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 a lot fewer heteroscedastic 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 high-dimensional nonparametric 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.

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 (see *e.g.* Ruppert, Wand and Carroll (2003)), and are increasingly used in econometrics. Our article considers the general problem of nonparametric regression density estimation, i.e., estimating the whole predictive density 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 economics, *e.g.* in the analysis of financial data where accurate estimation of the left tail probability is often the final goal

Villani: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden* and *Department of Statistics, Stockholm University. E-mail: mattias.villani@riksbank.se.* Kohn: *Faculty of Business, University of New South Wales, UNSW, Sydney 2052, Australia.* Giordani: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden.* 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 partly financially supported by a grant from the Swedish Research Council (Vetenskapsrådet, grant no. 412-2002-1007).

of the analysis (Geweke and Keane, 2007), but also in many other areas, such as machine learning (Bishop, 2006), where the predictive density is typically highly nonlinear and multimodal.

Our approach generalizes the popular finite mixture of Gaussians model (McLachlan and Peel, 2000) to the regression density case. Our 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. This model is called *Smoothly Mixing Regression* (SMR) in econometrics (Geweke and Keane, 2007). The SMR model is a mixture of regressions where the mixing probabilities are functions of the covariates which partition the space using stochastic (soft) boundaries.

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 has been 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 linear regressions can approximate essentially any function or a single density. Similarly, it is expected that the SMR should in principle be able to fit heteroscedastic data if the number of mixed regressions is large enough, but it is unlikely to be the most efficient model for that situation. Simulations in Section 3 show that this 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 would 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 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 and overfitting. In particular, variable selection makes the Metropolis-Hastings (MH) steps computationally tractable by reducing the effective number of parameters at each iteration. 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. Section 3 illustrates the methods using real and simulated examples which 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.

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 device 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. However, it is unclear to us how to extend their implementations in a practical way to allow for flexible heteroscedasticity, especially when the number of covariates is moderate to large. The empirical examples in Section 3 show that such extensions are necessary in some examples. To carry out the inference we develop efficient MCMC samplers that compare favourably to existing MCMC samplers for smooth homoscedastic mixtures case as well. A comparison with existing samplers is given in the working paper version of our paper (Villani, Kohn and Giordani, 2007 Appendix D).

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 finite mixture of Gaussians. For example, the simulations in Roeder and Wasserman (1997) suggest that mixtures with up to 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. Components j's *responsibility* 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.

Many of the models in the nonparametric literature are special cases of the SAGM model in (2.1) and (2.2). 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 of these articles use a multinomial probit mixing function. This means that the component probabilities must be computed by numerical integration, which makes the evaluation of predictive densities/likelihoods very time-consuming (Geweke and Keane, 2007). 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 free to take any value. 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.

There are at least two restrictions on the model that are useful in practice. First, we may restrict the heteroscedasticity to be the same across components: $\delta_1 = ... = \delta_m = \delta$. Given that we allow for nonparametric variance and mixing functions, the model will often be flexible enough even under this restriction. 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 both types of restrictions.

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). 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

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

The matrices $\partial z/\partial x$ and $\partial v/\partial x$ are typically of simple form, 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 extracts 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 3.2, where it is used to define the persistence of 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. 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. The covariates are standardized to have zero mean and unit variance to simplify the prior elicitation. 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.**

The prior decomposes as

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

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 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

$$\begin{aligned} \sigma_j^2 &\sim IG(\psi_{1j}, \psi_{2j}) \\ \alpha_{\mathcal{V}_j} | \mathcal{V}_j, \sigma_j^2 &\sim N(0, \tau_{\alpha_j}^2 \sigma_j^2 H_\alpha^{-1}) \end{aligned}$$

where IG denotes the inverse Gamma distribution and $\alpha_{\mathcal{V}_j^c} | \mathcal{V}_j$ is identically zero. H_α is a positive definite precision matrix, often equal to the identity matrix or a scaled version of the cross-product moment matrix $V'V$. 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 coefficients for that component's knot will then be poorly estimated, or may even be unidentified. To deal with this problem, we use the prior

$$\mathcal{V}_{kj} | \gamma \sim Bern[\omega_\alpha \pi_j(\kappa_k^v; \gamma)], \qquad (k = 1, ..., p; \;\; j = 1, ..., m), \tag{2.4}$$

where $0 \leq \omega_\alpha \leq 1$, 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 for example Kohn, Smith and Chan (2001), but it will require an extra MH step.

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

$$\begin{aligned}\delta_{\mathcal{W}_j}|\mathcal{W}_j &\sim N(0, \tau^2_{\delta_j} H_\delta^{-1}) \\ \mathcal{W}_{kj}|\gamma &\sim Bern[\omega_\delta \pi_j(\kappa^w_k;\gamma)], \quad (k=1,...,r,\ j=1,...,m).\end{aligned}$$

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 indicator in $\mathcal{Z}$ are assumed to be *iid* $Bern(\omega_\gamma)$. Let $\gamma_{\mathcal{Z}}$ denote the non-zero coefficients in the mixing function for a given $\mathcal{Z}$. The prior on γ is then assumed to be of the form

$$\gamma_{\mathcal{Z}}|\mathcal{Z} \sim N(0, \tau^2_\gamma H_\gamma^{-1}),$$

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

2.3. **Bayesian inference and model comparison.** We adopt a Bayesian approach to inference using MCMC to sample from the joint posterior distribution of the model parameters. We have experimented with several sampling schemes (see Appendix A to D in Villani et al. (2007) for a description of the algorithms and a comparison on the LIDAR data) and found that the scheme presented in Appendix A gives the best combination of efficiency and computing time. This algorithm includes variable selection in all three sets of covariates: mean, variance and mixing function.

Ideally we would like to use the marginal likelihood as a basis for model comparison. 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. It also gives a better assessment of the predictive performance that can be expected for future observations, and simplifies computations. To deal with the arbitrary choice of which observations to use for training and testing, we use B-fold cross-validation of the log predictive density score (LPDS):

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

where $\tilde{y}_b$ contains the n_b observations in the bth test sample, $\tilde{y}_{-b}$ denotes the remaining observations and $p(\tilde{y}_b|\tilde{y}_{-b}, x_i) = \int \prod_{i\in\mathcal{T}_b} p(y_i|\theta, x_i)p(\theta|\tilde{y}_{-b})d\theta$, where $\mathcal{T}_b$ is the index set for the observations in $\tilde{y}_b$. Here we have implicitly assumed independent observations conditional on θ and the covariates. The LPDS is easily computed by averaging $\prod_{i \in \mathcal{T}_b} p(y_i|\theta, x_i)$ over the posterior draws from $p(\theta|\tilde{y}_{-b})$. This can be computed from B complete runs with the posterior simulator, one for each training sample. For time series data it is typically false that the observations are independent conditional on the model parameters. For such data it is more natural to use the most recent observations in a single test sample. Ideally we would here like to re-estimate the model sequentially as we add more observations to the training sample, but this is too time-consuming, and we instead approximate the LPDS using a fixed training sample for all test observations.

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 Jeffreys' (1961) well-known rule-of-thumb for Bayes factors to assess the strength of evidence. It should be noted however that the original Bayes factor evaluates all the data 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 have chosen not to do so as the LPDS can then no longer be calibrated by Jeffreys scale of evidence.

3. Empirical illustrations

3.1. **Inverse problem.** Our first example is based on an inverse problem 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 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 the data we wish to estimate the density of $p(y|x)$. Figure 1 shows that this is a challenging regression density estimation problem as the density $p(y|x)$ is multimodal and heteroscedastic.

The prior $\tau_\alpha = \tau_\delta = 10$, $\tau_\gamma = 1000$ (the choice of τ_γ is explained below), and $\psi_1 = \psi_2 = 0.01$ (in the IG prior for the σ_j^2's) was used for all fitted SAGM models. We used truncated quadratic splines (see e.g. Ruppert et al, 2003) with 20 equally spaced knots, and variable selection among the knots with inclusion probabilities $\omega_\alpha = \omega_\delta = \omega_\gamma = 0.2$. Figure 1 displays the estimated 95% Highest Posterior Density (HPD) intervals in the predictive distribution, the mixing function and the predictive standard deviation as a function of x for four different models. The HPD intervals of the true density (obtained by simulation) are the black thin lines in the first column of Figure 1. 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 to capture the true density or the standard deviation. The SAGM(3) model in the second row of Figure 1 does an excellent job in capturing the true density and standard deviation. The same model is 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. In fact, with a nonparametric mixing function it is important that τ_γ is not made too small, for then the mixing function cannot change rapidly enough to fit the data (hence the choice of $\tau_\gamma = 1000$ for this data set). 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. 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.

3.2. **US Inflation.** Our second application is a nonlinear time series model for 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). We shall here illustrate 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

$$\begin{aligned} y_t|(s_t &= j, y_t^H) = c^{(j)} + \textstyle\sum_{i=1}^k \rho_i^{(j)} y_{t-i} + \varepsilon_t \\ var(\varepsilon_t|s_t &= j, y_t^H) = \sigma_j^2 \exp(\textstyle\sum_{i=1}^k \delta_i^{(j)} y_{t-i}), \end{aligned} \tag{3.1}$$

where $y_t^H = (y_{t-1}, ..., y_{t-k})'$. s_t follows 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 (soft) thresholds.

We now show that a very simple SAGM model with $k = 1$ lag and two linear components forms an interesting model for US inflation. We use the prior with $\tau_\alpha = \tau_\gamma =$

$\tau_\delta = 5$ and $\psi_1 = \psi_2 = 1$. Variable selection was used with $\omega_\alpha = \omega_\delta = \omega_\gamma = 0.5$ as prior inclusion probabilities. The model with a common variance function performed better than a model with separate δ's in the two components, so we present results for the common variance model. The upper left subgraph of Figure 2 displays the fit of a $SAGM(2)$ with one lag. The estimated model is clearly heteroscedastic: the posterior inclusion probability of y_{t-1} in the variance function is 0.995. 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 2). The two-component SAGM model outperformed the usual $AR(1)$ model in an out-of-sample forecast evaluation with the last ten years removed from the estimation sample (the Bayes factor is 5.55 in favor of the two-component model). Most of improved forecasting accuracy comes from the heteroscedastic disturbances rather than the changes in persistence (the Bayes factor comparing the model with two heteroscedastic components to the model with a single heteroscedastic component is only 2.25).

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

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

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 2. The mean persistence is roughly zero when inflation is low (the posterior inclusion probability of y_{t-1} in the low inflation component is 0.06), it then increases quite rapidly in the region $0\% - 3\%$ inflation to finally settle down around 0.9 when inflation is above 4%. Note that there is a corridor where inflation may even be locally explosive. 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).

3.3. **Simulated heteroscedastic data.**

We now investigate how well the smooth mixture of homoscedastic components can capture heteroscedastic data in finite samples, and in particular how this ability depends on the number of covariates. We simulated data from a single linear heteroscedastic component with $1, 2, 3$ and 5 additive covariates generated uniformly in the hypercube $[-1, 1]^p$. A zero mean was used to isolate the effects of the heteroscedasticity. 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, from the DGP, 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 2.3) 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 fit with one to five components. Figure 2 displays box plots of the difference in LPDS between the SMR models with a given number of components and the estimated SAGM(1) model. 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. Similar simulations also show that there is hardly any loss from fitting an SAGM model when the true DGP is an SMR model, see Villani et al. (2007) for details.

3.4. **LIDAR.** Our next data set has been used extensively in the nonparametric literature. The data comes from a technique that uses laser-emitted light to detect chemical compounds in the atmosphere (LIDAR, LIght Detection And Ranging, see Holst et al. (1996)). The response variable (logratio) consists of 221 observations on the log ratio of received light from two laser sources: one at the resonance frequency of the target compound, and the other from a frequency off this target frequency. The predictor is the distance travelled before the light is reflected back to its source (range). We will use the model with common δ in the components. The models with common δ and the models with separate δ's give essentially the same LPDS. Moreover, when the δ's are allowed to differ across components, the posterior distributions of the δ's are largely overlapping. The prior $\tau_\alpha = \tau_\delta = \tau_\gamma = 10$ and $\psi_1 = \psi_2 = 0.01$ was used, but other priors had very little impact on the fit and the LPDS.

The left column in Figure 4 displays the LIDAR data and the 68% and 95% Highest Posterior Density (HPD) regions in the predictive distribution $p(\mathsf{logratio} \mid \mathsf{range})$ from the SMR model with 3 linear component (top row) and $1, 2$ and 3 thin plate spline components (second to fourth row). See *e.g.* Green and Silverman (1994) for details on thin plate splines. We used 10 equally spaced knots in each of the mean, variance and mixing functions, and variable selection among the knots with $\omega_\alpha = \omega_\delta = \omega_\gamma = 0.2$ as prior inclusion probability. The SMR(3) models do fairly well, but fail to capture the small variance of logratio for the smallest values of range, and the predictive intervals also have a somewhat unpleasant visual appearance.

The right column of Figure 4 displays the fit of the SAGM models. The SAGM(3) model with linear components performs rather well. The best fit seems to be given by the SAGM model with a single nonparametric component. It is interesting to see that the overparametrized SAGM(2) and SAGM(3) models with nonparametric components do not seem to overfit. This is due to the self-adjusting mechanism provided by the variable selection: the more components that are added to the model, the fewer the knots in all components. For example, the SAGM(1) component has a highly non-linear mean, but the components in the SAGM(3) model with nonparametric components are essentially linear, all the knots in the SAGM(3) model have very small inclusion probabilities. The prior in (2.4) is very effective in removing a component's knots in regions of low mixing probability: almost all such knots have zero posterior inclusion probability. All the knots in the variance function of the SAGM models have posterior probabilities smaller than 0.1, suggesting strongly that the (log) variance function is linear in range. There is some evidence of smoothly changing nonlinearity in the (log odds) mixing function where most of the knots have posterior probabilities in the range 0.2-0.4. This is true for both SMR and SAGM models.

Table 1 displays the mean of the log predictive score (LPDS) over the $B = 5$ test samples as a function of the number of components. All three SAGM models with nonparametric components and the SAGM(3) model with linear components give very similar LPDS values. In particular, a single nonparametric heteroscedastic component is sufficient to fit the data. The SMR models need three components to come close to the LPDS of the SAGM model with a single nonparametric component, and even then do not quite reach it.

3.5. **US stock returns.** In our final example we analyze the distribution of 3673 daily returns on the S&P500 stock market index from January 21, 1991 to August 12, 2005. 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. This series is plotted in the left panel of Figure 5. Following Geweke and Keane (2007) we construct two predictors Return Yesterday y_{t-1} and a geometrically declining average of absolute returns, GeoAverage, which is defined as $(1-0.95)\sum_{s=0}^{\infty} 0.95^s \left|y_{t-2-s}\right|$.

Geweke and Keane (2007) conducted an out-of-sample evaluation of the conditional distribution $p(\mathsf{Return} \mid \mathsf{Return\ Yesterday}, \mathsf{GeoAverage})$ where the SMR model dramatically outperformed the popular t-GARCH(1,1) and several other widely used models for volatility in stock return data. Our aim here is to see if the SAGM can do a better job by having the heteroscedastic components capturing the heteroscedasticity in Return so that the mixture can concentrate more heavily on modelling the fat tails and possibly also skewness.

We fit SMR and SAGM models with the components modelled as two-dimensional thin plate spline surfaces. The mean of each component is restricted to be constant, in line with the literature on stock market data. Both the mixing and variance functions use 20 knots in $\mathbb{R}^2$ with the locations of the knots chosen by the algorithm in Villani et al. (2007, Appendix E). We apply variable selection among the knots with inclusion probabilities $\omega_\delta = \omega_\gamma = 0.2$. We used the prior $\tau_\alpha = 1$, $\tau_\delta = \tau_\gamma = 5$ and $\psi_1 = \psi_2 = 1$, but the predictive distribution is not sensitive to non-drastic changes in the prior hyperparameters. We report results from the model where the heteroscedasticity is common to all components as it outperformed the model with separate δ. We generated $30,000$ draws from the posterior, and used the last $25,000$ draws for inference.

Table 2 displays the LPDS for SMR and SAGM models evaluated on the 1000 last trading days as a single test sample. The best model is the SAGM(4) model which is more than 6 LPDS units better than the best SMR model (the Bayes factor is 415.72). This is decisive evidence in favor of the SAGM (Jeffreys, 1961). It is interesting to note that SAGM(1) is only slightly inferior to SAGM(4). This result is however particular to this specific test sample, which happens to be essentially free from outliers. To show this, we plot in Figure 5 (right panel) Return against GeoAverage (the main driver of the heteroscedasticity, see the standard deviation graphs in Villani et al. (2007, Figure 6)) in the training and test sample. It is clear from Figure 5 that a single heteroscedastic component will perform well in the test sample, but will most likely fail to capture the training observations with extreme returns but low GeoAverage value, *if* they had been in the test sample. To investigate this more formally we evaluate the LPDS using 5-fold cross-validation with the test samples systematically sampled through time (the first test sample consists of observation 1, 6, 11, etc.), even if this exercise may be regarded as somewhat unnatural for time series data. Table 3 shows that the SAGM(1) now performs substantially worse than, for example, the SAGM(3) model. The average LPDS difference between the best SAGM and the best SMR is now smaller (the Bayes factor comparing SAGM(3) to SMR(4) is 18.92), but the SAGM(3) model outperforms the SMR(4) in each of the five test samples.

We also consider the effect of using two additional covariates: Time and LastWeek, a moving average of the returns from the previous five trading days. The LPDS on the last 1000 observations is reported in Table 4. A comparison of Tables 2 and 4 shows that the two new covariates bring a very substantial improvment in predictive performance of the SAGM model, whereas the performance of the SMR is more or less unchanged. The relative support for the SAGM model is now dramatically stronger: the Bayes factor comparing the best fitting SMR and SAGM models is $7.26 \cdot 10^7$ in favor of the SAGM model.

Figure 6 display quantile-quantile plots (QQ-plots) of the normalized residuals (see e.g. Leslie et al., 2007) for the SMR and SAGM models with two covariates. The normalized residuals are defined as $\Phi^{-1}[\hat{F}(x_i)]$, for $i = 1, ..., n$, where $\hat{F}(x_i)$ is the posterior expectation of the predictive distribution function at x_i. The QQ-plot graphs the empirical quantiles of the normalized residuals against the quantiles of the standard normal density. Deviations from the 45° degree line signal a lack of fit. The models with one component both do a poor job in the tails of the distribution (not shown in Figure 6 for scaling considerations). Adding another component to the SAGM(1) model gives a substantial improvement in fitting the tails, and the fit of the SAGM(3) is excellent. The SMR model improves as more components are added, but even the SMR(4) model cannot fully capture the tails of the distribution.

The estimated mixing function for the SMR and SAGM models in Villani et al. (2007, Figure 7) clearly reveal that the SMR model is using all the components to capture the heteroscedasticity in the data. The SAGM model has a more efficient division of labor: a dominant global component with a probability exceeding 0.5 for all values of the covariates captures the bulk of the heteroscedasticity and the other much more local components take care of the heavy tails. The implications of this can be seen in the posterior mean of the 1% quantile of the predictive density, the so called Value-at-Risk (VaR). Figure 7 displays the VaR as a function of the two covariates for the SMR(4) and SAGM(4) models. For some covariate values, the difference between the two models is larger than 1%, which is quite substantial for daily returns.

Appendix A. MCMC sampling for the SAGM model

This appendix describes our preferred Metropolis-Hasting scheme for sampling from the joint posterior distribution of the SAGM parameters. See Villani et al. (2007) for details on other algoritms and a comparison of the algorithms' performance .

We use a general method for constructing tailored MH proposal densities to sample from the full conditional posterior of the two pairs $(\gamma, \mathcal{Z})$ and $(\delta, \mathcal{W})$. The technique was initially proposed by Gamerman (1997) for the exponential family without variable selection, and later extended by Nott and Leonte (2004) to allow for variable selection with the exponential family. We will here briefly describe an extension of this method to a more general setting, see Villani et al. (2007, Appendix A) for details. The basic idea of this *variable dimension K-step Newton method* is as follows. Suppose that the model can be written $p(y_i|\theta_i)$, where $\theta_i = \beta' x_i$, x_i is a q-dimensional covariate vector and $\beta = (\beta_1, ..., \beta_q)'$. The distribution $p(y_i|\theta_i)$ is not restricted to the exponential family. Let $\mathcal{J} = (j_1, ..., j_q)$ be a vector of binary variable selection indicators, such that $j_i = 0$ if $\beta_i = 0$, and let $j_i = 1$ otherwise. The two sets of parameters θ and $\mathcal{J}$ are proposed jointly. There are many ways to propose $\mathcal{J}$. A simple but often useful option is to randomly select one of the indicators and always propose a change of the chosen indicator, see Nott and Kohn (2005) for more advanced proposals. The proposal for θ is drawn conditional on $\mathcal{J}$ from the multivariate t-distribution with $c > 2$ degrees of freedom:

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

where the second argument of the density is the covariance matrix. $\hat{\theta}$ is obtained by taking a fixed number of steps (K) with Newton's method from the current point θ_c toward the mode of $p(\theta|y, \mathcal{J}_p)$. Note that K is often set to a small number, so $\hat{\theta}$ is rarely the (conditional) mode, but typically quite close to it. We have found that $K \leq 3$ is more than sufficient for good convergence, which makes the algorithm very fast. We may further speed up the algorithm by evaluating the gradient and Hessian on a (random) subset of the observations. In many cases we can also replace the Hessian by its expected value (Fisher scoring). Because of the variable selection, θ_c and θ_p may be of different dimensions, so that Newton's original method needs to be generalized, see Villani et al. (2007) for details. The key idea is that the functional $\theta_i = \beta' x_i$ is one-dimensional and is expected to change only slightly in value as variables are added or removed from the model. Note also that variable selection has the advantage of keeping down the dimension of θ in every iteration of the algorithm, which speeds up the algorithm and increases the MH acceptance probability.

We now describe the updating steps of the sampling scheme 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 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. V_j denotes the $n_j \times p$ submatrix containing the rows of V corresponding to the jth component's observations given an allocation s. 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 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 here 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 first consider the case without covariate selection. 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

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

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. The extension of the K-step Newton proposal to the case with covariate selection is straightforward; see Villani et al. (2007) for details.

Updating γ and $\mathcal{Z}$

γ and $\mathcal{Z}$ are updated using the K-step Newton method. We describe the updating step for the case without variable selection, and we refer to Villani et al. (2007) for details on the variable selection case. 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. 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 Newton algorithm. 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 is

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

Unless otherwise stated, the reported results in this article were generated by $10,000$ draws after a burn-in of $2,000$ draws. We used $c_\delta = 10$ and $c_\gamma = 10$ degrees of freedom in multivariate-t Newton-based proposal densities for δ and γ. The component allocation is initialized with the k-means clustering algorithm.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

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

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

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

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

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

[15] Griffin, J. E., and Steel, M. F. J. (2007). Bayesian nonparametric modelling with the dirichlet process regression smoother, unpublished manuscript.

[16] Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements, *Environmetrics*, **7**, 401-416.

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

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

[19] 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.

[20] 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.

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

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

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

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

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

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

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

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

[36] 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.

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

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

Table 1. LIDAR data. Average log predictive density score (LPDS) over the 5 cross-validation samples.

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
SMR	−1579.16	−1430.39	−1413.96	−1410.50	−1410.92
SAGM	−1404.95	−1409.02	−1407.99	−1404.47	−1409.06

Table 2. SP500 data - two covariates. Log predictive density score (LPDS) on the last 1000 observations.

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
SMR	−1058.85	−955.97	−945.69	−942.01	−942.02
SAGM	−955.24	−944.22	−939.07	−939.81	−939.51

Table 3. SP500 data - two covariates. Average log predictive density score (LPDS) over the 5 cross-validation samples.

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
SMR	−1579.16	−1428.05	−1412.02	−1412.83	−1414.11
SAGM	−1393.92	−1398.92	−1396.63	−1395.31	−1401.87

Table 4. SP500 data - four covariate model. Log predictive density score (LPDS) on the last 1000 observations.

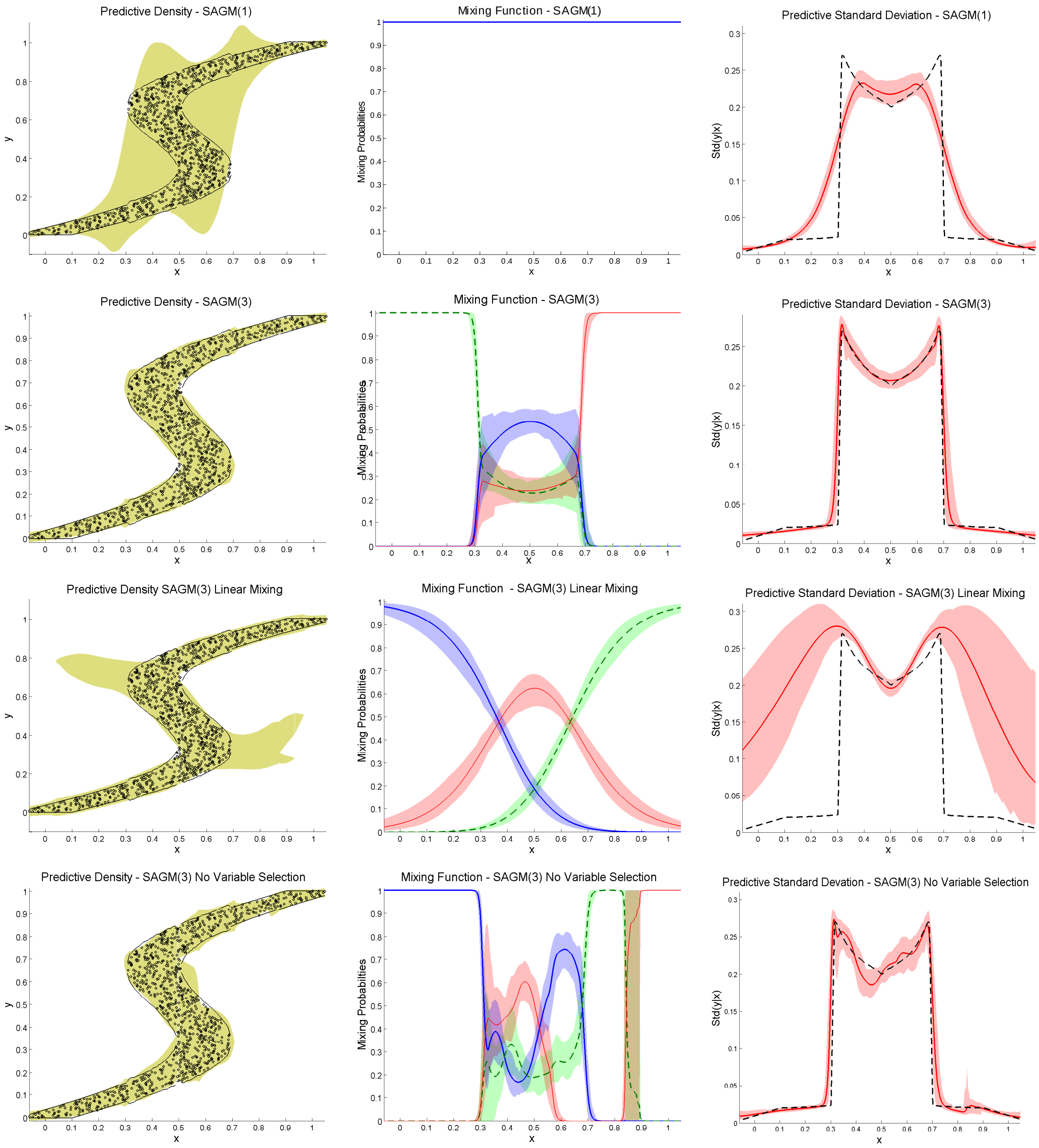

FIGURE 1. Inverse problem data. First column displays the data and the 95 percent HPD intervals in the predictive density. The second and third columns present the mixing and predictive standard deviation function, respectively. The rows correspond to four different SAGM models.

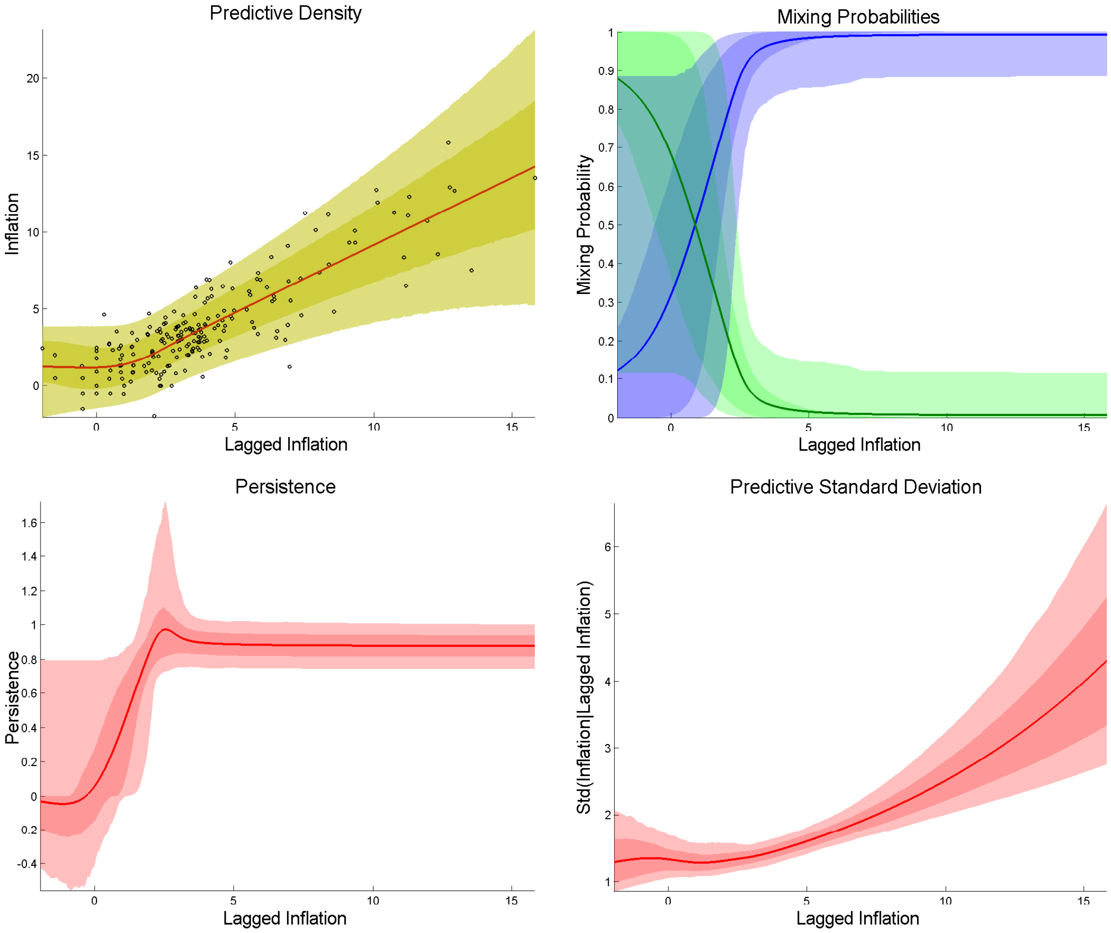

FIGURE 2. US inflation data. The upper left graph displays the data with the 68 and 95 percent HPD intervals in the predictive density of the SAGM(2) model. The other graphs depict the posterior distribution of the mixing probabilities, the persistence and the standard deviation of the predictive distribution.

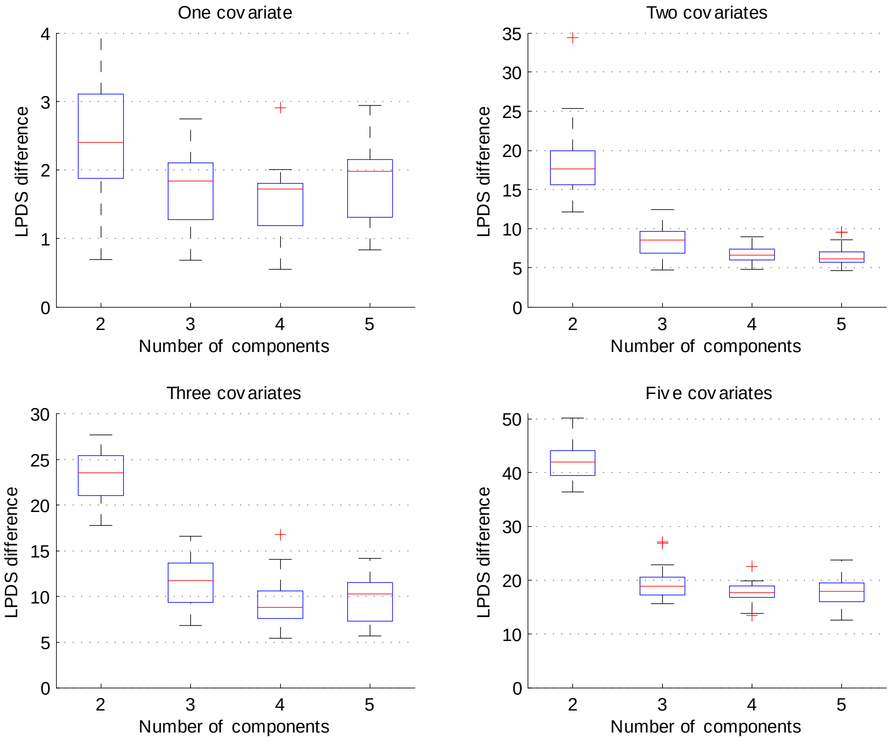

Figure 3. Simulated heteroscedastic data. Box plots of the difference in log predictive score (LPDS) between the estimated SAGM(1) model and the SMR model as a function of the number of components in the SMR model.

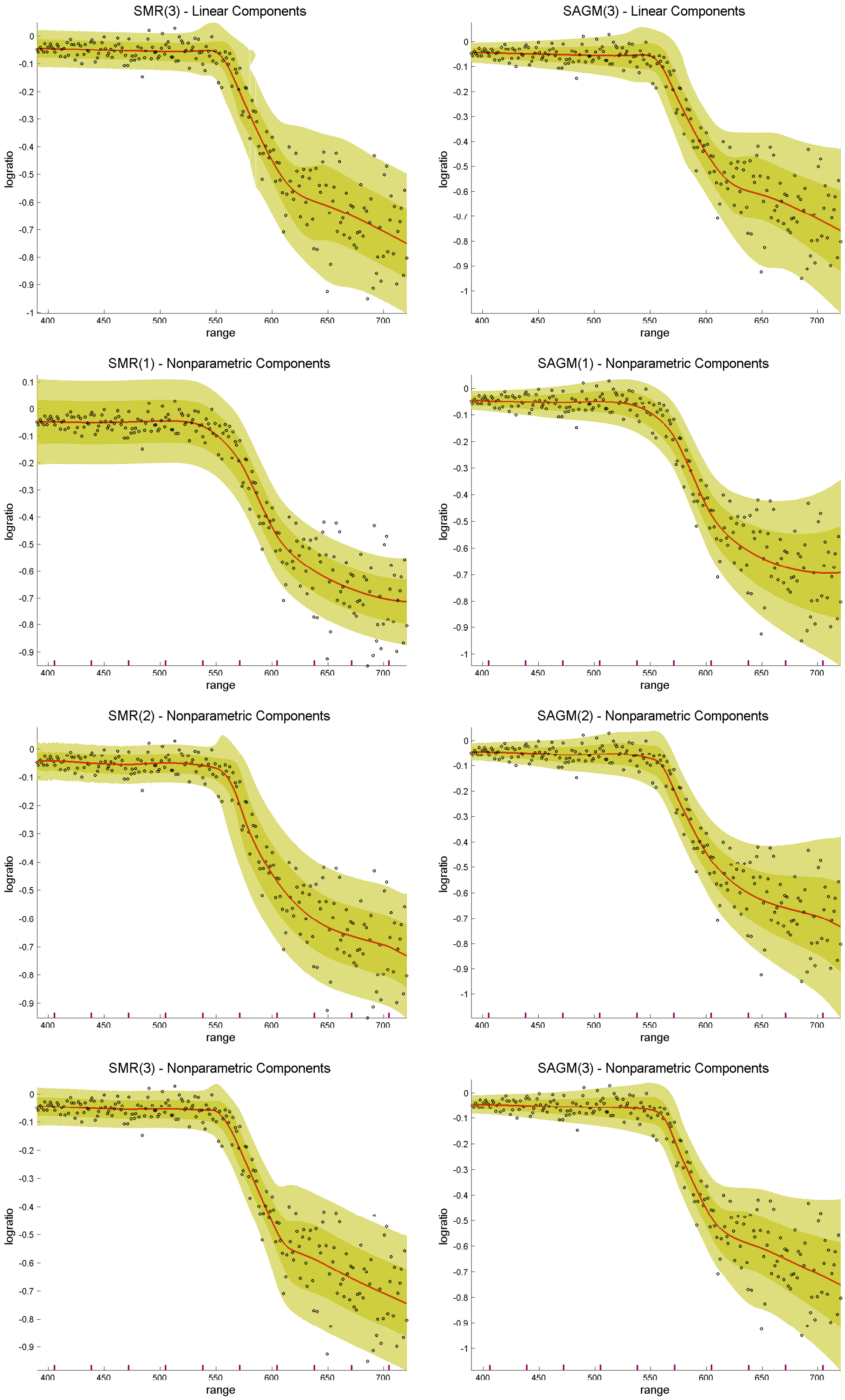

FIGURE 4. The LIDAR data overlayed on 68 and 95 percent HPD predictive intervals. The solid red line is the predictive mean. The thicker tick marks on the horizontal axis locate the knots of the thin plate splines.

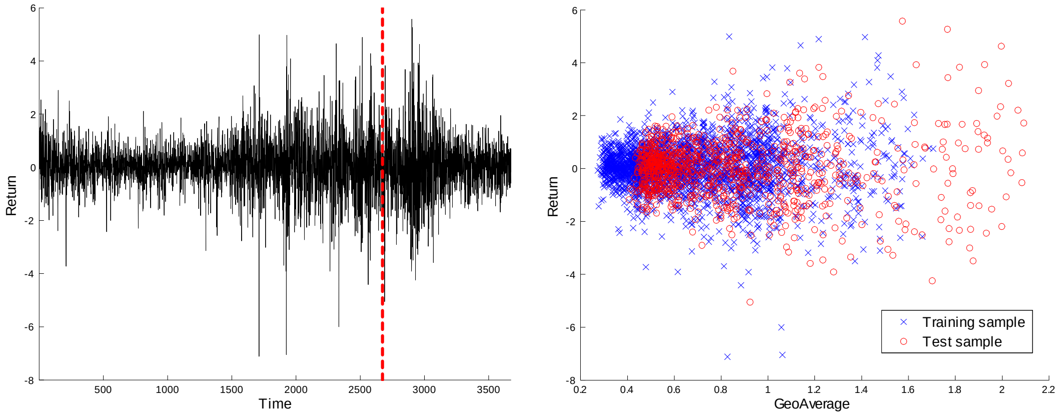

Figure 5. SP500 data. Left: Time plot of Return. Right: Return vs GeoAverage.

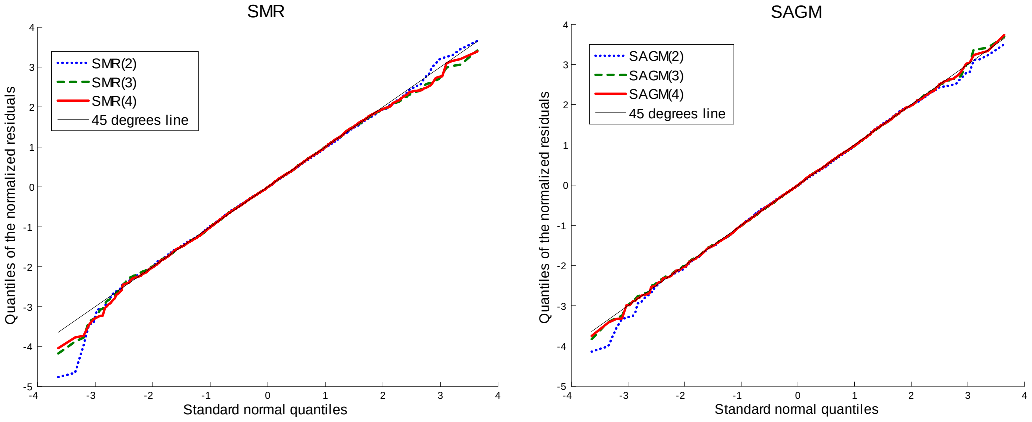

Figure 6. SP500 data. QQ-plots of the normalized residuals.

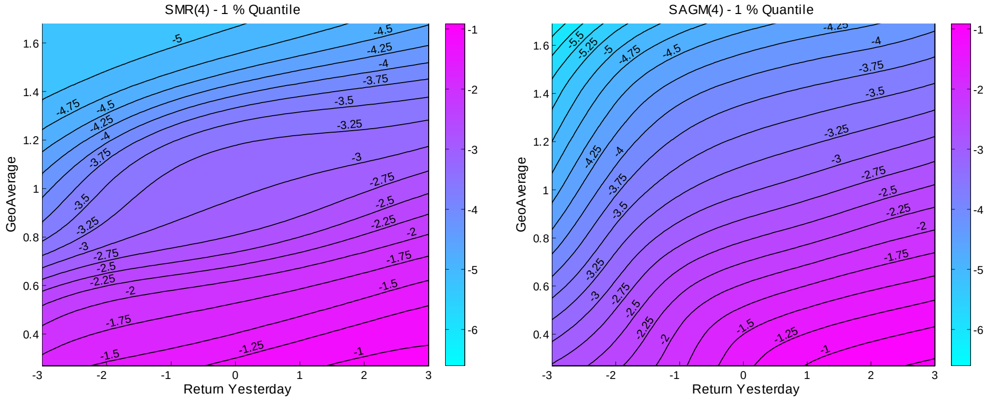

Figure 7. SP500 data. Value at risk (VaR). 1 percent quantile of the predictive density.