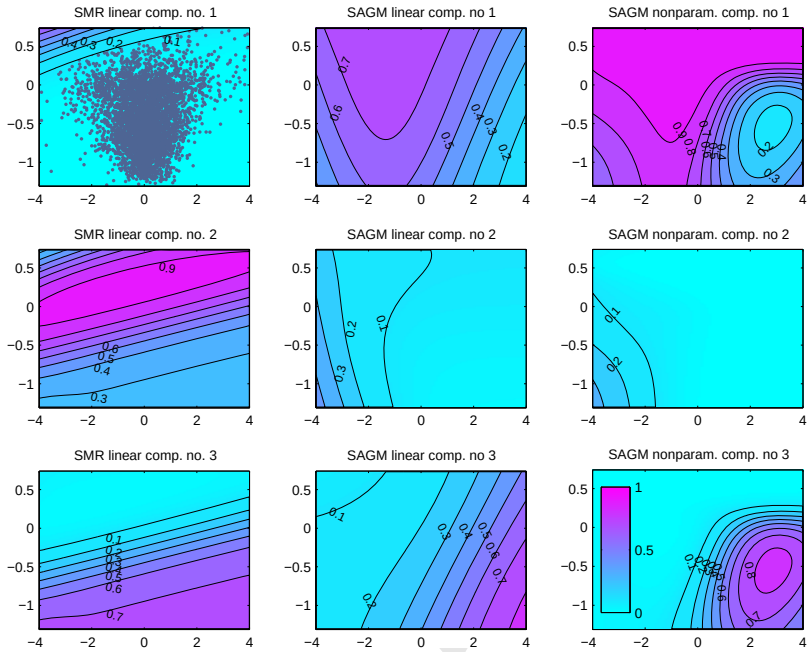

FIGURE 7. U.S. stock return data. Contour plots of the posterior mean of the mixing function as a function of the two covariates RLastDay (horizontal axis) and CloseAbs95 (vertical axis). The left column depicts the mixing function for the SMR(3), the middle column for the SAGM(3) with linear components, and the rightmost column displays the results for the SAGM(3) with nonparametric spline surfaces.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

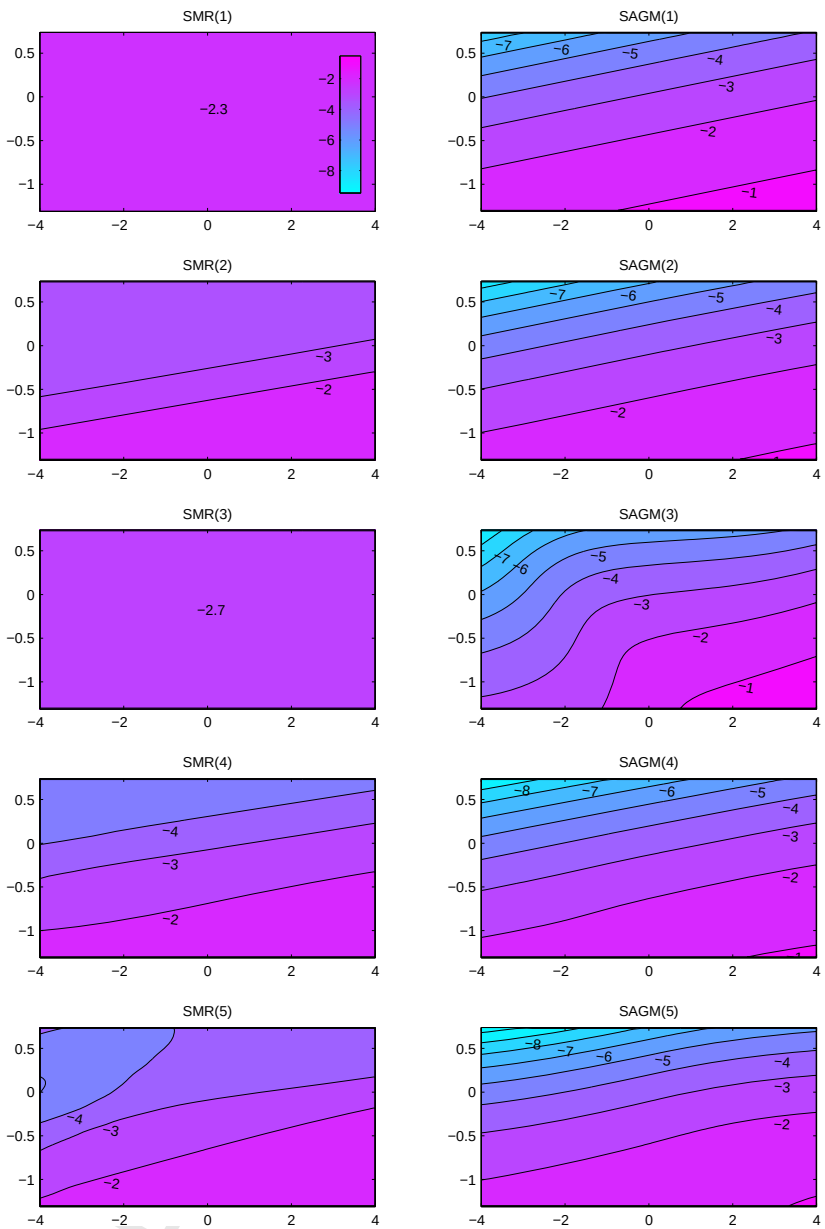

FIGURE 8. U.S. stock return data. Contour plots of the posterior mean of the 1% predictive quantile for the SMR (left) and SAGM (right) as a function of the two covariates RLastDay (horizontal axis) and CloseAbs95 (vertical axis) for different number of components in the mixture.

ACCEPTED MANUSCRIPT

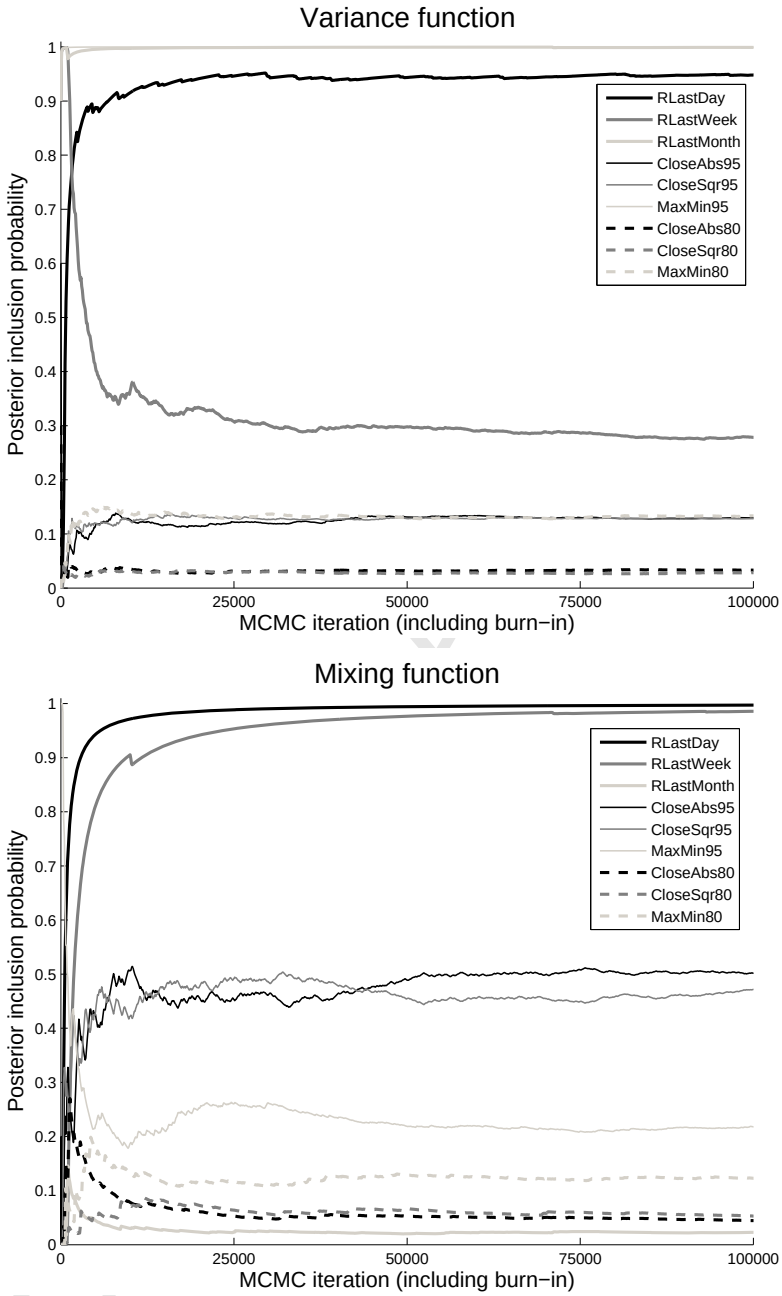

FIGURE 9. U.S. stock return data. Cumulative estimates of the posterior inclusion probabilities in the variance function (top) and mixing function (bottom) as a function of the number of MCMC iterations. No burn-in was removed in the figure.