SVERIGES RIKSBANK
WORKING PAPER SERIES
211

Nonparametric Regression Density Estimation Using Smoothly Varying Normal Mixtures

Mattias Villani, Robert Kohn and Paolo Giordani

SEPTEMBER 2007

WORKING PAPERS ARE OBTAINABLE FROM

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

Nonparametric Regression Density Estimation Using Smoothly Varying Normal Mixtures

Mattias Villani, Robert Kohn and Paolo Giordani*

Sveriges Riksbank Working Paper Series
No. 211
September 2007

Abstract

We model a regression density nonparametrically 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, Predictive inference, Splines, Value-at-Risk, Variable selection.

*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).

1. Introduction

Nonlinear and nonparametric regression models are widely used in statistics, see e.g. Ruppert, Wand and Carroll (2003) for an introduction. 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 many applications such as 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), and so called *inverse problems* in 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) to the regression density case. Our model is an extension of 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 (experts) where the mixing probabilities are functions of the covariates. This model partitions the space of covariates using stochastic (soft) boundaries. The early machine learning literature used ME models with many simple experts (constant or linear).

Some recent statistical literature takes the opposite approach of using a small number of more complex experts. The most common approach has been to use basis expansion methods (polynomials, splines) to allow for nonparametric experts, see *e.g.* Wood, Jiang and Tanner (2002). 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 experts very flexible and to use extra experts only when they are required.

The ME model with homoscedastic experts can in principle fit heteroscedastic data if the number of experts is large enough. See for example Jiang and Tanner (1999a,b) for some results on approximating the mean function and the density of a generalized linear model by a ME, but it is unlikely to be the most efficient model for that situation. Simulations in Section 3 show that the ME model can have difficulties in modelling heteroscedastic data, and that its predictive performance quickly deteriorates as the number of covariates grows. If the experts themselves are heteroscedastic, we would clearly need fewer of them.

Our article generalizes the ME model by using Gaussian *heteroscedastic* experts with the three components of each expert, i.e. the means, variances and the mixing probabilities, modeled 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 expert probabilities. The centering of the spline basis functions (knots) is 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 ME models as it allows the estimation method to automatically downweight or remove basis functions from an expert in the region where the expert 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 size of the probability of each expert when deciding whether to include a basis function in an expert. 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 experts in the model to reduce sensitivity to the prior.

The first Bayesian paper on ME models 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 homoscedastic Gaussian ME approaches. 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. Our simulations in Section 3 show that such extensions are necessary in some examples. To carry out the inference we develop efficient MCMC samplers which compare favourably to existing MCMC samplers in the (homoscedastic) ME case as well. A comparison with existing samplers is given in Appendix D.

2. The Mixture of Heteroscedastic Experts Model

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 *Mixture of Heteroscedastic Experts (MHE) 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 group/expert membership for the ith observation, v_i is a p-dimensional vector of covariates for the conditional mean of observation i with coefficients, α_j, that vary across the m experts, and w_i is an r-dimensional vector of covariates for the conditional variance of observation i. Expert j's *responsibility/competence* for the ith observation is modelled by a multinomial logit (softmax) gating 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 of regressors for observation i, and $\gamma_1 = 0$ for identification. The three sets of regressors, 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 gating 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 gating functions, respectively. We shall denote the original vector of covariate observations from which the basis expansions (v_i, w_i, z_i) were constructed by x_i.

Many of the models in the nonparametric literature are special cases of the MHE 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 polynomials expansions of the covariates. Both of these articles use a multinomial probit gating function. This means that the expert probabilities must be computed by numerical integration, which makes

the evaluation of predictive densities/likelihoods very time-consuming. 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 MHE 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 MHE, 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 expert 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 gating functions, respectively.

There are at least two restrictions on the model that are useful in practical work. First, we may restrict the heteroscedasticity to be the same across experts: $\delta_1 = ... = \delta_m = \delta$. Given that we allow for nonparametric variance and gating 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 experts. That is, either a covariate has a non-zero coefficient in all of the experts or its coefficient is zero for all experts. Our posterior sampling algorithms handle both types of restrictions.

We use the following notation. Let $Y = (y_1, ..., y_n)'$ denotes 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 $Z = (z_1, ..., z_n)'$ denote the $n \times p$, $n \times r$ and $n \times q$ dimensional matrices of covariates expanded from X. The covariates are standardized to zero mean and unit variance to simplify the prior elicitation. Let $s = (s_1, ..., s_n)'$ denote the n-vector of

expert 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 experts. Let $\mathcal{V} = (\mathcal{V}_1, ..., \mathcal{V}_m)$, where $\mathcal{V}_j$ contains the variable selection indicators for the jth expert. 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 expert j is then

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

and IG denotes the inverse Gamma distribution and $\alpha_{\mathcal{V}_j^c}|\mathcal{V}_j$ is the zero vector with probability one. 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 nonparametric ME literature. An a priori positioning of a knot at location κ in covariate space runs the risk that some of the experts may have very low competence in the neighborhood of that point ($\pi_j(\kappa; \gamma) \approx 0$ for at least some j). The coefficients for that knot will then be poorly estimated, or may even be unidentified, for those low-competence experts. To deal with this problem, we use the prior

$$\mathcal{V}_{kj}|\gamma \sim Bern[\omega_\alpha \pi_j(\kappa_k^v; \gamma)], \quad (k = 1, ..., p; \ j = 1, ..., m), \tag{2.3}$$

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 expert's responsibility for the knot decreases. In the limit where the jth expert has zero responsibility for κ_k^v, that knot is automatically excluded from expert 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:

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

The prior on the gating 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 gating function for a given $\mathcal{Z}$. The prior on γ is then assumed to be of the form

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

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 sampling from the joint posterior distribution of the model parameters. Appendices A to C describe two efficient algorithms which automatically include variables selection in all three sets of covariates: mean, variance and gating.

Appendix D shows that our algorithms compare favorably to existing samplers that have been proposed for the (homoscedastic) ME model.

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, but see also the time series example in Section 3.4. 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. Alternatively, importance sampling can be used to compute the LPDS using only draws from the full posterior $p(\theta|y)$ (Gelfand, 1995). This leads to the estimate $\hat{p}(\tilde{y}_b|\tilde{y}_{-b}, x) = 1/[N^{-1} \sum_{i=1}^{N} \{p(\tilde{y}_b|x_b, \theta^{(i)})\}^{-1}]$, for $b = 1, ..., B$, where $\{\theta^{(i)}\}_{i=1}^{N}$ are the MCMC draws from the full posterior $p(\theta|y)$.

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 a prototype of an inverse problem from robotics, e.g. how to set the angles of a robot arm to move the end effector to a specific position. Bishop (2006) generates data from the following simple model to illustrate such a problem: $a_i = b_i + 0.3\sin(2\pi b_i) + u_i$, where b_i are equally spaced points on the interval $[0, 1]$ and $u_i \overset{iid}{\sim} U(0, 1)$. Now, let $y_i = b_i$ and $x_i = a_i$. We generated 1000 observations from this model and fitted several different MHE models to it. The data are plotted in the first column of Figure 1. 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 models. We used truncated quadratic splines (Hastie, Tibshirani and Friedman, 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 gating 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 MHE with a single expert, which clearly is not flexible enough to cannot capture the true density

or the standard deviation. The MHE(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 gating function (the mean and variance are still nonparametric). The terrible fit of this model clearly demonstrates the importance of a flexible gating function. In fact, with a nonparametric gating function it is important that τ_γ is not made too small, for then the gating 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 MHE(3) with nonparametric mean, variance and gating 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 this is needed. Estimating τ_γ will not help here either.

3.2. **Simulated heteroscedastic data.** We now investigate how well the ME model with homoscedastic experts 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 expert 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 ME and MHE models with linear experts. We use cross-validation (see Section 2.3) here even if we know the true DGP to simplify the comparisons of strength of evidence with the real data examples later in this section. 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 ME and MHE models were fit with one to five experts. Figure 2 displays box plots of the difference

in LPDS between the ME models with a given number of experts and the estimated MHE(1) model. With a single covariate the predictive performance of the ME models with $m \geq 3$ is fairly close to that of MHE(1). As the number of covariates grows, the ME model has increasing difficulty in fitting the data, relative to the MHE(1) model, and it seems that its predictive performance cannot be improved by adding more than five experts. There are already some signs of overfitting with five experts. Even with two covariates the evidence is decisively in favor of the MHE(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 ME relative to the MHE(1) was much inferior to the case with five covariates.

We also investigated the consequences of fitting an MHE model when the true DGP is an ME model. 250 data sets were simulated from a five-covariate ME(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 ME(2) and MHE(2) models using 5-fold cross-validation exactly as above. The ME(2) had a higher LPDS than the MHE(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 ($LPDS_{MHE}$-$LPDS_{ME}$) ranged from -1.368 to 0.366, with a median of -0.640, suggesting that the over-parametrized MHE(2) had at best only a marginally worse predictive performance than the true ME(2) model. Note also that variable selection could have been used to exclude covariates in the variance function of MHE, which should improve its performance relative to the ME model.

3.3. **LIDAR data.** Our first real data set comes from a technique that uses laser-emitted light to detect chemical compounds in the atmosphere (LIDAR, LIght Detection And Ranging, 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 experts. 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 experts, the posterior distributions of the δ's are largely over-lapping. 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 3 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 ME model with 3 linear expert (top row) and $1, 2$ and 3 thin plate spline experts (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 gating functions, and variable selection among the knots with $\omega_\alpha = \omega_\delta = \omega_\gamma = 0.2$ as prior inclusion probability. The ME(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 3 displays the fit of the MHE model. The MHE(3) model with linear experts performs rather well. The best fit seems to be given by the MHE model with a single nonparametric expert. It is interesting to see that the overparametrized MHE(2) and MHE(3) models with nonparametric experts do not seem to overfit. This is due to the self-adjusting mechanism provided by the variable selection: the more experts that are added to the model, the fewer the knots in all experts. For example, the MHE(1) expert has a highly non-linear mean, but the experts in the MHE(3) model with nonparametric experts are essentially linear, all the knots in the MHE(3) model have very small inclusion probabilities. The prior in (2.3) is very effective in removing experts' knots in low competence regions, almost all such knots have zero posterior inclusion probability. All the knots in the variance function of the MHE 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) gating function where most of the knots have posterior probabilities in the range 0.2-0.4. This is true for both ME and MHE 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 experts. All three MHE models with nonparametric experts and the MHE(3) model with linear experts give very similar LPDS values. In particular, a single nonparametric heteroscedastic expert is sufficient to fit the data. The ME models need three experts to come close to the LPDS of the MHE model with a single nonparametric expert, and even then do not quite reach it.

3.4. **US stock returns data.** Our second real data example analyzes the distribution of 3674 daily returns on the S&P500 stock market index from January 1, 1990 to January 30, 2004. 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 4. 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 $(1 - 0.95) \sum_{s=0}^{\infty} 0.95^s \, |y_{t-2-s}|$.

Geweke and Keane (2007) conducted an out-of-sample evaluation of the conditional distribution $p(\text{Return} \mid \text{Return Yesterday, GeoAverage})$ where the ME 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 MHE can do a better job by having the heteroscedastic experts capturing the heteroscedasticity in Return so that the mixture can concentrate more heavily on modelling the fat tails.

We fit ME and MHE models with the experts modelled as two-dimensional thin plate spline surfaces. The mean of each expert is restricted to be constant, in line with the literature on stock market data. 20 knots in $\mathbb{R}^2$ are used in both the variance and gating functions. The locations of the 20 knots were chosen by the algorithm in 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 experts as it outperformed the model with separate δ.

Table 2 displays the LPDS for ME and MHE models evaluated on the 1000 most recent trading days as a single test sample. The best model is the MHE(4) model which is more than 6 LPDS units better than the best ME model (the Bayes factor is 415.72). This is decisive evidence in favor of the MHE (Jeffreys, 1961). It is interesting to note that MHE(1) is only slightly inferior to MHE(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 4 (right panel) Return against GeoAverage (the main driver of the heteroscedasticity, see the standard deviation graphs in Figure 6 below) in the training and test sample. It is clear from Figure 4 that a single heteroscedastic expert 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 MHE(1) now performs substantially worse than, for example, the MHE(3) model. The average LPDS difference between the best MHE and the best ME is now smaller (the Bayes factor comparing MHE(3) to ME(4) is 18.92), but the MHE(3) model outperforms the ME(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 MHE model, whereas the performance of the ME is more or less unchanged. The relative support for the MHE model is now dramatically stronger: the Bayes factor comparing the best fitting ME and MHE models is $7.26 \cdot 10^7$ in favor of the MHE model.

Figure 5 display quantile-quantile plots (QQ-plots) of the normalized residuals (see e.g. Leslie et al., 2007) for the ME and MHE 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 expection 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 expert both do a poor job in the tails of the distribution (not shown in Figure 5 for scaling considerations). Adding another expert to the MHE(1) model gives a substantial improvement in fitting the tails, and the fit of the MHE(3) is excellent. The ME model improves as more experts are added, but even the ME(4) model is not able to fully capture the tails of the distribution.

Figure 6 displays contour plots of the posterior mean of the predictive standard deviation (SD) as a function of the two covariates. The estimated SD changes a lot as more experts are added to the ME model, whereas in the MHE model the SD is much more stable as more experts come into play, suggesting that the SD can be captured quite well with a single heteroscedastic expert. It takes four homoscedastic experts to come close to the SD function produced by the MHE(1) model.

To understand better the differences in interpretation between the ME and MHE models, Figure 7 displays the contours of the posterior mean of the gating function for the ME(3) (left column) and MHE(3) (right column) models. The experts in the ME model have been ordered by their variances in descending order from top to bottom. Figure 7 shows that the ME model is using the experts to capture the heteroscedasticity in the data (compare with Figure 6). The interpretation of the MHE model is quite

different, with a global expert (expert no. 2) capturing the bulk of the heteroscedasticity. Expert 1 and 3 in the MHE are much more local and take care of the heavy tails. This has important consequences for the stock trader which we now explore through Value-at-Risk (VaR) analysis. VaR is usually defined as the 1% quantile in the distribution of returns. It provides traders with a form of probabilistic bound on how much money they risk losing from one trading day to another. Figure 8 displays contour plots of the 1% quantile of the predictive distribution. As for the predictive SD, the VaR varies a lot as more experts are added to the ME model. For the MHE the situation is again more stable, but there are larger differences between the MHE models in Figure 8 than between the MHE models in Figure 6. This suggests that while one heteroscedastic expert is enough to capture the variance of the S&P500, at least another expert is needed to model the heavy tails. Finally, we note that there are large differences between even the ME(4) and MHE(4) in the modelling of the 1% quantile of the predictive distribution, for some covariate values the difference is more than 1% from one trading day to the other, which is quite substantial for daily returns.

Appendix A. Variable dimension K-step Newton proposals for Metropolis-Hastings updates

This appendix describes a general method for constructing tailored proposal densities for the Metropolis-Hasting algorithm which are used in Appendices B and C to draw from the non-standard conditional posteriors of $(\gamma, \mathcal{Z})$ and $(\delta, \mathcal{W})$. We first briefly sketch the algorithm when the parameters do not change in dimension. Let θ be a vector of parameters with a non-standard density $p(\theta|y)$ from which we want to sample using the Metropolis-Hastings algorithm. $p(\theta|y)$ may be a conditional posterior density and the following algorithm can then be used as a step in a Metropolis-within-Gibbs algorithm. Gamerman (1997) showed how to modify the Fisher scoring procedure to produce effective proposal densities. Gamerman's procedure runs as follows. Assume

that the target density $p(\theta|y)$ can be written

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

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 MHE variance function, then $\varphi_i = w_i'\delta$. Assume also that the gradient and Hessian of the log posterior are available in closed form. We can now use Newton's method to iterate K steps from the current point θ_c toward the mode of $p(\theta|y)$, thereby obtaining $\hat{\theta}$ and the Hessian at $\hat{\theta}$. Note that $\hat{\theta}$ may not be the mode but is typically close to it already after only $K \leq 3$ Newton iterations. 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 $c > 2$ degrees of freedom:

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

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

Nott and Leonte (2004) generalized Gamerman's (1997) Fisher scoring algorithm to allow for covariate selection in generalized linear models within the exponetial family. Their method was also used in Leslie et al. (2007). We present their algorithm in a more general setting which is not restricted to the exponential family. The p-dimensional parameter vector θ is accompanied by a vector of binary covariate selection indicators $\mathcal{J} = (j_1, ..., j_p)$. Here we need to propose θ and $\mathcal{J}$ simultaneously, and we will do so from 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 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 then becomes

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

It should be noted that the proposal density in 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 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. The idea is to exploit 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 can then be written

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

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). Note also that after the first Newton iteration the parameter no longer changes in dimension, and the generalized Newton algorithm in (A.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.

Appendix B. Gibbs sampler for the MHE model

This appendix describes the updating steps of the Gibbs sampler in detail. We make use of the following transformation from a heteroscedastic regression to a homoscedastic one with δ as the heteroscedasticity parameter vector

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

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 expert is immediate. We use the following notation. Let n_j denote the number of observations allocated to the jth expert for a given s. V_j denotes the $n_j \times p$ submatrix containing

the rows of V corresponding to the jth expert'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

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

where $\tilde{V}_j$ is the covariate matrix for the jth expert 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 (see Appendix A) to generate from it. The gradient and Hessians 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 experts 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 follows from a direct application of the general method in Appendix A. Note also that variable selection has the advantage of keeping down the dimension of δ in every iteration of the algorithm, which speeds up the algorithm and increases the MH acceptance probability.

Updating γ and $\mathcal{Z}$

We first consider the case without covariate selection. The full conditional posterior of the multinomial logit parameters $\gamma = (\gamma_2', ..., \gamma_m')'$ is of the form

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

which is a non-standard density. It is again possible to derive the gradient and Hessian of this conditional posterior density in closed form, and use the K-step Newton proposal to sample 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 expert 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. Note that D does not enter the Hessian, so the Hessian is equal to its expected value. To handle covariate selection in the gating function we can apply the generalized K-step Newton algorithm in Appendix A. The matrix A_k in the generalized Newton update is here 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 Newton algorithm. The matrix B_k and the vector g_k in Appendix A are defined analogously. Note also that when the prior for $\mathcal{V}$ depends on the value of the gating function at the knots (see Section 2.2), then the conditional posterior of γ equals the expression in (B.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 experts.

Updating s

The expert indicator, s_i $(i = 1, ..., n)$ are independent conditional on the other model parameters, and can therefore be drawn all at once. 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 $10,000$ Gibbs sampling draws after a burn-in of $2,000$ draws. We use $K = 3$ Newton steps in the updating of δ and γ, and $c_\delta = 10$ and $c_\gamma = 10$ degrees of freedom in multivariate-t Newton-based proposal densities for δ and γ. The expert allocation is initialized with the k-means clustering algorithm.

Appendix C. A collapsed sampler for the MHE model

An alternative algorithm, which we refer to as the *collapsed Gibbs sampler*, simulates from the joint posterior using the 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 h may 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. The latter simulation is straightforward and we will only give the details of sampling from $p(\gamma, s, \delta|\mathcal{D})$. Liu, Wong and Kong (1995) prove in a general setting that sampling schemes based on collapsing (integrating out) are expected to be more efficient than pure Gibbs sampling schemes, and we present some support for this claim in Appendix D. The collapsed Gibbs sampler is for most problems more time-consuming than the Gibbs sampler in Appendix B and the increased efficiency must be weighed against increased computing time. 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.

Updating γ

This step is exactly 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 expert 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

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

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 $p(Y|X, s, \gamma, \delta)$ splits up into m marginal likelihoods, one for each expert. We refer to $p(Y|X, s, \gamma, \delta)$ as the marginal likelihood and $p(Y_j|X_j, s, \gamma, \delta)$ as

expert j's marginal likelihood. $p(Y_j|X_j, s, \gamma, \delta)$ can be efficiently computed as follows. Let R_j be the upper 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 (C.1) nm times for a single update of all expert allocations. Fortunately, the change from one computation to the next consists of a simple re-allocation of a single observation from one expert 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 expert j^* to expert j. This requires that we modify the Cholesky 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 expert 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 expert j^*, which is called an *update* of the Choleski with $\tilde{v}_i$).

Even with the 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 expert as before, then the acceptance probability is unity), and ii) whenever a change of expert 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 quite substantial reduction since in the typical case $n_c << n$. There are many ways to propose s. Among them is to propose from

the gating 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.

Appendix D. A comparison of MCMC algorithms on the LIDAR data

Table 5 reports performance summaries of the MCMC algorithms for the MHE(3) model with linear experts for the LIDAR data. The inefficiency factor (IF) in Table 5 is a commonly used measure of numerical efficiency for MCMC samplers and it is defined as $1 + 2\sum_{k=1}^{K} \rho_k$, where ρ_k is the autocorrelation at the kth lag in the MCMC chain for a given parameter and K is an upper limit of the lag length such that $\rho_k \approx 0$ for all $k > K$. The IF approximates the ratio of the numerical variance of the posterior mean from the MCMC chain to that from hypothetical iid draws. The collapsed Gibbs sampler where every s_i is generated directly from its conditional posterior (Collapse) is the most efficient, but it is also the most time-consuming.[1] The collapse Gibbs samplers with MH updating of the s_i (Collapse-Gating proposes s from the gating function, Collapse-Adapt is the adaptive scheme discussed in Appendix C) are substantially faster and, at least the adaptive version is almost as efficient as the pure collapsed algorithm. The fastest algorithm is the Geweke-Keane sampler which is based on the multinomial probit gating function augmented with latent utilities. The Geweke-Keane algorithm gives large IFs

[1]We used Matlab 7 on a 2 GHz Pentium M processor.

when used on the LIDAR data, however. The fact that latent variable augmented sampling schemes can be highly inefficient has been documented in the literature. One recent example for the probit regression is Del Moral, Doucet and Jasra (in press). In our experience, the Geweke-Keane sampler is least efficient when at least some of the gating function parameters are large in absolute value, i.e. when the experts are fairly sharply separated in covariate space. Note also that these inefficiencies spill over to the mean parameters. It may be that these inefficiencies are not too bad if one only cares about the predictive density, but they will matter when the model is interpreted. It is interesting to compare the time to obtain the equivalent to a 1000 iid draws from the algorithms. The median of this time over the parameters is 430.23 seconds for Geweke-Keane, 72.54 for Gibbs Logit, and 363.94, 110.96, and 105.01, for the three Collapse samplers. The best compromise between computing time and efficiency on the LIDAR data is therefore obtained from the Gibbs sampler, closely followed by Collapse-Gating and Collapse-Adapt. The results in Table 5 were obtained using $K = 3$ Newton steps in the updating of δ and γ. We also ran the Gibbs sampler with $K = 1$ for δ and $K = 3$ for γ. This sampler is faster (152.23 iterations per second), with no notable efficiency loss (the IFs and the MH acceptance probabilities were unchanged). Finally, we also tried $K = 1$ for both δ and γ. Here the MH acceptance probability for γ droped to 45.66% and the IFs doubled for γ; the IFs for the other parameters were essentially unchanged. This sampler generates 188.23 draws per second. We have found in general that $K = 1$ is sufficient for δ, whereas the parameters in γ may require $K = 3$, at least when the experts clearly divides the covariate space (c.f. the experience with the Geweke-Keane sampler above).

Appendix E. A simple 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 p.s.d. 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 E.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 in comparison to lower density regions; $\alpha = 1/p$ is a good choice. We typically use a root-finding algorithm to search for the global radius ϵ that gives exactly a pre-specified number of knots.

References

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

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

[9] Gelfand, A. E. (1995). Model determination using sampling-based methods, in *Markov Chain Monte Carlo in Practice*, eds. Gilks, W. R., Richardson, S., and Spiegelhalter D. J., Chapman & Hall, London.

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

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

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

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

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

[15] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*, Springer, New York.

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

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

[24] 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.

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

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

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

[28] 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.

[29] 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.

[30] 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.

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

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

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

[34] 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		
	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$
ME	26.564	59.137	63.162	48.399	61.571	62.985
MHE	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.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
ME	-1579.16	-1430.39	-1413.96	-1410.50	-1410.92
MHE	-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.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
ME	-1058.85	-955.97	-945.69	-942.01	-942.02
MHE	-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.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
ME	-1579.16	-1428.05	-1412.02	-1412.83	-1414.11
MHE	-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.

Parameter	Geweke-Keane	Gibbs	Collapse	Collapse-Gating	Collapse-Adapt
α_{11}	160.20	5.19	3.42	5.33	6.12
α_{21}	295.54	7.45	4.24	8.40	7.92
α_{12}	141.87	11.45	6.56	17.48	10.35
α_{22}	93.07	7.43	4.33	10.09	7.36
α_{31}	31.92	2.97	2.30	2.84	2.81
α_{32}	29.02	2.75	2.17	2.68	2.50
σ_1	4.17	3.34	1.97	2.36	2.52
σ_2	16.52	14.77	8.11	14.95	10.29
σ_3	11.33	12.58	5.68	5.82	6.03
γ_{11}	764.12	12.33	10.95	15.92	15.05
γ_{21}	1603.80	19.61	20.21	53.68	22.79
γ_{12}	868.54	17.18	15.99	16.97	17.90
γ_{22}	1711.63	5.45	6.81	5.90	6.83
δ	13.57	15.43	10.47	9.89	10.72
MH acc. prob. δ	91.22	91.48	99.79	99.81	99.82
MH acc. prob. γ	–	64.88	64.75	64.42	64.60
Iterations per sec	273.04	130.28	16.816	82.418	72.758

TABLE 5. LIDAR data. Inefficiency factors and computing times with the different MCMC algorithms for the MHE(3) model with linear experts.

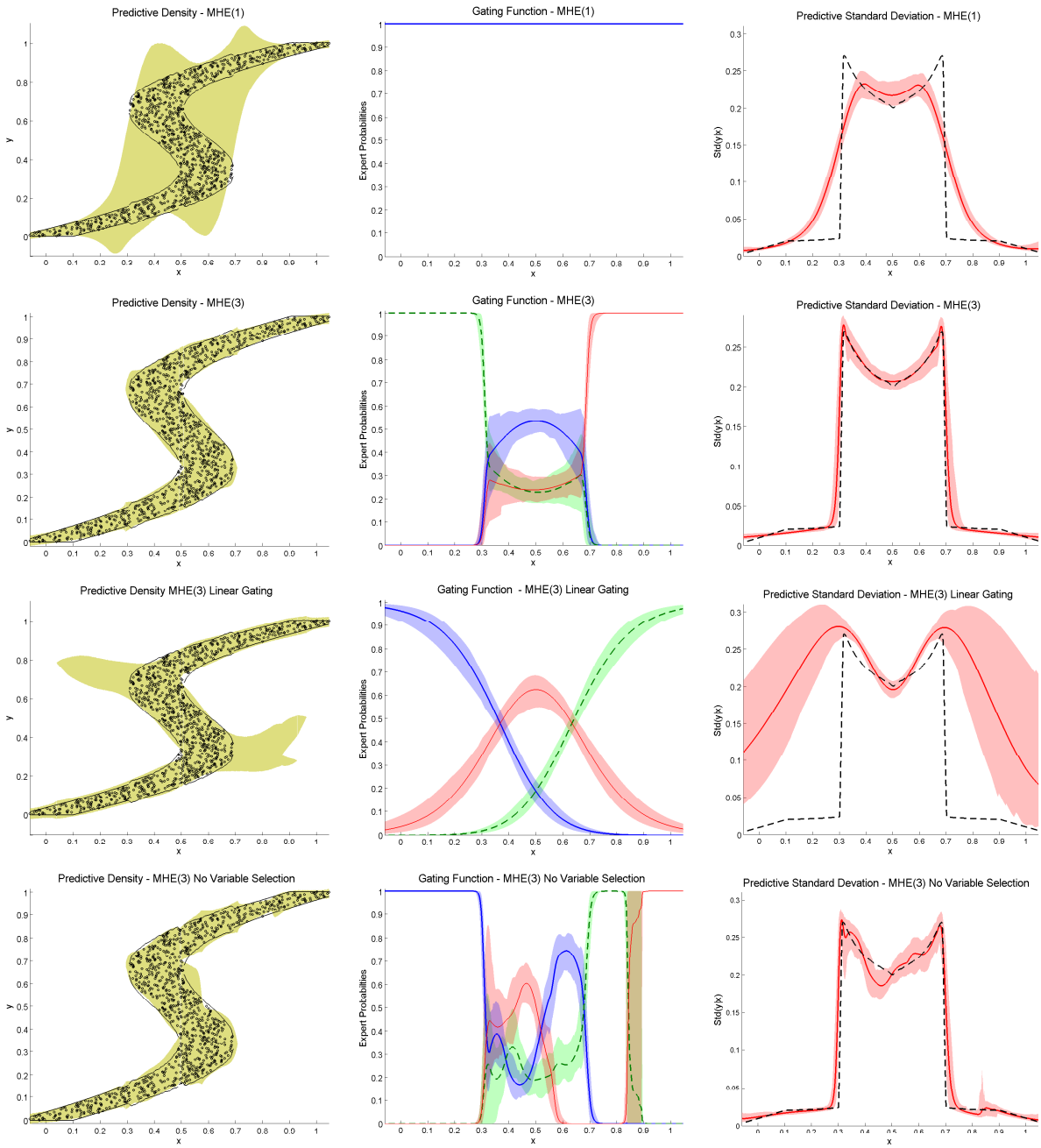

FIGURE 1. Inverse problem data. First column displays the data and the 95 percent HPD intervals in the predictive density. The second and third column depict the gating and predictive standard deviation function, respectively. The rows correspond to four different MHE models.

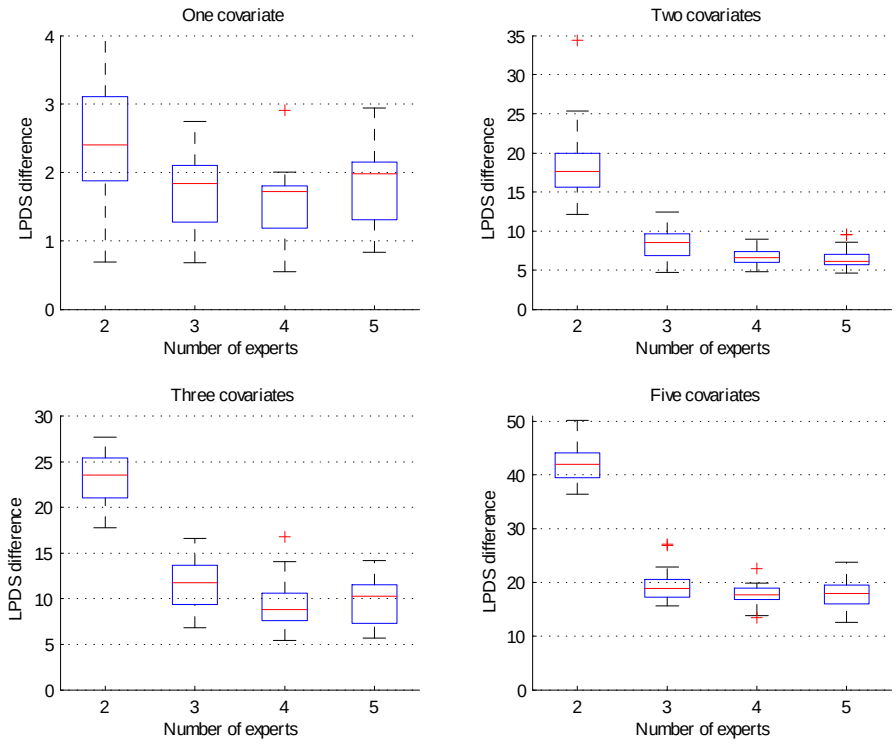

FIGURE 2. Simulated heteroscedastic data. Box plots of the difference in log predictive score (LPDS) between the estimated MHE(1) model and the ME model as a function of the number of expert in the ME model.

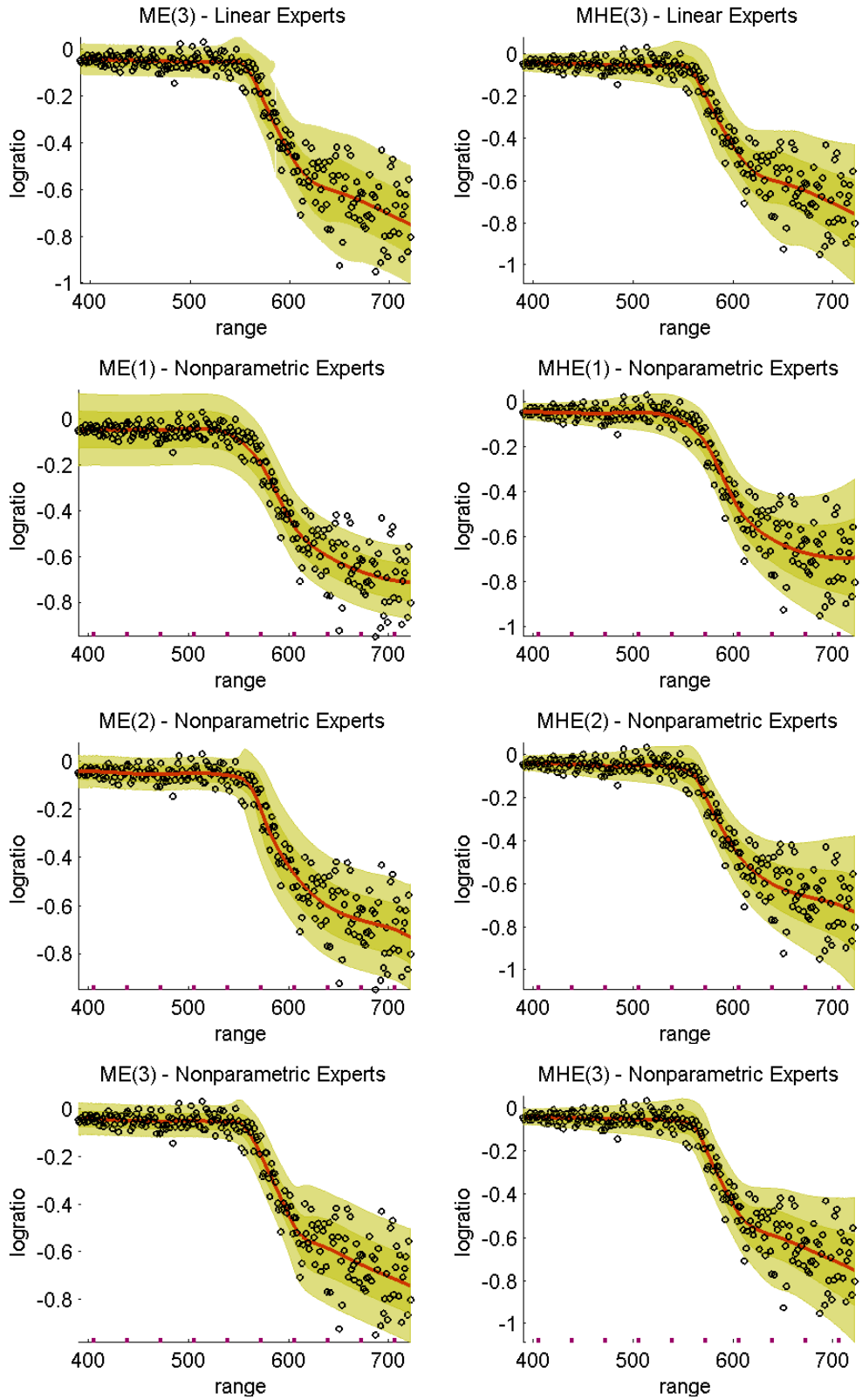

FIGURE 3. The LIDAR data overlayed on 68 and 95 percent HPD predictive intervals. The solid red line is the predictive mean. The thicker tick marks on the horizontal axis locate the knots of the thin plate splines.

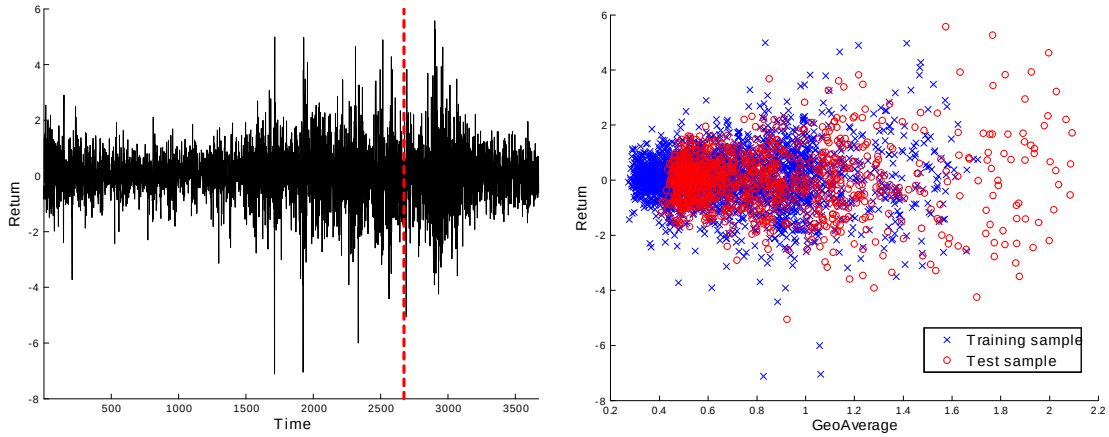

FIGURE 4. SP500 data. Time plot of Return with training and test sample separated by vertical dashed line (left) and scatterplot of Return vs GeoAverage (right).

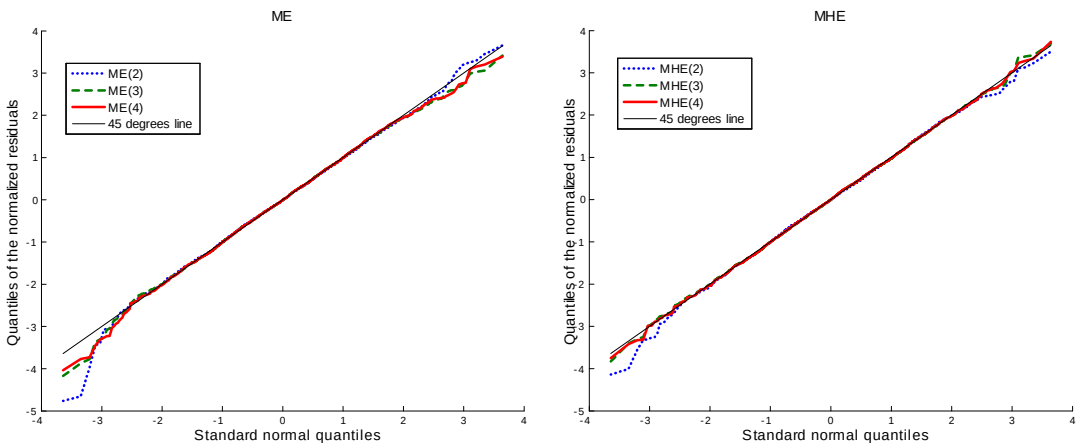

FIGURE 5. SP500 data. QQ-plots of the normalized residuals.

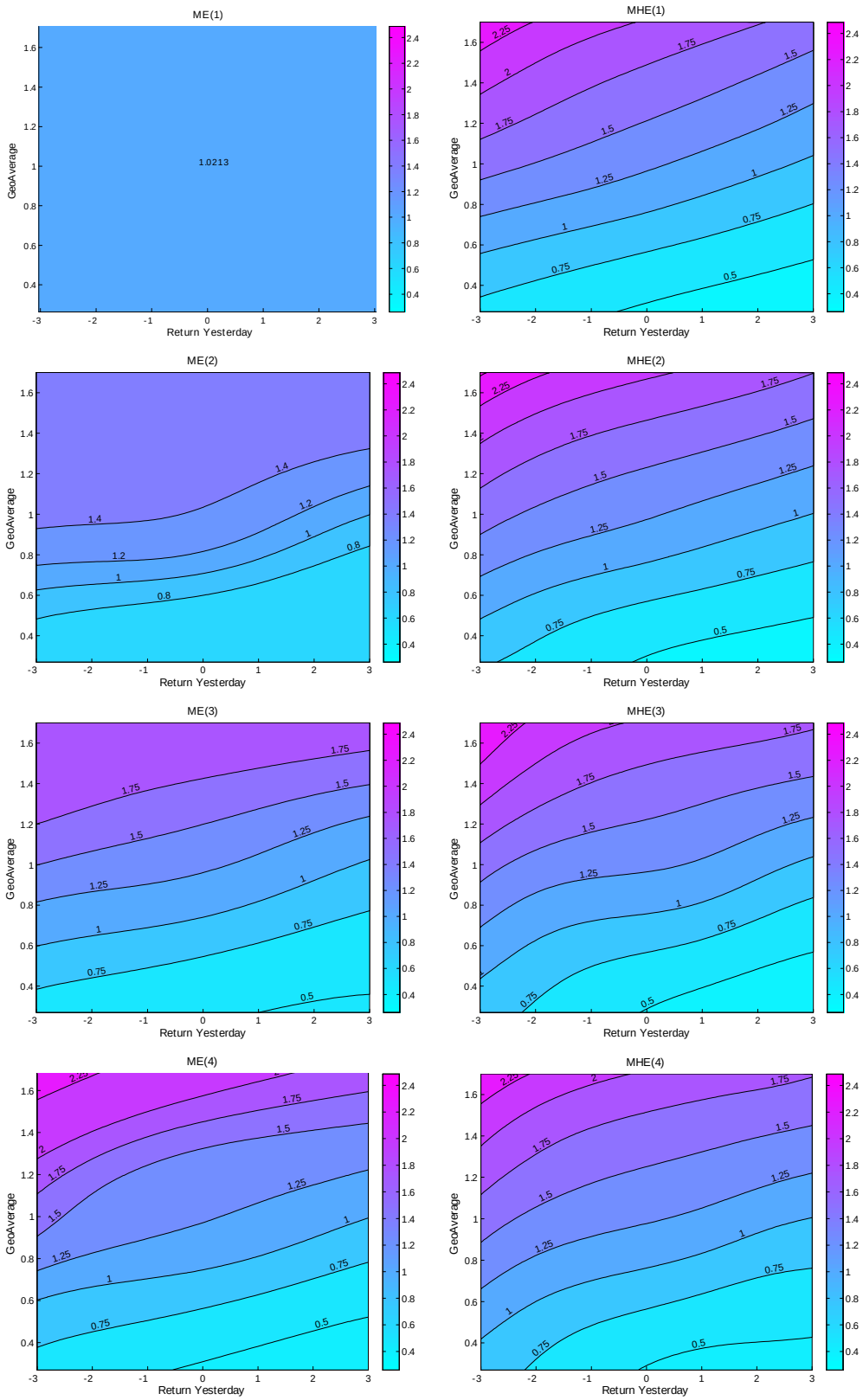

FIGURE 6. SP500 data. Contour plots of the predictive standard deviation as a function of the covariates for the ME (left column) and MHE (right column) models.

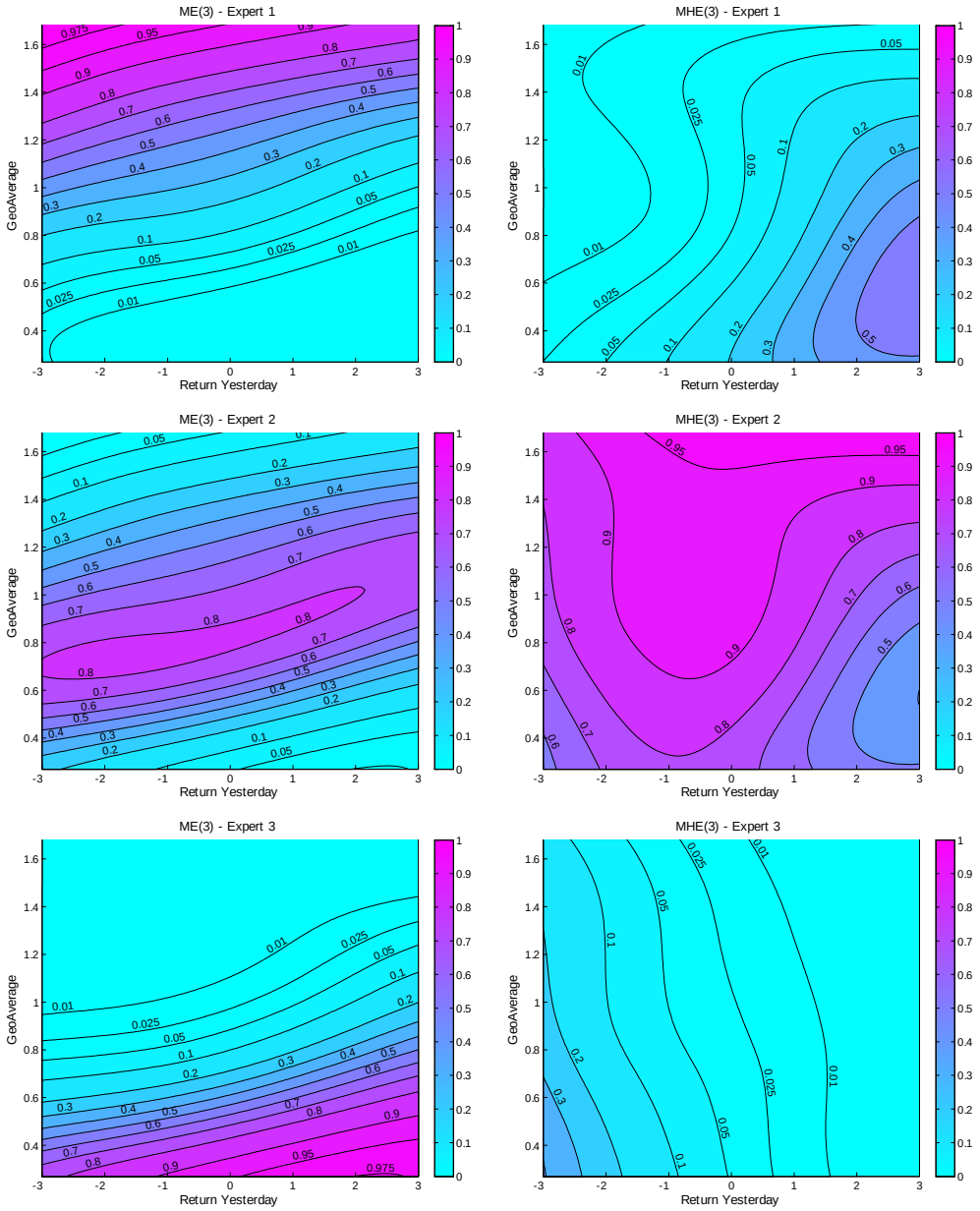

FIGURE 7. SP500 data. Posterior mean of the gating function for the ME(3) (left column) and the MHE(3) (right column) models. The experts in the ME(3) model are ordered in decreasing variance from top to the bottom.

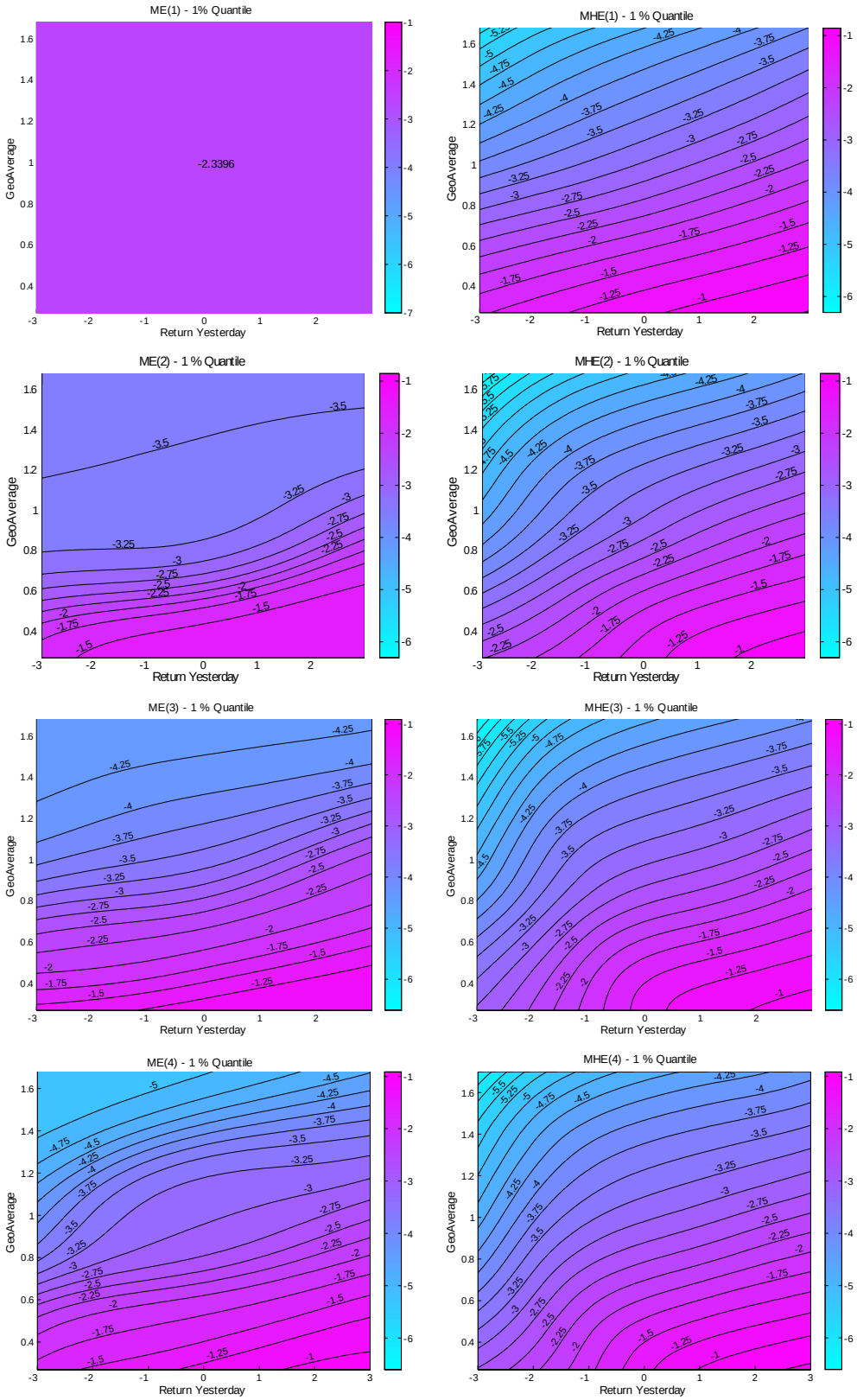

FIGURE 8. SP500 data. Value at risk (VaR) analysis. Contour plots of the 1 percent quantile of the predictive distribution.

Earlier Working Papers:

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

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

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

ISSN 1402-9103 E-Print, Stockholm 2007