Bayesian Modeling of Conditional Densities

Feng Li

Bayesian Modeling of Conditional Densities

Feng Li

Abstract

This thesis develops models and associated Bayesian inference methods for flexible univariate and multivariate conditional density estimation. The models are flexible in the sense that they can capture widely differing shapes of the data. The estimation methods are specifically designed to achieve flexibility while still avoiding overfitting. The models are flexible both for a given covariate value, but also across covariate space. A key contribution of this thesis is that it provides general approaches of density estimation with highly efficient Markov chain Monte Carlo methods. The methods are illustrated on several challenging non-linear and non-normal datasets.

In the first paper, a general model is proposed for flexibly estimating the density of a continuous response variable conditional on a possibly high-dimensional set of covariates. The model is a finite mixture of asymmetric student-t densities with covariate-dependent mixture weights. The four parameters of the components, the mean, degrees of freedom, scale and skewness, are all modeled as functions of the covariates. The second paper explores how well a smooth mixture of symmetric components can capture skewed data. Simulations and applications on real data show that including covariate-dependent skewness in the components can lead to substantially improved performance on skewed data, often using a much smaller number of components. We also introduce smooth mixtures of gamma and log-normal components to model positively-valued response variables. In the third paper we propose a multivariate Gaussian surface regression model that combines both additive splines and interactive splines, and a highly efficient MCMC algorithm that updates all the multi-dimensional knot locations jointly. We use shrinkage priors to avoid overfitting with different estimated shrinkage factors for the additive and surface part of the model, and also different shrinkage parameters for the different response variables. In the last paper we present a general Bayesian approach for directly modeling dependencies between variables as function of explanatory variables in a flexible copula context. In particular, the Joe-Clayton copula is extended to have covariate-dependent tail dependence and correlations. Posterior inference is carried out using a novel and efficient simulation method. The appendix of the thesis documents the computational implementation details.

Keywords: Bayesian inference; density estimation; smooth mixtures; surface regression; copulas; Markov chain Monte Carlo.

ISBN 978-91-7447-665-1

Printed in Sweden by US-AB, Stockholm 2013

Distributor: Department of Statistics, Stockholm University

To my parents

献给我的父母

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: Li, F., Villani, M. and Kohn, R. (2010), "Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities", *Journal of Statistical Planning and Inference* **140**(12), 3638–3654.

PAPER II: Li, F., Villani, M. and Kohn, R. (2011), "Modeling conditional densities using finite smooth mixtures", *in* K. Mengersen, C. Robert and M. Titterington, eds, 'Mixtures: estimation and applications', John Wiley & Sons, Chichester, pp. 123–144.

PAPER III: Li, F. and Villani, M. (2013), "Efficient Bayesian multivariate surface regression", *Scandinavian Journal of Statistics* **in press** .

PAPER IV: Li, F. (2013), "Modeling covariate-contingent correlation and tail-dependence with copulas", *Manuscript* .

Contents

Acknowledgements

Time has flown since I began this journey on the first day. My odyssey during my PhD study now concludes here. I was extremely lucky for having the opportunity to engage with all the wisdom and for being a member at Department of Statistics, Stockholm University.

I would like to express my deepest sense of gratitude to my supervisor, Professor Mattias Villani. I could never have done this without you. Your unlimited knowledge and support with patience, kindness and generousness guided me through step-by-step to the right direction. Your very high academic standard inspires me all the time.

I am grateful to my coauthor Professor Robert Kohn from University of New South Wales, regarding my first and second papers, who gave helpful comments to my papers. I wish to express my gratitude to my assistant supervisor Professor Daniel Thorburn for the fruitful discussions, mostly during lunches, and beneficial comments to my thesis.

I would also like to show my appreciation to Professor Fan Yang Wallentin, who introduced me to the PhD study in this lovely city Stockholm. Special thanks go to Professor Dietrich von Rosen. It was such a joy to have so many interesting conversations with you. Professor Dietrich von Rosen and Tatjana von Rosen also gave me valuable suggestions and help for my future career.

Thanks to Department of Statistics at Stockholm University for financial support during my period of study. My gratitude goes to all of my colleagues, former and present at the department. Professor Gebrenegus Ghilagaber is always very supportive to my academic initiatives. Thanks to Håkan who constantly supplied me the best computer for running simulations.

Thanks go to all my friends during my PhD journey in Sweden. Without you guys, the trip would not have been so colorful. I would like to mention a few friends in particular. Bertil, I am so thankful for your hospitality that every time I visit you, and so much fun we have together. Matias, do not forget those days in Southampton and Kyoto and our nice bull session. My office mates, Annika and Karin, I have truly enjoyed all the moments with you. Yuli and Chengcheng, thanks a lot for sharing the wonderful time on- and off-campus about statistics or about cuisines. Time would not stop me but promote me to think of those good old days. Bergrún, Ellinor, Jessica, Linda, Nicklas, Pär, Sofia, Tea, Olivia and all other people, I would never forget those gym days, movie nights, and all the fantastic, crazy and joyful entertaining time with you.

Finally, I am deeply indebted to my family, mum and dad, who are perpetually understanding and encouraging no matter where I am. I owe a great debt of gratitude to my lovely little sister, Minmin, who is always around with my parents during my time abroad.

Stockholm, 2013

Feng

1. Introduction and background

1.1 Motivating flexible Bayesian modeling

Statistical methods have been developed rapidly in the past twenty years. One driving factor of this development is that more and more complicated high-dimensional data require sophisticated data analysis methods. A noticeably successful case is the machine learning field which is now wildly used in industry. Another reason are the dramatic advancements in the statistical computational environment. Computationally expensive methods that in the past couldonly be run on expensive super computers are now possible to run on a standard PC. This has created an enormous momentum for Bayesian analysis where complex models are typically analyzed with modern computer-intensive simulation methods.

Traditional linear models with Gaussian assumptions are challenged by the new large complicated datasets, which have in turn generated interest in new approaches with flexible model with less restrictive assumptions. Moreover, research has shifted the attention from merely modeling the mean and variance of the data to sophisticated modeling of skewness, tail-dependence and outliers. However such work demands efficient inference tools. The development of highly efficient Markov chain Monte Carlo (MCMC) methods has reduced the barrier. Moreover, the Bayesian approach provides a natural way for prediction, model comparison and evaluation of complicated models, and has the additional advantage of being intimately connected with decision making.

1.2 Bayesian inference

In Bayesian statistics, inference of an unknown quantity θ combines data information y with prior beliefs about θ via Bayes' formula

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\mathrm{d}\theta}$$

where $p(y|\theta)$ is the likelihood function and $p(\theta)$ is the prior knowledge of θ and $\int p(y|\theta)p(\theta)\mathrm{d}\theta$ is also know as the *marginal likelihood* or *prior predictive distribution*. In many simple statistical models with vague priors, Bayesian inference draws similar conclusions to those obtained from a traditional frequentist approach, see e.g. Gelman et al. (2004). The Bayesian approach is however more easily to extended to more complicated models using MCMC simulation techniques.

In all but the most simplistic models, the posterior distribution is analytically intractable and Markov chain Monte Carlo (MCMC) algorithms are used for sampling the posterior distribution

$p(\theta|y)$. The Metropolis-Hastings algorithm draws from the Bayesian posterior distribution of θ by generating random draws from a proposal distribution and accepts each draw with a certain probability. The efficiency of Metropolis-Hastings algorithm depends how well the proposal distribution approximates the true posterior. The Gibbs sampler is a special case of Metropolis-Hastings algorithm in which the proposal draws are simulated from the full conditional posterior and are accepted with probability one. When drawing from the posterior in complicated models one usually needs to mix different algorithms. Metropolis-Hastings within Gibbs is one of such combinations where the subsets of the posterior parameter vector θ are sampled using the Gibbs sampler with each parameter subset drawn via Metropolis-Hastings algorithm.

1.3 Density estimation

In statistics, density estimation is the procedure of estimating an unknown density $p(y)$ from observed data. The very early stage of density estimation techniques traces back to the usage of histograms, later followed by kernel density estimation in which the shape of the data is approximated through a kernel function with a smoothing parameter (*bandwidth*), see e.g. Silverman (1986). However due to the difficulty in specifying the bandwidth in kernel density estimation, mixture models have become a popular alternative approach, see Frühwirth-Schnatter (2006) for a textbook treatment. The mixture densities are usually written as

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

where $\sum_{k=1}^{K} \omega_k = 1$ for non-negative mixture weights ω_k and $p_k(x|\theta_k)$ are the component densities. When $n < \infty$, the mixture is said to be finite. If $K = \infty$, it is called an infinite mixture, the Dirichlet process mixture being the most prominent example, see e.g. Hjort et al. (2010).

One important property is that the moments of the mixture density are easily obtained through the moments of its mixture components. If the m:th central moment exists for all of its component densities, the m:th central moment for the finite mixture density exists and is of the form

$$E((y-\mu)^m|\theta) = \sum_{k=1}^{K} \sum_{i=1}^{m} \omega_k \binom{m}{i} E((y-\mu_k)^i|\theta_i)$$

where μ_k is the mean of k:th density component. Mixture densities can be used to capture data characteristics such as multi-modality, fat tails, and skewness. Zeevi (1997) uses mixture densities to approximate complicated densities. See Figure 1.1 for an example with a mixture of normal densities. For other properties of mixtures, see Frühwirth-Schnatter (2006).

1.3.1 Conditional density estimation

The conditional density estimation concentrates on modeling the relationship between a response y and set of covariates x through a conditional density function $p(y|x)$. In the simplest case, the

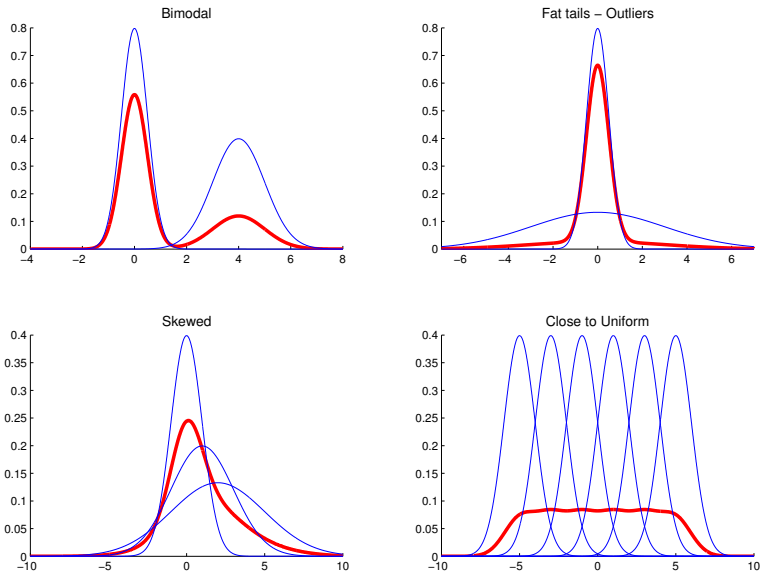

Figure 1.1: Using mixture of normal densities (thin lines) to mimic a flexible density (bold line)

Gaussian linear regression $y = x'\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ is trivially equivalent to modeling $p(y|x)$ by a Gaussian density with mean function $\mu = x'\beta$ and constant variance σ^2.

Mixtures of conditional densities is the obvious extension of mixture models to the conditional density estimation problem:

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

where $p_i(y|x)$ is the conditional density in i:th mixture component. A simple case is the mixture of homoscedastic Gaussian regression models with constant mixture weights. The limitation of this model is that it restricts the shape of the distribution to be the same for all x. A smooth mixture is a finite mixture density with weights that are smooth functions of the covariates

$$\omega_k(x) = \frac{\exp(x'\gamma_k)}{\sum_{i=1}^{K} \exp(x'\gamma_i)}.$$

This model allows the density shape to be different for different x values. Villani et al. (2009) propose the mixture of heteroscedastic Gaussian model with smooth weight functions. Norets (2010) shows that large classes of conditional densities can be approximated in the Kullback-Leibler distance by finite smooth mixtures of normal regressions.

In conditional density estimation, an important focus is modeling the regression mean $E(y|x)$. A spline is a popular approach for nonlinear regression that models the mean as a linear combination of a set of nonlinear basis functions of the original regressors,

$$y = f(x) + \varepsilon = x'\beta + \sum_{i=1}^{k} x(\xi_i)'\beta_i + \varepsilon$$

where k is number of basis functions $x(\xi)$ used and ξ_i is the location of i:th basis function, often referred to as a knot. Each basis function is defined by a knot ξ_i in covariates space and the knots determine the points of flexibility of the fitted regression function. In the case with multiple covariates $x_1,...,x_q$ it is common to assume additivity

$$y = \sum_{j=1}^{q} f_j(x_j) + \varepsilon,$$

where $f_j(x_j)$ are spline functions. The more general surface model does not assume additivity and uses a multi-dimensional basis function with interactions among the covariates. It is possible to have both additive and interactive splines in the regression.

1.3.2 Multivariate density estimation

The multivariate density estimation and conditional density estimation are analogues of their univariate cases except that the densities $p(\boldsymbol{Y})$ and $p(\boldsymbol{Y}|\boldsymbol{X})$ are multivariate. Therefore, kernel density estimators can be naturally extended to the multivariate case with a multivariate bandwidth matrix, but optimizing the bandwidth matrix is much more difficult. Alternatively, one may use mixture of multivariate densities. Smooth mixture of multivariate regression models and multivariate splines are extensions of conditional density estimation from univariate case to multivariate case. In addition to the methods mentioned above, copula is a more general choice for multivariate density estimation because of its unique feature that a copula function separates the multivariate dependence from its marginal functions, and it is possible to use both continuous and discrete marginal models.

1.3.3 Copula density estimation

In the multivariate density estimation, research diverts into different directions. One of them is to explore the multivariate dependence using *copulas* (Sklar, 1959). Let $F(y_1,...,y_M)$ be a multi-dimensional distribution function with marginal distribution functions $F_1(y_1),\cdots,F_M(y_M)$. Then there exists a function C such that

$$\begin{aligned} F(y_1,...,y_M) =& C(F_1(y_1),...,F_M(y_M)) \\ =& C\left(\int_{-\infty}^{y_1} f_1(z_1)dz_1,...,\int_{-\infty}^{y_M} f_M(z_M)dz_M\right) = C(u_1,...,u_M) \end{aligned}$$

where $C(\cdot)$ is the copula function and $f(\cdot)$ is the density of the marginal distribution $F(\cdot)$. Furthermore, if $F_i(y_i)$ are all continuous for $i \in \{1,...,M\}$, then C is unique. The derivative $c(u_1,...,u_M) = \partial^M C(u_1,...,u_M)/(\partial u_1...\partial u_M)$ is the copula density that corresponds to the multivariate density function.

A nice feature of the copula construction is that it separates the marginal distributions $f_1(y_1),...,f_M(y_M)$ from the dependence structure given by the copula function. For instance, the Gaussian copula

which is obtained from a Gaussian density function can be combined with non-Gaussian, or even discrete, marginal distributions, see e.g. Pitt et al. (2006). In addition, a richer class of multivariate distributions via copula is possible to construct through methods like Laplace transform, mixtures of conditional distributions, and convolution *etc*, with appealing properties.

The dependence properties of copulas have been theoretically studied by Joe (1997) and others. Given a bivariate distribution function $F(y_1, y_2)$ and its copula function $C(u_1, u_2)$, the correlation between two marginal densities can be measured by Kendall's τ

$$\tau = 4\int\int F(y_1, y_2)\mathrm{d}F(y_1, y_2) - 1 = 4\int\int C(u_1, u_2)\mathrm{d}C(u_1, u_2) - 1.$$

Unlike Pearson's correlation that can only measure linear dependence, Kendall's τ is a rank correlation that is invariant with respect to strictly increasing transformations, i.e. the marginal densities do not affect the Kendall's τ if they are strictly continuous. This property makes Kendall's τ a more desirable measure of association for multivariate non-Gaussian distributions. The same property holds for Spearman's ρ. See Joe (1997) for other characteristics of Kendall's τ for different copula densities. For example, for copulas generated via the Laplace transform, which are also known as Archimedean copulas, Kendall's τ can be written as

$$\tau = 1 - 4\int_0^\infty s(\phi'(s))^2\mathrm{d}s$$

where $\phi'(s)$ is the first order derivative of the Laplace transform $\phi(s)$.

In addition to correlation, dependence in the tail is also important in many applications. Tail-dependence measures the extent to which several variables simultaneously take on extreme values. The lower tail-dependence λ_L and the upper tail-dependence λ_U can be defined in terms of copulas in the bivariate case

$$\lambda_L = \lim_{u\to 0^+} Pr(X_1 < F_1^{-1}(u)|X_2 < F_2^{-1}(u)) = \lim_{u\to 0^+} \frac{C(u,u)}{u},$$
$$\lambda_U = \lim_{u\to 1^-} Pr(X_1 > F_1^{-1}(u)|X_2 > F_2^{-1}(u)) = \lim_{u\to 1^-} \frac{1-C(u,u)}{1-u}.$$

Not all multivariate copulas generate tail-dependence. The Gaussian copula, for example, has no tail-dependence and the student's t copula generates a rather restrictive tail-dependence as a results of only having a single degrees of freedom parameter for all the modeled variables. In the bivariate copula family, the Joe-Clayton copula has explicit parameters for the lower and upper tail-dependence.

A copula function satisfies the inequalities $L \leq C(u_1, ..., u_M) \leq U$ where $L = \sum_{i=1}^{M} u_i - M + 1$ is Fréchet–Hoeffding lower bound and $U = \min\{u_1, ..., u_M\}$ is Fréchet–Hoeffding upper bound. Note that U is also a copula but L is a copula if $M = 2$. Furthermore, in the bivariate case, if the copula is close to the upper bound, it shows strong positive dependence and if the copula is close to the lower bound, it shows strong negative dependence (Nelsen, 2006).

The conditional density estimation of $p(\boldsymbol{Y}|\boldsymbol{X})$ in terms of a copula is expressed as

$$p(\boldsymbol{Y}|\boldsymbol{X}) = c(u_1|\mathbf{x}_1, ..., u_M|\mathbf{x}_M) \times \prod_{i=1}^{M} p_i(y_i|\mathbf{x}_i)$$

where $p_i(y_i|\mathbf{x}_i)$ is the conditional density in i:th marginal model with covariate vector $\mathbf{x}_i$. The inference for a copula model is similar to the inference methods used for other multivariate models. In particular, the likelihood for copula is written as

$$\prod_{j=1}^{n} c(u_{j1}, ..., u_{jM}) \times \prod_{i=1}^{M} \mathscr{L}_i$$

where $\mathscr{L}_i$ is the likelihood in i:th marginal model.

1.4 Regularization

Variable selection is a technique that is commonly used in regression models. Historically the purposes for using variable selection are to select meaningful covariates that contributes to the model, inhibit ill-behaved design matrices, and to prevent model over-fitting. Methods like backward and forward selections are standard routines in most statistical software packages. However the drawbacks are obvious in those techniques, e.g. the selection depends heavily on the starting points, which becomes more problematic with high dimensional data with many covariates.

Most current methods rely on Bayesian variable selection via MCMC, as introduced by Smith & Kohn (1996); George & McCulloch (1997). A standard Bayesian variable selection approach is to augment the regression model with a variable selection indicator $\mathscr{I}$ for each covariate

$$\mathscr{I}_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0, \end{cases}$$

where β_j is the jth covariate in the model. More informally, this can be expressed as

$$\mathscr{I}_j = \begin{cases} 1 & \text{if the variable } j\text{enters the model} \\ 0 & \text{otherwise.} \end{cases}$$

Variable selection is then obtained by sampling the posterior distribution of all regression coefficient jointly with the variable selection indicators, thereby yielding the marginal posterior probability of variable inclusion $p(\mathscr{I}|\text{Data})$. More recent improved algorithms include (Brown et al., 1998) for large covariate sets and the adaptive scheme for Bayesian variable selection in (Nott & Kohn, 2005). See O'Hara & Sillanpää (2009) for a review of Bayesian variable selection approaches.

For the purpose of overcoming problems with overfitting, shrinkage estimation can also be used as an alternative, or even complementary, approach to variable selection. A shrinkage estimator shrinks the regression coefficients towards zero rather than eliminating the covariate completely. One way to select a proper value of the shrinkage is by cross-validation, which is costly

with big data and complicated models. In the Bayesian approach, the shrinkage parameter is usually automatically estimated together with other parameters in the posterior inference. The *lasso* (least absolute shrinkage and selection operator) (Tibshirani, 1996) approach can be viewed as shrinkage estimator with a Laplace prior (Park & Casella, 2008). Lasso can be shown to perform both shrinkage and variable selection at the same time.

1.5 Bayesian predictive inference and model comparison

Two types of prediction are commonly used in predictive inference. Let Y_b be the testing dataset for evaluating the predictions, and Y_{-b} the training dataset used for estimation. The prediction of Y_b given Y_{-b} is called *in-sample prediction* if $Y_b \in Y_{-b}$ and *out-of-sample prediction* if $Y_b \notin Y_{-b}$. Assuming that the data observations are independent conditional on the model parameters θ, the predictive density can be written

$$p(Y_b|Y_{-b}) = \int \prod_{j=1}^{n} p(Y_{j,b}|\theta)p(\theta|Y_{-b})\mathrm{d}\theta$$

where $p(\theta|Y_{-b})$ is the posterior based on the training dataset Y_{-b} and $\prod_{j=1}^{n} p(Y_{j,b}|\theta)$ is the likelihood for the observations conditional on the model parameters. The predictive density can be viewed as a weighted average of the likelihood with $p(\theta|Y_{-b})$ as the weight function. In time series, the predictive distribution for predicting p period ahead is written differently due the dependence of time,

$$p(Y_{(T+1):(T+p)}|Y_{1:T}) = \prod_{i=1}^{p} \int p(Y_{T+i}|\theta, Y_{1:(T+i-1)})p(\theta|Y_{1:(T+i-1)})\mathrm{d}\theta.$$

Bayesian model comparison have historically been based on the marginal likelihood. It is well-known, however,that the marginal likelihood is very sensitive to the specification of prior. This sensitivity is apparent already from its definition since the marginal likelihood is the expected likelihood where the expectation is taken with respect to the prior. Due to this prior sensitivity, it is becoming more common to have model comparisons based on the log predictive density score (LPDS)

$$\text{LPDS} = \frac{1}{B}\sum_{i=1}^{B} \log p(Y_{b_i}|Y_{-b_i})$$

in which the dataset are partitioned into B subsets, $Y_{b_1}, ..., Y_{b_B}$. The LPDS sacrifices a part of the data, uses that data to train the prior into a more robust posterior, and then uses that posterior to integrate out the model parameters. In cross-sectional data, the data can be partitioned randomly or with a systematic pattern. In time series it is more common to use the past data as the training data and predict the future.

2. Summary of papers

Paper I: Flexible modeling of conditional distributions using smooth mixtures of asymmetric student *t* densities

In this paper we propose a general model for flexibly estimating the density of a continuous response variable conditional on a possibly high-dimensional set of covariates.

The paper introduces a new model class with mixtures of flexible asymmetric student *t* densities (split-*t*)with covariate-dependent mixture weights, also referred to as a smooth mixture. The properties of the split-*t* are studied. The four parameters of the mixture components - the mean, degrees of freedom, scale and skewness -are all modeled as functions of covariates. The modeling philosophy is the *complex-and-few* approach where enough flexibility is used within the mixture components, so that the number of components can be kept to a minimum.

Inference is Bayesian and the computation is carried out using Markov chain Monte Carlo simulation. We use a tailored Metropolis-Hastings-within-Gibbs algorithm for sampling the posterior distribution of the parameters. The number of components in the mixture model are selected via a Bayesian version of out-of-sample cross-validation. To enable model parsimony, a variable selection prior is used in each set of covariates and among the covariates in the mixing weights. We use variable-dimension finite-step Newton proposals in the Metropolis-Hastings algorithm to update coefficients and variable selection indicators efficiently.

The model is applied to analyze the distribution of daily stock market returns of the S&P500 index conditional on nine covariates including the historical returns and volatility measures such as a geometrically decaying average of past absolute returns. The out-of-sample evaluation shows that mixtures of few asymmetric student *t* densities outperforms widely used GARCH models and other recently proposed mixture models during the recent financial crisis. We also investigated estimation stability over different subsamples for the popular *Value-at-Risk* measure.

Paper II: Modeling conditional densities using finite smooth mixtures

In this paper we explore the flexibility of modeling conditional densities using finite smooth mixtures, with particular emphasis on skewed data. We explore how well a smooth mixture of symmetric components can capture skewed data. Simulations and applications on real data show that including covariate-dependent skewness in the components can lead to substantially improved

performance on skewed data, often using a much smaller number of components. Furthermore, variable selection is effective in removing unnecessary covariates in the skewness, which means that there is little loss in allowing for skewness in the components when the data are actually symmetric. We also explore the use of splines in the mixture components and demonstrate the efficiency of variable selection in smooth mixtures on a well known environmental data set from the nonparametric regression literature.

In the simulation study, we analyze the relative performance of smooth mixtures adaptive Gaussian densities and split-t densities by comparing the estimated conditional densities $q(y|x)$ with the true data-generating densities $p(y|x)$ using estimates of both the Kullback-Leibler divergence and the L_2 distance. We find that smooth mixtures with a few complex components can greatly outperform smooth mixtures with many simpler components. Moreover, variable selection is effective in down-weighting unnecessary aspects of the components and makes the results robust to mis-specification of the number of components, even when the components are complex.

We also introduce smooth mixtures of gamma and log-normal components to model positively-valued response variables where the parameters are reparametrized in terms of mean and variance. This reparametrization makes the prior specification easier for practitioners. A large set of model with gamma and log-normal components are compared on a dataset of electricity expenditures in 1602 Australian households.

Paper III: Efficient Bayesian multivariate surface regression

In this paper we further investigate nonparametric modeling for multivariate conditional density estimation using a Gaussian multivariate regression with a mean surface modeled flexibly using a spline surface.

Methods for choosing a fixed set of knot locations in additive spline models are fairly well established in the statistical literature. While most of these methods are in principle directly extendable to non-additive surface models, they are less likely to be successful in that setting because of the curse of dimensionality, especially when there are more than a couple of covariates.

We propose a regression model for a multivariate Gaussian response that combines both additive splines and interactive splines, and a highly efficient MCMC algorithm that updates all the knot locations jointly. We use shrinkage priors to avoid overfitting with different estimated shrinkage factors for the additive and surface part of the model, and also different shrinkage parameters for the different response variables. This makes it possible for the model to adapt to varying degrees of nonlinearity in different parts of the data in a parsimonious way.

We compare the performance of the traditional fixed knots approach to our approach with freely estimated knot locations using simulated data with different number of covariates and for varying degrees of nonlinearity in the true surface. We use shrinkage priors with estimated shrinkage both for the fixed and free knot models, but no variable selection.

We also compare three types of MCMC updates of the knots: i) one-knot-at-a-time updates using a random walk Metropolis proposal with tuned variance, ii) one-knot-at-a-time updates with

the tailored Metropolis-Hastings step, and iii) full block updating of all knots using the tailored Metropolis-Hastings step. The massive efficiency and speed gains from updating all the blocks jointly using a tailored proposal when our algorithm is used comparing to other algorithms.

Moreover, the sensitivity study of the posterior inferences with respect to variations in the prior shows the free knots model is also more robust in the sense that it performs consistently well across different datasets.

Our surface model is illustrated in a finance application where a firms leverage is modeled as a function of the proportion of fixed assets, the firm's market value in relation to its book value, firm sales and profits. It is shown that our approach is computationally efficient, and that allowing for freely estimated knot locations can offer a substantial improvement in out-of-sample predictive performance.

Paper IV: Modeling covariate-contingent correlation and tail-dependence with copulas

In this paper we propose a general approach for modeling a covariate-dependent copula. The copula parameters as well as the parameters in the marginal models are linked to covariates. Our method allows for variable selection among the covariates in the marginal models and in the copula parameters. Posterior inference is carried out using an efficient MCMC simulation method.

We first introduce the reparametrized Joe-Clayton copula where the correlation and lower tail-dependence parameters are used as explicit copula parameters. Our parameterization reduces the effort for specifying the prior information in our Bayesian approach. Most importantly, this parameterization make it possible to directly link correlations and tail-dependence to covariates via separate link functions. We also study some new properties for this copula.

We describe the prior specification for the model in details and we also consider a special situation where the model parameters are variationally dependent of each other. Our solution involves introducing a conditional link function, which is demonstrated in our application to make the MCMC algorithm more robust and gives higher acceptance probability in Metropolis- Hastings algorithm.

We illustrate our covariate-dependent copula model with daily returns from the S&P100 and S&P600 daily stock market indices during the period from September 15, 1995 to January 16, 2013. In the marginal models, we use an asymmetric student's t density in all margins with all four parameters in the model linked to covariates. The use of covariates in the correlation and lower-tail dependence parameters in the copula is shown to improve out-of-sample predictive performance. Moreover, variable selection also enhances the model's predictive performance, and provides interesting insights into which covariates are associated with lower-tail dependence and correlation between the variables.

References

BROWN, P. J., VANNUCCI, M. & FEARN, T. (1998). Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 627–641.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite mixture and Markov switching models*. Springer Verlag.

GELMAN, A., STERN, H. & RUBIN, D. (2004). *Bayesian data analysis*. CRC press.

GEORGE, E. I. & MCCULLOCH, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica* **7**, 339–374.

HJORT, N. L., HOLMES, C., MÜLLER, P. & WALKER, S. G. (2010). *Bayesian nonparametrics*, vol. 28. Cambridge University Press.

JOE, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London.

NELSEN, R. (2006). *An introduction to copulas*. Springer Verlag.

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

NOTT, D. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.

O'HARA, R. B. & SILLANPÄÄ, M. J. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis* **4**, 85–117.

PARK, T. & CASELLA, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

PITT, M., CHAN, D. & KOHN, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554.

SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis*, vol. 26. Chapman & Hall/CRC.

SKLAR, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229–231.

SMITH, M. & KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* , 267–288.

VILLANI, M., KOHN, R. & GIORDANI, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* **153**, 155–173.

ZEEVI, A. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks* .

Journal of Statistical Planning and Inference
2010, Vol. **140**, No. 12, 3638–3654
DOI: 10.1016/j.jspi.2010.04.031

FLEXIBLE MODELING OF CONDITIONAL DISTRIBUTIONS USING SMOOTH MIXTURES OF ASYMMETRIC STUDENT T DENSITIES

FENG LI, MATTIAS VILLANI, AND ROBERT KOHN

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

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

1. INTRODUCTION

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

Villani et al. (2009) propose the smooth adaptive Gaussian mixture (SAGM) model as flexible model for regression density estimation. Their model is a finite mixture of Gaussian densities with the mixing probabilities, the component means and component variances modeled as functions of the covariates x, with Bayesian variable selection in all three sets of covariates. See Frühwirth-Schnatter (2006) for a comprehensive introduction to mixture models.

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

Li (corresponding author): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: feng.li@stat.su.se Villani: Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden and Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: mattias.villani@riksbank.se Kohn: Australian School of Business, University of New South Wales, UNSW, Sydney 2052, Australia. E-mail: r.kohn@unsw.edu.au.

single complex component can often give a better and numerically more stable fit in substantially less computing time than a model with many simpler components. Moreover, simulations and real applications in Villani et al. (2009) show that a simple-and-many approach can fail to fit heteroscedastic data even with a very large number of components, especially in situations with more than one or two covariates. Having heteroscedastic components in the mixture is therefore crucial for accurately modeling heteroscedastic data.

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

2. THE MODEL AND PRIOR

2.1. **Smooth mixtures.** Our model is a finite mixture density with weights that are smooth functions of the covariates,

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

where $p_k(y|x)$ is the kth component density with weight $\omega_k(x)$. The component densities are asymmetric student t densities described in detail in the next section. The weights are modeled by a multinomial logit function

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

with $\gamma_1 = 0$ for identification. The covariates in the components can in general be different from the covariates in the mixture weights. Jiang (1999); Jiang & Tanner (1999) show that smooth mixtures with sufficiently many (generalized) linear regression components can approximate any density in the exponential family with arbitrary smooth mean functions. See also Zeevi (1997) for approximation of densities with mixture models.

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

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

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

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

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

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

where

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

is the kernel of a student t density with variance $\phi^2\nu/(\nu-2)$ *and* $c = 2[(1+\lambda)\phi\sqrt{\nu}Beta(\frac{\nu}{2},\frac{1}{2})]^{-1}$ *is the normalization constant.*

The location parameter μ is the mode, $\phi > 0$ is the scale parameter, and $\lambda > 0$ is the skewness parameter. When $\lambda < 1$ the distribution is skewed to the left, when $\lambda > 1$ it is skewed to the right, and when $\lambda = 1$ it reduces to the usual symmetric student-t density (Figure 1, left). The skewness of split-t can approach infinity as ν approaches 3 and when ν approaches infinity, the maximum skewness approaches 1 (Figure 1, right). The split-t distribution reduces to the two-piece normal distribution in Gibbons (1973) and John (1982) as $\nu \to \infty$. The split-t density has the advantage that its interpretation is simple since it is equal to the well-known symmetric student t density on either side of the mode, but any other asymmetric t density can equally well be used in our MCMC methodology, see Section 3.1.

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

$$S(y) = \frac{E\left[y - E(y)\right]^3}{V(y)^{3/2}}$$

$$K(y) = \frac{E\left[y - E(y)\right]^4}{V(y)^2} - 3,$$

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

Lemma 2. *If* $y \sim t(\mu, \phi, \lambda, \nu)$ then

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

where

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

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

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

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

where

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

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

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

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

but any smooth link function can equally well be used in the MCMC methodology. Additional flexibility can be obtained by letting a subset of the covariates be a non-linear basis expansions, e.g. additive splines or splines surfaces (Ruppert et al., 2003) as in Villani et al. (2009), but this is not pursued here. A strength of our approach is that the four regression coefficient vectors: β_μ, β_ϕ, β_ν and β_λ are all treated in a unified way in the MCMC algorithm. Whenever we refer to a regression coefficient vector without subscript, β, the argument applies to any of the regression coefficient vector of the split-t parameters in (3).

This split-t model will often be flexible enough to fit the data, but there are datasets that require a smooth mixture model, for example when the data are multimodal for some covariates values. A second example occurs when the wrong link function is used in one of the split-t parameters,

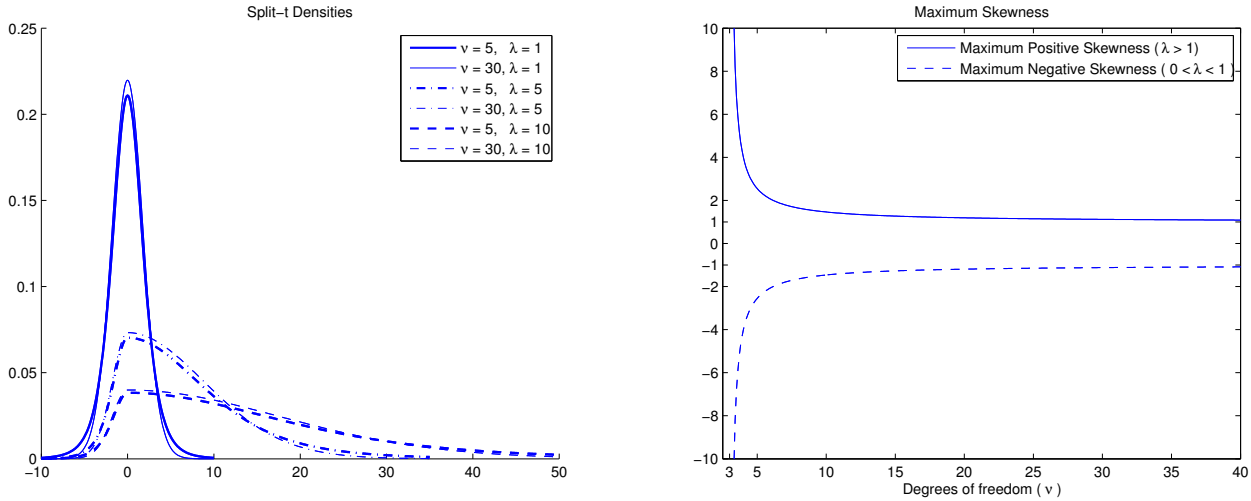

FIGURE 1. Graphical display of the split-t densities. The left-hand side are the split-t densities with location parameter $\mu = 0$ and skewness parameter $\lambda = 1.8$. The right-hand side is the maximum skewness of the split-t as a function of degrees of freedom.

where the mixture can then correct for this erroneous choice. A third example is when there are outliers in the data that cannot be accommodated by a student t density.

A smooth mixture of split-t densities is a model with a large number of parameters, however, and is therefore likely to over-fit the data unless model complexity is controlled effectively. We use Bayesian variable selection in all four split-t parameters, and in the mixing function. This can lead to important simplifications of the split-t components. Not only does this control complexity for a given number of components, but it also simplifies the existing components if an additional component is added to the model (the LIDAR example in Villani & Kohn (2007) illustrates this well). Increasing the number of components can therefore in principle even reduce the number of effective parameters in the model.

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

2.3. **The prior.** Although the MCMC methodology (see Section 3.2) allows any prior distribution, we shall now present an easily specified prior that depends only on a few hyper-parameters. First, we standardize the covariates by subtracting the mean and dividing by the standard deviation. This allows us to assume prior independence between the intercept and the remaining regression coefficients, and the intercepts have the interpretation of being the (possibly transformed) split-t parameters at the mean of the original covariates. Since there can be a large number of covariates in the model, our strategy is to incorporate available prior information via the intercepts, and to treat the remaining regression coefficients more informally. Assuming a normal prior for μ

implies a normal prior on $\beta_{\mu 0}$. The other three split-t parameters ϕ, λ and ν are assumed to follow independent log-normal priors with means m^* and s^*, where m^* and s^* are different for the different split-t parameters. This translates into a normal prior on the intercept with mean

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

and variance

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

The regression coefficients β_μ, β_ϕ, β_ν and β_λ are assumed to be independent a priori. We allow for Bayesian variable selection by augmenting each parameter vector β by a vector of binary covariate selection indicators $\mathscr{I} = (i_1, ..., i_p)$ such that $\beta_j = 0$ if $i_j = 0$. Let $\beta_{\mathscr{I}}$ denote the subset of β selected by $\mathscr{I}$. We assume the following prior for each β vector

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

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

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

The prior on the mixing function decomposes as

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

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

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

and $\gamma_{\mathscr{Z}^c} = 0$ with probability one. We use $\tau_\gamma^2 = 10$ as default value. Finally, $p(s|\gamma, \mathscr{Z})$ is given by the multinomial logit model in (2). To reduce the number of parameters and to speed up the MCMC algorithm we restrict the columns of $\mathscr{Z}$ to be identical, i.e. make the assumption that a

covariate is either present in the mixing function in all components, or does not appear at all, but the extension to general $\mathscr{X}$ is straightforward, see Villani et al. (2009).

3. INFERENCE METHODOLOGY

3.1. **The general MCMC scheme.** We use MCMC methods to sample from the joint posterior distribution, and draw the parameters and variable selection indicators in blocks. Villani et al. (2009) experimented with several different algorithms in a related setting and the algorithm outlined below is similar to their preferred algorithm. The details of the algorithm are given in Appendix A. The method used to select the number of components is discussed in Section 3.3.

The algorithm is a Metropolis-within-Gibbs sampler that draws parameters using the following six blocks:

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

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

Mixture models have well-known identification problems, the most serious one being the so-called label switching problem, which means that the likelihood is invariant with respect to permutations of the components in the mixture, see e.g. Celeux et al. (2000), Jasra et al. (2005) and Frühwirth-Schnatter (2006). The aim of our article is to estimate the predictive density, so that label switching is neither a numerical nor conceptual problem (Geweke, 2007). If an interpretation of the mixture components is required, then it is necessary to impose some identification restrictions on some of the model parameters, e.g. an ordering constraint (Jasra et al., 2005).

The number of components is assumed known in our MCMC scheme. A Bayesian analysis via mixture models with an unknown number of components is possible using, e.g. Dirichlet process mixtures (Escobar & West, 1995), reversible jump MCMC (Richardson & Green, 1997) and birth-and-death MCMC (Stephens, 2000). However, one major drawback is that the posterior distribution of the number of components for a given data set typically depends heavily on the priors. In order to avoid that, we instead compare and select models based on the out-of-sample LPDS (see details in Section 3.3). Our *complex-and-few* approach is also helpful in this aspect as it keeps the number of components to a minimum (see Section 4).

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

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

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

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

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

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

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

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

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

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

The proposal density at the current point $g_1(\beta_c|\mathscr{I}_c, \beta_p)$ is a multivariate t-density with mode $\widetilde{\beta}$ and covariance matrix equal to the negative inverse Hessian evaluated at $\widetilde{\beta}$, where $\widetilde{\beta}$ is the point obtained by iterating R steps with the Newton algorithm, this time starting from β_p. A simple way to propose $\mathscr{I}_p$ is to randomly select a small subset of $\mathscr{I}_c$ and then always propose a change of the selected indicators. This proposal can be refined in many ways, using, e.g. the adaptive scheme in Nott & Kohn (2005), where the history of $\mathscr{I}$-draws is used to adaptively build up a proposal for each indicator. It is important to note that β_c and β_p may now be of different dimensions, so the original Newton iterations no longer apply. We will instead generate β_p using the following generalization of Newton's method. The idea is that when the parameter vector β

changes dimensions, the dimension of the functionals $\varphi_c = x'\beta_c$ and $\varphi_p = x'\beta_p$ stay the same, and the two functionals are expected to be quite close. A generalized Newton update is

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

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

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

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

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

X_r is the matrix with the covariates that have non-zero coefficients in β_r, and all expressions are evaluated at $\beta = \beta_r$. For the prior gradient this means that $\partial \ln p(\beta)/\partial \beta$ is evaluated at β_r, including all zero parameters, and that the sub-vector conformable with β_{r+1} is extracted from the result. The same applies to the prior Hessian (which does not depend on β however, if the prior is Gaussian). Note that we only need to compute the scalar derivatives $\partial \ln p(y_i|\phi_i)/\partial \phi_i$ and $\partial^2 \ln p(y_i|\phi_i)/\partial \phi_i^2$.

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

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

3.3. **Model comparison.** The key quantity in Bayesian model comparison is the marginal likelihood. The marginal likelihood is sensitive to the choice of prior, however, and this is especially true when the prior is not very informative, see e.g. Kass (1993) for a general discussion and

Richardson & Green (1997) in the context of density estimation. By sacrificing a subset of the observations to update/train the vague prior we remove much of the dependence on the prior, and obtain a better assessment of the predictive performance that can be expected for future observations. To deal with the arbitrary choice of which observations to use for estimation and model evaluation, one can use B-fold cross-validation of the log predictive density score (LPDS):

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

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

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

where $\mathscr{T}_b$ is the index set for the observations in $\tilde{y}_b$, and the LPDS is easily computed by averaging $\prod_{i\in\mathscr{T}_b}p(y_i|\theta,x_i)$ over the posterior draws from $p(\theta|\tilde{y}_{-b})$. This requires sampling from each of the B posteriors $p(\theta|\tilde{y}_{-b})$ for $b=1,...,B$, but these MCMC runs can all be run in isolation from each other and are therefore ideal for parallel computing on widely available multi-core processors.

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

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

with each term in the decomposition

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

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

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

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

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

models between 3 and 5 is considered strong evidence in favor of one model, and a difference of more than five LPDS points is very strong evidence.

4. MODELING THE DISTRIBUTION OF DAILY STOCK MARKET RETURNS

4.1. **S&P500 data and priors.** Modeling the volatility/variability in financial data has been an highly active research area since the seminal paper by Engle (1982) introduced the ARCH model (see, e.g. Baillie (2006) for a survey of the field), and there are large financial markets for volatility-based instruments. Financial data, such as stock market returns, are typically heavy tailed and subject to volatility clustering, i.e. a time-varying variance that evolves in a very persistent fashion. We here model the entire distribution of daily returns from the S&P500 stock market index, $p(y_t|x_t)$, where $y_t = 100\ln(p_t/p_{t-1})$ is the daily return at time t, p_t is the closing S&P500 index on day t, and x_t contains the covariate observations at time t. By focusing on the whole distribution of returns we are able to compute, e.g. the posterior distribution of the *Value-at-Risk* (VaR), i.e. the 1% quantile of the return distribution, which is of fundamental interest to financial analysts, see Villani et al. (2009) for an example based on the S&P500 datasets.

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

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

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

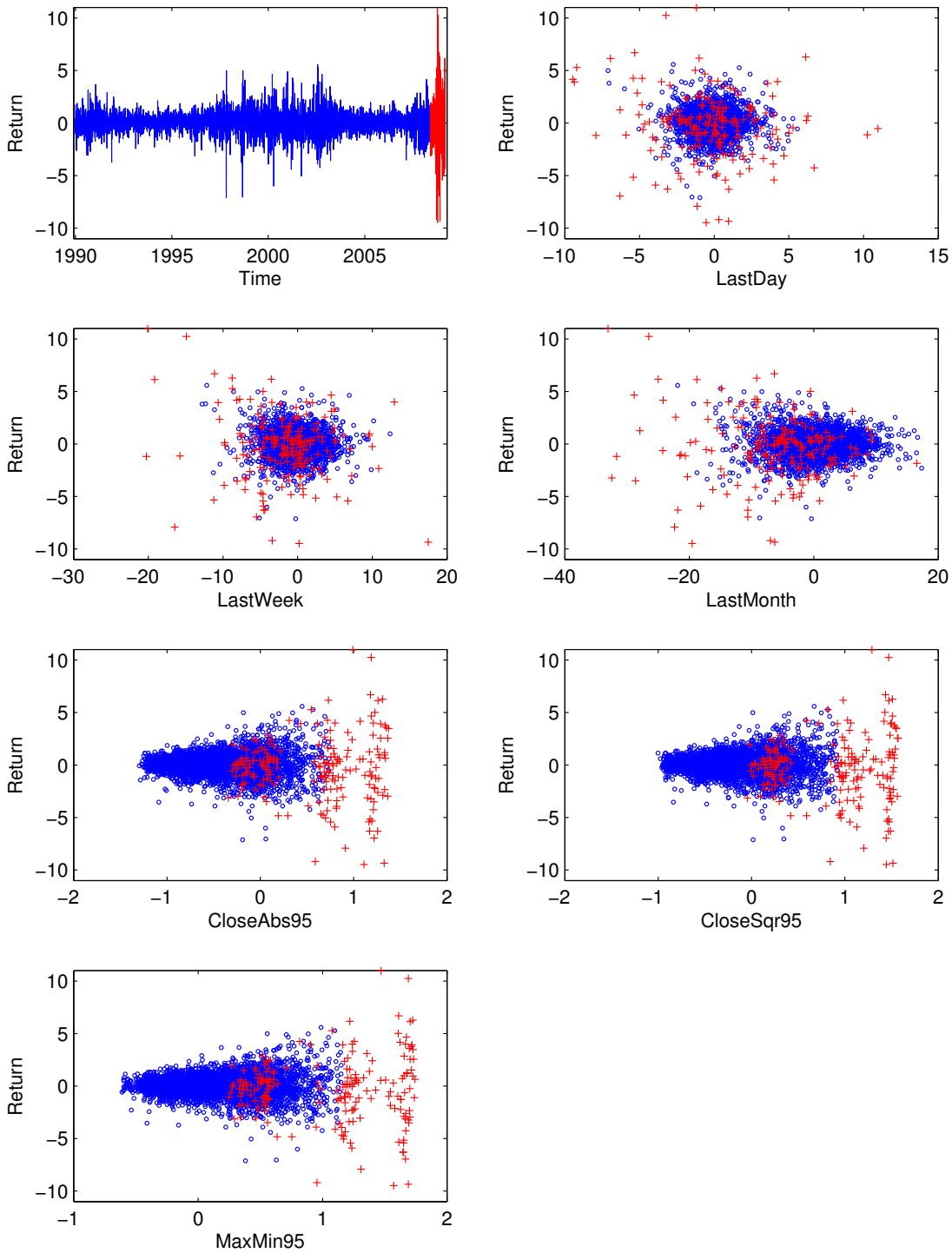

FIGURE 2. Graphical display of the S&P500 data from January 1, 1990 to May 29, 2008 (blue lines and circles) and May 30, 2008 to March 13, 2009 (red lines and crosses). The subgraph in the upper left position is a time series plot of Return, the other subgraphs are scatter plots of Return against a covariate.

and CloseAbs95, a geometrically decaying average of past absolute returns

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

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

Villani et al. (2009) demonstrate that the SAGM model with its heteroscedastic components outperforms the SMR in Geweke & Keane (2007). Villani et al. (2009) also introduce seven additional covariates and show that they substantially improve the out-of-sample performance of the SAGM. We will concentrate on this nine-variable model. The seven additional covariates are: LastWeek and LastMonth, a moving average of the returns from the previous five and 20 trading days, respectively. The variable CloseAbs80, the same variable as CloseAbs95 but with $\rho = 0.80$, is also added to the covariate set, and so is the square root of $(1-\rho)\sum_{s=0}^{\infty}\rho^s y_{t-2-s}^2$, for $\rho = 0.80$ and 0.95 (CloseSqr80 and CloseSqr95). Finally, Villani et al. (2009) include a measure of volatility that has been popular in the finance literature: $(1-\rho)\sum_{s=0}^{\infty}\rho^s(\ln p_{t-1-s}^{(h)} - \ln p_{t-1-s}^{(l)})$, where $p_t^{(h)}$ and $p_t^{(l)}$ are the highest and lowest values of the S&P500 index at day t. This measure has been shown both theoretically and empirically to carry more information on the volatility than changes in closing quotes (Alizadeh et al., 2002). We consider both $\rho = 0.8$ (MaxMin80) and $\rho = 0.95$ (MaxMin95). As in Villani et al. (2009), all variables except LastDay, LastWeek and LastMonth enter the model in logarithmic form.

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

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

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

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

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

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

components when K is large. This reinforces the conclusion stressed in Villani et al. (2009) that having heteroscedastic components is crucial for modeling heteroscedastic data.

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

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

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

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

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

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

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

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

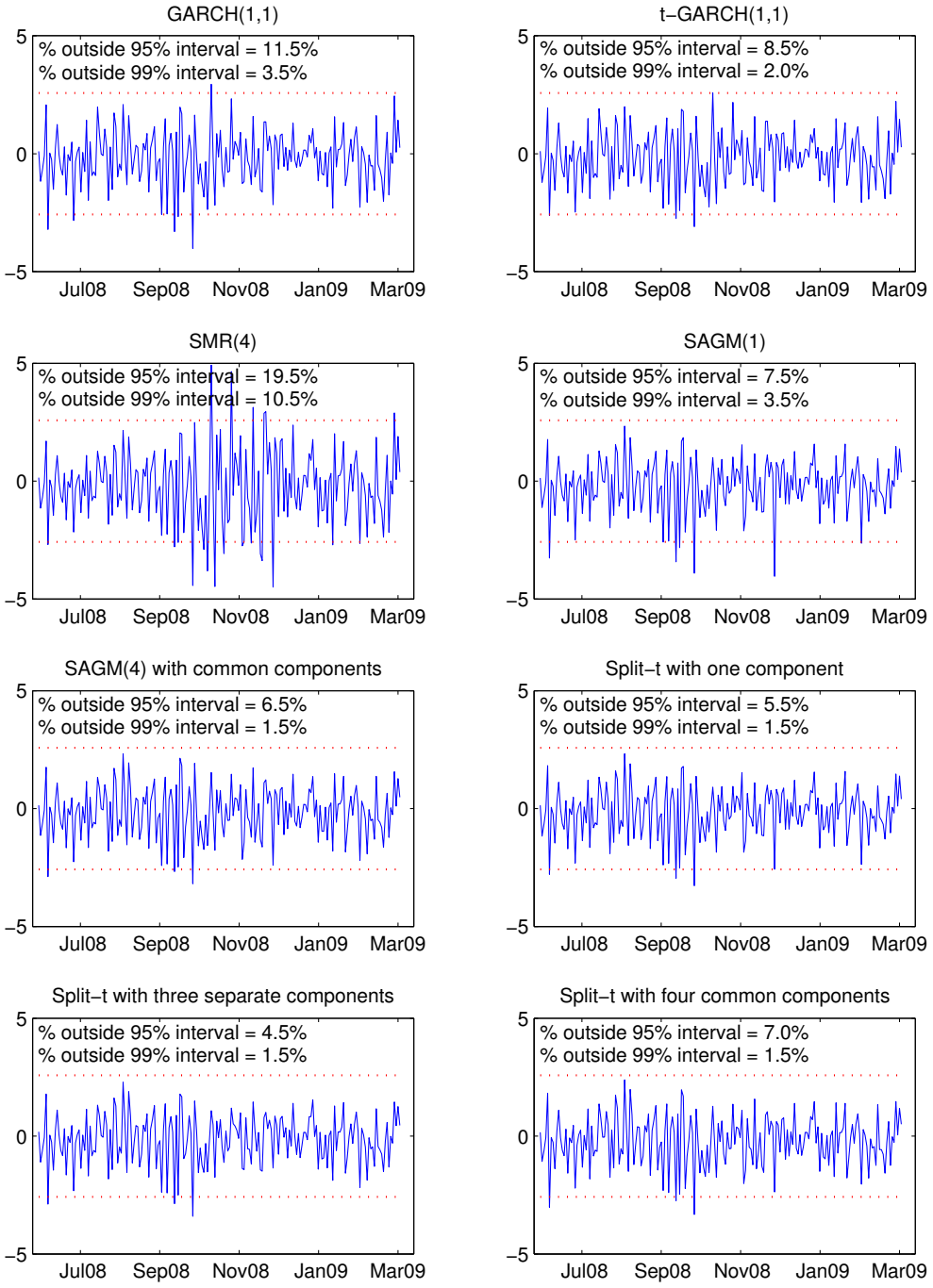

FIGURE 3. Plot of the 199 normalized residuals in the evaluation sample over time (solid lines). The dotted lines are the 99% probability intervals under the $N(0,1)$ reference distribution. Each sub-graph displays the percentage of normalized residuals outside the 95% and 99% probability intervals of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

TABLE 4. Posterior summary of the one-component split-t model[✠].

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

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

are outside the 99% reference interval. The remaining four models in Figure 3 have rather similar seemingly homoscedastic and independent residuals, and they all have close to the right coverage. The one-component split-t model is doing remarkably well during this very difficult time period.

We now take a more detailed look at the inferences from the one-component split-t model. Table 4 presents summaries of the posterior distribution. The results from the variable selection among the covariates in the scale parameter is very similar to the results for the variance function in Villani et al. (2009): the covariates MaxMin95, LastWeek and LastMonth have a posterior inclusion probability close to one, and all other covariates are essentially excluded. There is support for some small skewness in the model, but no covariates enter λ. The degrees of freedom at the posterior mean is $\exp(2.482) = 11.96$, (assuming all other covariates at their mean) which is not very heavy tailed, but LastWeek enters the model with probability 0.638 and with a large negative coefficient, so the degrees of freedom is very small for the largest values of LastWeek (recall that LastWeek$\in [-1,1]$). The last column of Table 4 gives the inefficiency factor (IF) for all parameters with inclusion probabilities larger than 0.02. It is clear that the MCMC algorithm is very efficient, almost all parameters have IFs smaller than 10. The MH acceptance probabilities for the regression coefficients in μ, ϕ, ν and λ are as high as 95%, 81%, 75% and 94%, respectively.

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

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

Figure 4 presents box plots of the posterior distribution of the number of included parameters, i.e. $p(\sum_{k=1}^{K}(\sum_{q=1}^{Q}\sum_{p=1}^{P}\mathscr{I}_{kqp} + \sum\mathscr{I}_{mix}))$, where $\mathscr{I}_{kqp}$ is the Bayesian variable selection indicator for the pth variable in the qth parameter in the kth component density and $\sum\mathscr{I}_{mix}$ is the sum of variable selection indicators in the mixing function. Figure 4 shows that the SMR(4) has 26 effective parameters on average, while SAGM(1), which performs better than any SMR model, has only five effective parameters on average. Moreover, the one-component split-t model contains only four more effective parameters than SAGM(1), but the split-t model has much high LPDS.

Figure 4 (bottom right) also shows that the proportion of posterior included parameters to potential parameters is close to 0.5 in the SAGM and split-t models with a large number of components. This result is in part a reflection of our choice of a Bernoulli(0.5) prior for the variable selection indicators. This prior implies that the prior on the number of effective parameters is a binomial distribution with mean $N/2$ and standard deviation $\sqrt{N/4}$, where N is the number of potential parameters in the model. For models with large N the prior is therefore fairly tightly centered on a large number of effective parameters. Other priors on the variable selection indicators are straightforward to implement, however, e.g. the uniform prior in Denison et al. (2002) or the hierarchical prior in Kohn et al. (2001).

To investigate the stability of the predictive densities for different sets of sample sizes we estimate the one-component split-t model using five samples with an increasing number of observations. The samples consist of the first 1000, 2000, 3000, 4000 trading days and then finally using the full sample between Jan 1, 1990 and Mar 13, 2009. Figure 5 displays the conditional predictive densities for the three sets of covariates values present on the 4648th, 4725th, and 4753th trading day where MaxMin95 is 0.2503, 0.9043, and 1.737, respectively (hence representing states of low, medium and high volatility). Figure 5 shows that the 1% quantiles (*VaR*) of the return distribution do not change significantly over five estimation samples.

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

5. Conclusions

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

Acknowledgments

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

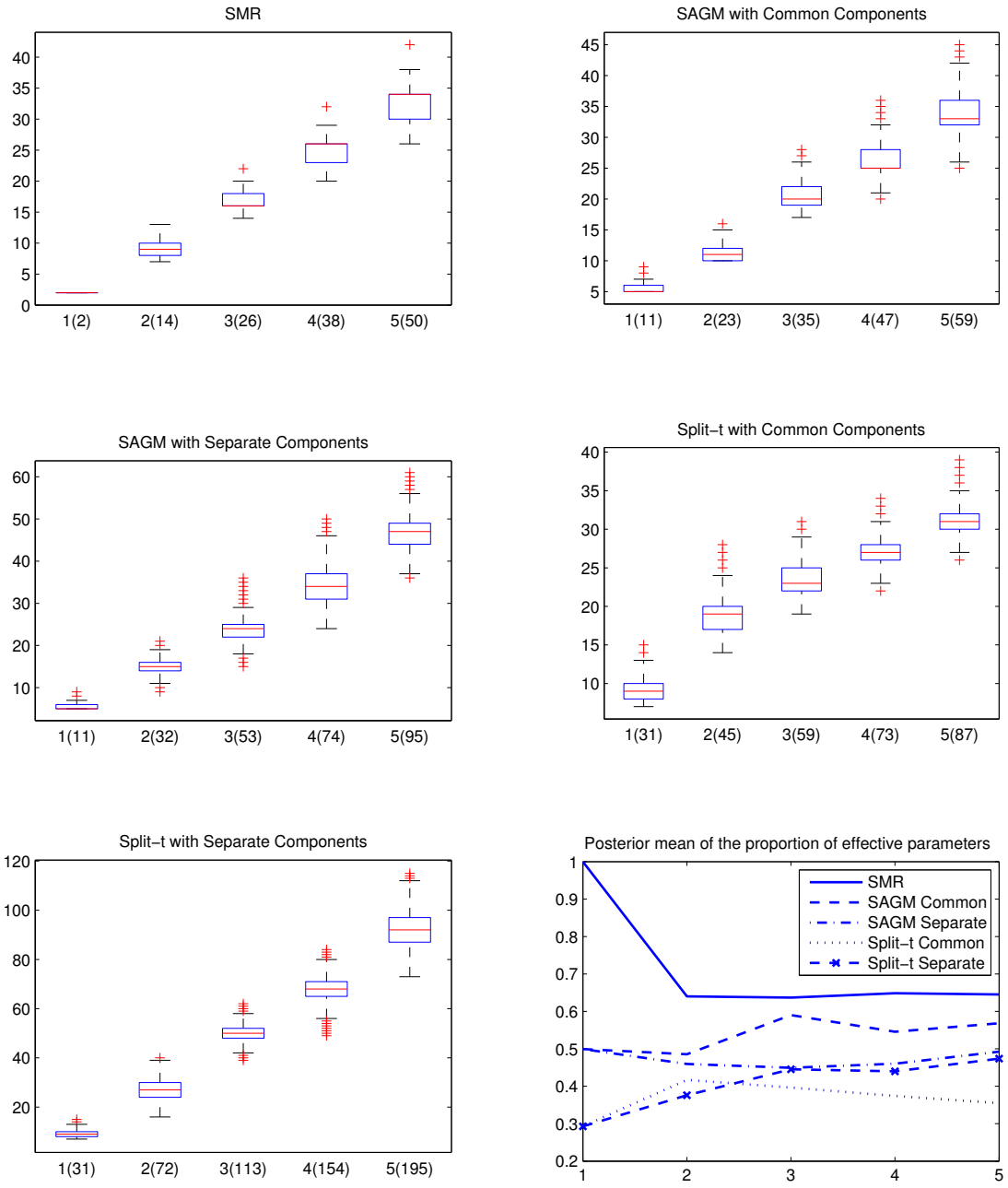

FIGURE 4. The posterior distribution of the number of included parameters. On first five subplots, the horizontal axis measures the number of components (with the number of potential parameters in parentheses) and the vertical axis is the total number of effective parameters in the model. All models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

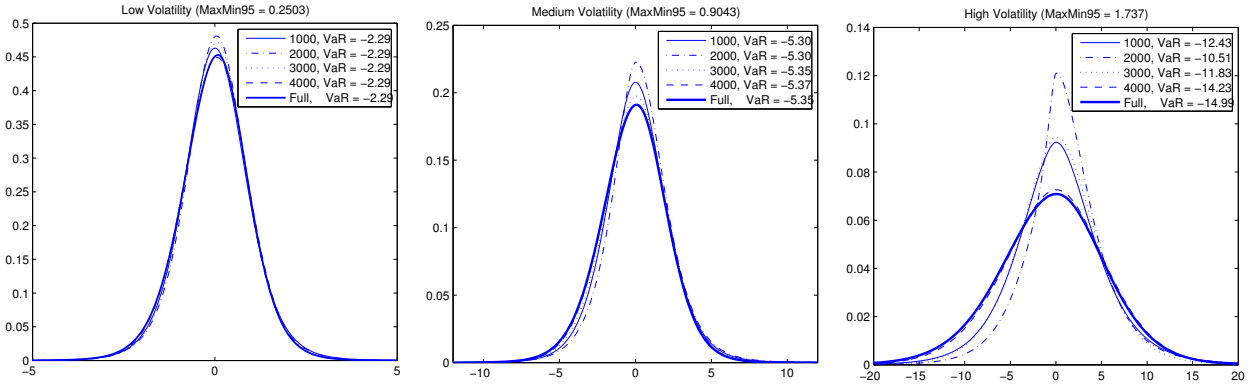

FIGURE 5. Investigating estimation stability over different subsamples. The subgraphs show predictive densities for different sets values on the covariates (low, medium and high volatility). The model is estimated on the first 1000, 2000, 3000, 4000 trading days starting from Jan 1, 1990 and the full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

APPENDIX A. MCMC IMPLEMENTATION

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

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

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

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

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

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

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

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

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

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

Gradient and Hessian wrt μ

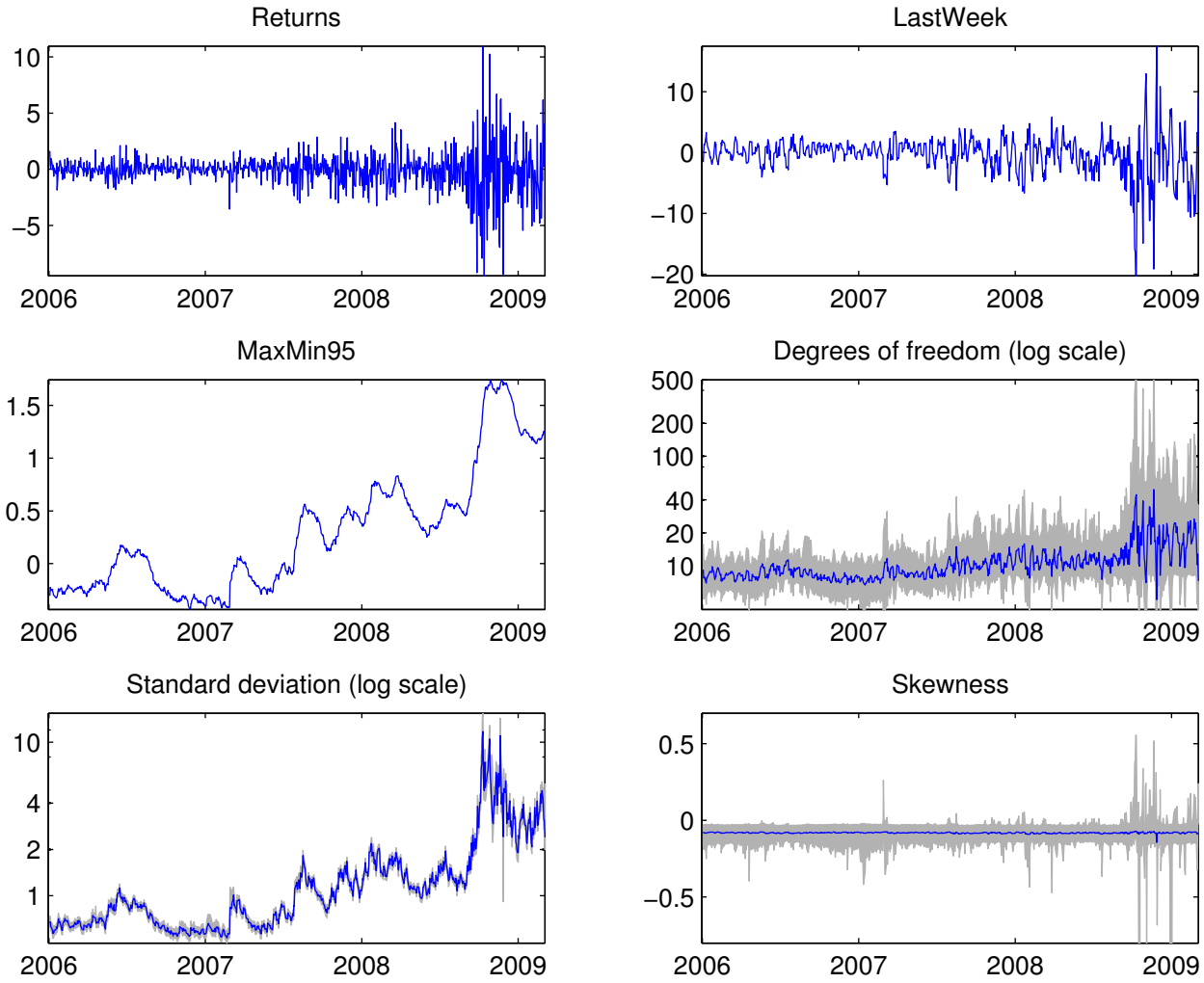

FIGURE 6. Time series plot of the posterior median and 95% probability intervals for some moments of the return distribution. The time series of returns and two of the key covariates are also plotted. The posterior distribution is based on the full sample up to March 13, 2009. The distribution of the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

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

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

Gradient and Hessian wrt ϕ

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

Gradient and Hessian wrt ν

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

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

Gradient and Hessian wrt λ

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

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

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

REFERENCES

ABRAMOWITZ, M. & STEGUN, I., eds. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical table*. Courier Dover Publications, New York.

ALIZADEH, S., BRANDT, M. W. & DIEBOLD, F. X. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance* **57**, 1047–1091.

BAILLIE, R. T. (2006). *Handbook of Econometrics*, vol. 1, chap. Modeling volatility. Palgrave Macmillan, New York, pp. 737–764.

BERKOWITZ, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business and Economic Statistics* **19**, 465–474.

BOLLERSLEV, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics* **69**, 542–547.

CELEUX, G., HURN, M. & ROBERT, C. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association* **95**, 957.

DENISON, D., HOLMES, C. C., MALLICK, B. K. & SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Jone Wiley & Sons, Chichester.

DIEBOLT, J. & ROBERT, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 363–375.

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

ESCOBAR, M. & WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the american statistical association* **90**.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite mixture and Markov switching models*. Springer Verlag.

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

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society* **54**, 1317–1339.

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

GEWEKE, J. & KEANE, M. (2007). Smoothly mixing regressions. *Journal of Econometrics* **138**, 252–290.

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

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

JACOBS, R., JORDAN, M., NOWLAN, S. & HINTON, G. (1991). Adaptive mixtures of local experts. *Neural computation* **3**, 79–87.

JASRA, A., HOLMES, C. C. & STEPHENS, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20**, 50–67.

JEFFREYS, H. (1961). *Theory of probability, 3rd ed*. Oxford, New York.

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

JIANG, W. & TANNER, M. A. (1999). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation* **11**, 1183–98.

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

JORDAN, M. & JACOBS, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation* **6**, 181–214.

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

KASS, R. & RAFTERY, A. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

KOHN, R., SMITH, M. & CHAN, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322.

NOTT, D. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.

NOTT, D. J. & LEONTE, D. (2004). Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics* **13**, 362–382.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **59**, 731–792.

RUPPERT, D., WAND, M. & CARROLL, R. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *The Annals of Statistics* **28**, 40–74.

VILLANI, M. & KOHN, R. (2007). Nonparametric regression density estimation using smoothly varying normal mixtures. *Sveriges Riksbank Research* **211**.

VILLANI, M., KOHN, R. & GIORDANI, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* **153**, 155–173.

WOOD, S., JIANG, W. & TANNER, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513.

ZEEVI, A. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks* .

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

Book chapter in *Mixtures: Estimation and Applications*
Edited by Kerrie L. Mengersen, Christian P. Robert and D. Michael Titterington.
First edition (2011), pp. 123–144
DOI: 10.1002/9781119995678.ch6

MODELING CONDITIONAL DENSITIES USING FINITE SMOOTH MIXTURES

FENG LI, MATTIAS VILLANI, AND ROBERT KOHN

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

KEYWORDS: Bayesian inference, Markov chain Monte Carlo, Mixture of Experts, Variable selection.

1. INTRODUCTION

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

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

Li: Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. Villani: Research Division, Sveriges Riksbank and Department of Statistics, Stockholm University. Kohn: Australian School of Business, University of New South Wales, UNSW, Sydney 2052, Australia.

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

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

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

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

selection in a mixture context is illustrated in Section 4.2 on the LIDAR data. Reducing the number of effective parameters by variable selection also helps the MCMC algorithm to converge faster and mix better.

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

2. THE MODEL AND PRIOR

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

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

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

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

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

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

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

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

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

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

where

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

Note that each mixture components has its own set of parameters. We will suppress the component subscript k in the remainder of this section, but, unless stated otherwise, all parameters are component-specific. SAGM uses a linear link function for the mean and log link for the variance, but any smooth link function can equally well be used in our MCMC methodology. Additional

flexibility can be obtained by letting a subset of the covariates be a non-linear basis expansions, e.g. additive splines or splines surfaces (Ruppert et al., 2003) as in Villani et al. (2009); see also the LIDAR example in Section 4.2.

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

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

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

where

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

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

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

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

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

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

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

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

where

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

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

Similarly, the log-normal components are of the form

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

where

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

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

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

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

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

We standardize the covariates to have zero mean and unit variance, and assume prior independence between the intercept and the remaining regression coefficients. The intercepts then have the interpretation of being the (possibly transformed) density parameters at the mean of the original covariates. The strategy in Villani et al. (2010) is to specify priors directly on the parameters of the mixture component, e.g. the degrees of freedom ν in the split-t components, and then back out the implied on the intercept β_{ν_0}. For example, a normal prior for a parameter with identity link (e.g. μ in the split-t model) trivially implies a normal prior on $\beta_{\mu 0}$; a log-normal prior with mean m^* and variance s^{*2} for a parameter with log link (e.g. ϕ in the split-t model) implies a normal prior $N(m_0, s_0^2)$ for β_{ϕ_0} where

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

The regression coefficients vectors are assumed to be independent a priori. We allow for Bayesian variable selection by augmenting each parameter vector β by a vector of binary covariate selection indicators $\mathscr{I} = (i_1, ..., i_p)$ such that $\beta_j = 0$ if $i_j = 0$. Let $\beta_{\mathscr{I}}$ denote the subset of β selected by $\mathscr{I}$. In a Gaussian linear regression one can use a g-prior (Zellner, 1986) $\beta \sim N[0, \tau_\beta^2 (X'X)^{-1}]$ on the full β and then condition on the restrictions imposed by $\mathscr{I}$. Setting

$\tau^2 = n$, where n is the number of observations, gives the unit-information prior, i.e. a prior that carries information equivalent to a single observation from the model. More generally, the unit information prior is $\beta \sim N[0, \tau_\beta^2 \mathscr{I}^{-1}(\beta)]$ where

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

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

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

The prior on the mixing function decomposes as

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

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

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

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

3. Inference methodology

3.1. **The general MCMC scheme.** We use MCMC methods to sample from the joint posterior distribution, and draw the parameters and variable selection indicators in blocks. The algorithm below is the preferred algorithm from the experiments in Villani et al. (2009). The number of components is determined by a Bayesian version of cross-validation discussed in Section 3.3.

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

(1) $\{(\beta_\mu^{(k)}, \mathscr{I}_\mu^{(k)})\}_{k=1,...,K}$
(2) $\{(\beta_\phi^{(k)}, \mathscr{I}_\phi^{(k)})\}_{k=1,...,K}$
(3) $\{(\beta_\lambda^{(k)}, \mathscr{I}_\lambda^{(k)})\}_{k=1,...,K}$
(4) $\{(\beta_\nu^{(k)}, \mathscr{I}_\nu^{(k)})\}_{k=1,...,K}$

(5) $s = (s_1, ..., s_n)$ (6) γ and $\mathscr{I}_Z$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

X_r is the matrix with the covariates that have non-zero coefficients in β_r, and all expressions are evaluated at $\beta = \beta_r$. For the prior gradient this means that $\partial \ln p(\beta)/\partial \beta$ is evaluated at β_r, including all zero parameters, and that the sub-vector conformable with β_{r+1} is extracted from the result. The same applies to the prior Hessian (which does not depend on β however, if the prior is Gaussian). Note that we only need to compute the scalar derivatives $\partial \ln p(y_i|\phi_i)/\partial \phi_i$ and $\partial^2 \ln p(y_i|\phi_i)/\partial \phi_i^2$.

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

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

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

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

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

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

4. APPLICATIONS

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

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

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

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

TABLE 1. Priors in the simulation study

	μ	ϕ	ν	λ
Mean	0	1	10	1
Std	10	1	7	0.8

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

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

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

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

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

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

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

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

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

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

4.2. **LIDAR data .** Our first real data set comes from a technique that uses laser-emitted light to detect chemical compounds in the atmosphere (LIDAR, LIght Detection And Ranging). The response variable (logratio) consists of 221 observations on the log ratio of recieved light from two laser sources: one at the resonance frequency of the target compound, and the other from a frequency off this target frequency. The predictor is the distance travelled before the light is reflected back to its source (range). The original data comes from Holst et al. (1996) and has been analyzed by for example Ruppert et al. (2003) and Leslie et al. (2007). Our aim is to model the predictive density $p(\mathsf{logratio} \mid \mathsf{range})$.

Leslie et al. (2007) show that a Gaussian model with nonparametric mean and variance can capture this data set quite well. We will initially use the SAGM model in Villani et al. (2009) with the mean, variance and mixing functions all modelled nonparametrically by thin plate splines (Green & Silverman, 1994). Ten equidistant knots in each component are used for each of these three aspects of the model. We use a version of SAGM where the variance functions of the components are proportional to each other, i.e. only the intercepts in the variance functions are allowed to be different across components. The more general model with completely separate variance functions gives essentially the same LPDS, and the posterior distributions of the component variance functions (identified by order-restrictions) are largely over-lapping. We use the variable selection prior in Villani et al. (2009) where the variable selection indicator for a knot κ in the kth

mixture component is distributed as $Bernoulli\left[\pi_\beta \cdot \omega_k(\kappa)\right]$. This has the desirable effect of down-weighting knots in regions where the corresponding mixture component has small probability. We compare our results to the smoothly mixing regression (SMR) in Geweke & Keane (2007) which is a special case of SAGM where the components' variance functions are independent of the covariates and any heteroscedasticity is generated solely by the mixture. We use a prior with $m^* = 0$ and $s^{*2} = 10$ in the mean function, and $m^* = 1$ and $s^{*2} = 1$ in the variance function (see Section 2.3). Given the scale of the data, these priors are fairly non-informative. As documented in Villani et al. (2009) and Li et al. (2010), the estimated conditional density and the LPDS are robust to variations in the prior.

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

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

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

of range, a sizeable skewness which is also visible in Figure 1. The LPDS of the one-component split-t model is 64.014, which is only slightly worse than SAGM(1).

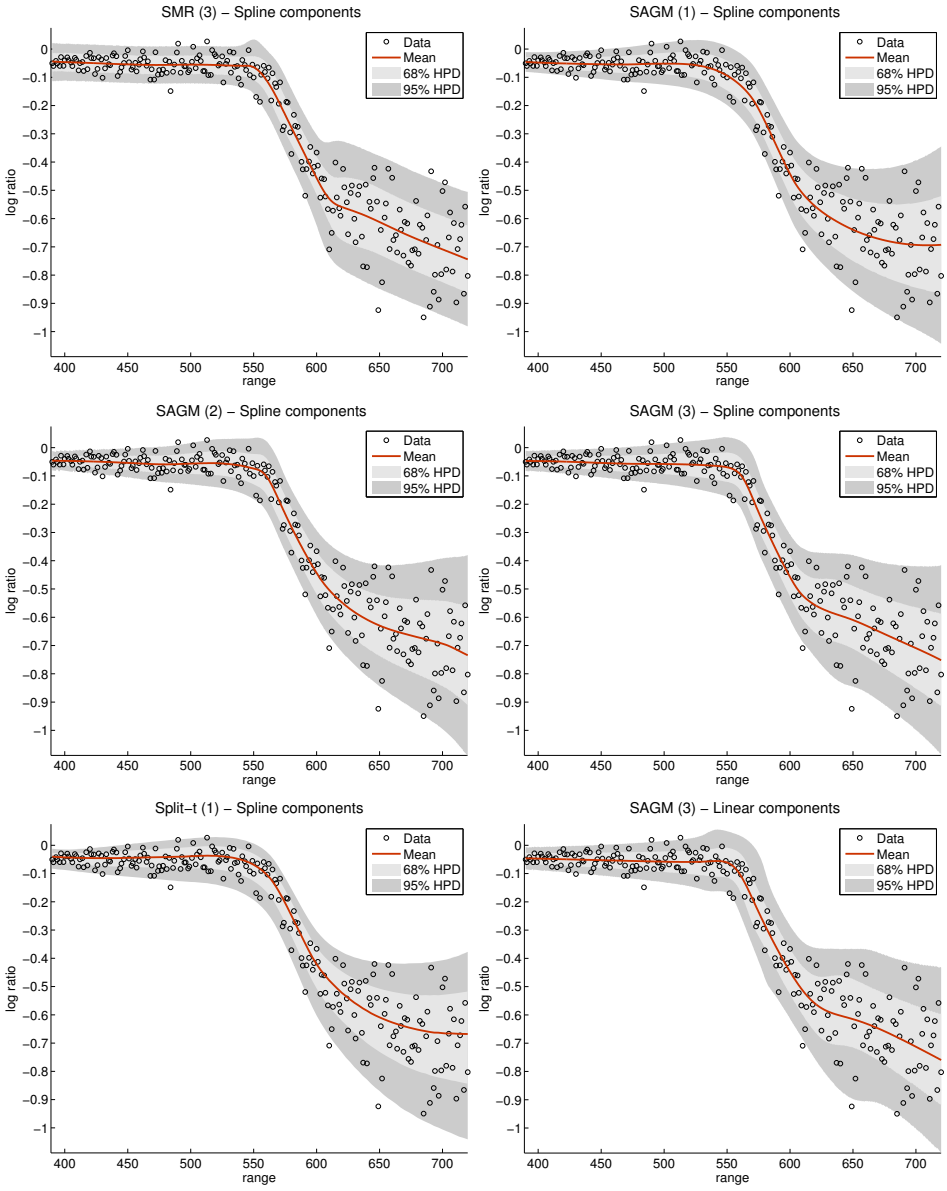

FIGURE 1. Assessing the in-sample fit of the smooth mixture models for the LIDAR data. The figure displays the actual data overlayed on HPD predictive regions. The solid line is the predictive mean.

4.3. **Electricity expenditure data.** Our second example uses a data set with electricity expenditures in 1602 households from South Australia (Bartels et al., 1996). Leslie et al. (2007) analyze

this data set and conclude that a heteroscedastic regression with errors following a Dirichlet process mixture fits the data well. They also document that the response variable is quite skewed. We consider both in-sample and out-of-sample performance of smooth mixture models, using the data set in Leslie et al. (2007) without interactions. The thirteen covariates used in our application are defined in Table 4

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

TABLE 4. The electricity bills regressors (subsets)

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

The prior means of μ and ϕ are set equal to the median and the standard deviation of the response variable, respectively. This data snooping is innocent as we set the standard deviation of μ and ϕ to 100, so the prior is largely non-informative. The prior mean and standard deviation of the skewness parameter, λ are both set to unity. This means that we are centering the prior on a symmetric model, but allowing for substantial skewness a priori. The prior mean of the degrees of freedom is set to 10 with a standard deviation of 7, which is wide enough to include both the Cauchy and essentially the Gaussian distributions. Since the data sample is fairly large, and we base model choice on the LPDS, the results are insensitive to the exact choice of priors.

We first explore the out-of-sample performance of several smooth mixture models using five-fold cross-validation of the LPDS. The five subsamples are chosen by sampling systematically from the data set. Table 5 displays the results for a handful of models. Every model is estimated both under the assumption of separate parameters and when all parameters except the intercepts are common across components; see Section 2.2.

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

The best model overall is the split-t model with three separate components, closely followed by the gamma model also with two separate components. It seems that this particular data set has

TABLE 5. Log predictive density score (LPDS) from fivefold cross-validation of the electricity expenditure data.

Model		$K=1$	$K=2$	$K=3$
SMR	*separate*	$-2,086$	$-2,027$	$-2,020$
	common	–	$-2,030$	$-2,020$
SAGM	*separate*	$-2,040$	$-2,022$	$-2,024$
	common	–	$-2,022$	$-2,017$
Split-normal	*separate*	$-2,014$	$-2,012$	$-2,015$
	common	–	$-2,064$	$-2,251$
Student's t	*separate*	$-2,025$	$-2,014$	$-2,014$
	common	–	$-2,029$	$-2,022$
Split-t	*separate*	$-2,034$	$-2,006$	$-1,996$
	common	–	$-2,073$	$-2,041$
Gamma	*separate*	$-2,007$	$-2,002$	$-2,003$
	common	–	$-2,008$	$-2,009$
Log-normal	*separate*	$-2,011$	$-2,006$	$-2,009$
	common	–	$-2,007$	$-2,010$

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

a combination of skewness and heavy-tailedness which is better modeled by a mixture than by a single skewed and heavy-tailed component.

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

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

Figure 3 depicts the conditional predictive densities $p(y|x)$ from three of the models: split-$t(1)$ (the most feature-rich one-component model), student-$t(2)$ (the best mixture of symmetric

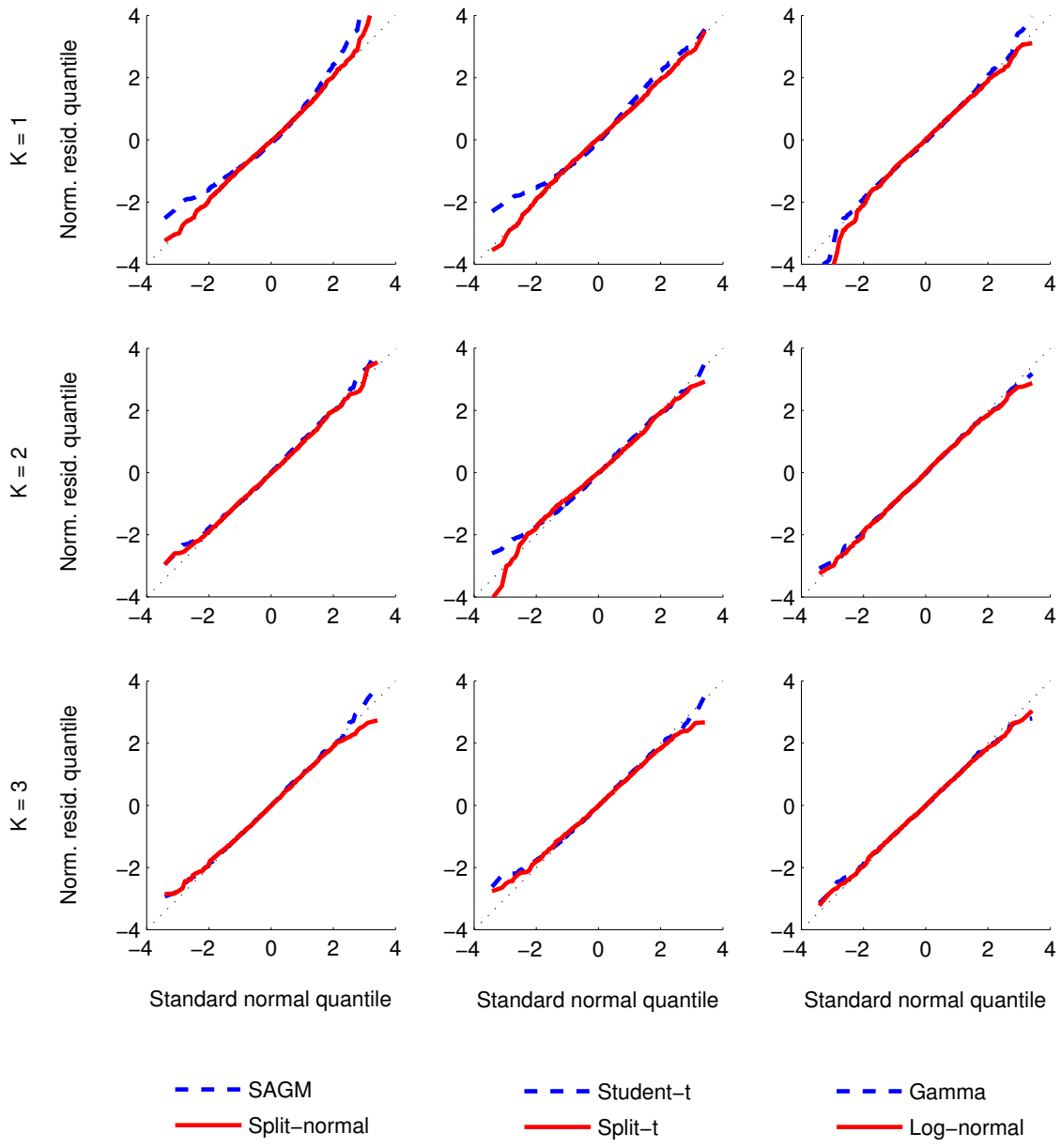

FIGURE 2. Quantiles plots of the normalized residuals resulting from SAGM and split-normal (first column); student's t and split-t (second column); gamma and log-normal (third column) with one to three separate components respectively. If the model is correct, the normalized residuals should be on the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

densities model with a minimal number of components) and Gamma(1) (the most efficient model with a minimum number of potential parameters). The predictive densities are displayed for three

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

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

different conditioning values of the most important covariates: log(rooms), log(income), sheonly and whtgel. All other covariates except the one indicated below the horizontal axis are fixed at their sample means. It is clear from Figure 3 that the predictive densities are very skewed, but also that the different models tend to produce very different types of skewness. The predictive densities from the 2-component student-t model are unimodal except for median and high values of whtgel where the two components are clearly visible.

5. CONCLUSIONS

We have presented a general model class for estimating the distribution of a continuous variable conditional on a set of covariates. The models are finite smooth mixtures of component densities where the mixture weights and all component parameters are functions of covariates. The inference methodology is a fully unified Bayesian approach based on a general and efficient MCMC algorithm. Easily specified priors are used and Bayesian variable selection is carried out to obtain model parsimony and guard against over-fitting. We use the log predictive density score to determine the number of mixture components. Simulation and real examples show that using fairly complex components in the mixture is a wise strategy and that variable selection is an efficient approach to guard against over-fitting.

ACKNOWLEDGMENT

We would like to thank Denzil Fiebig for the use of the electricity expenditure data. Robert Kohn's research was partially supported by ARC Discovery grant DP0988579. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

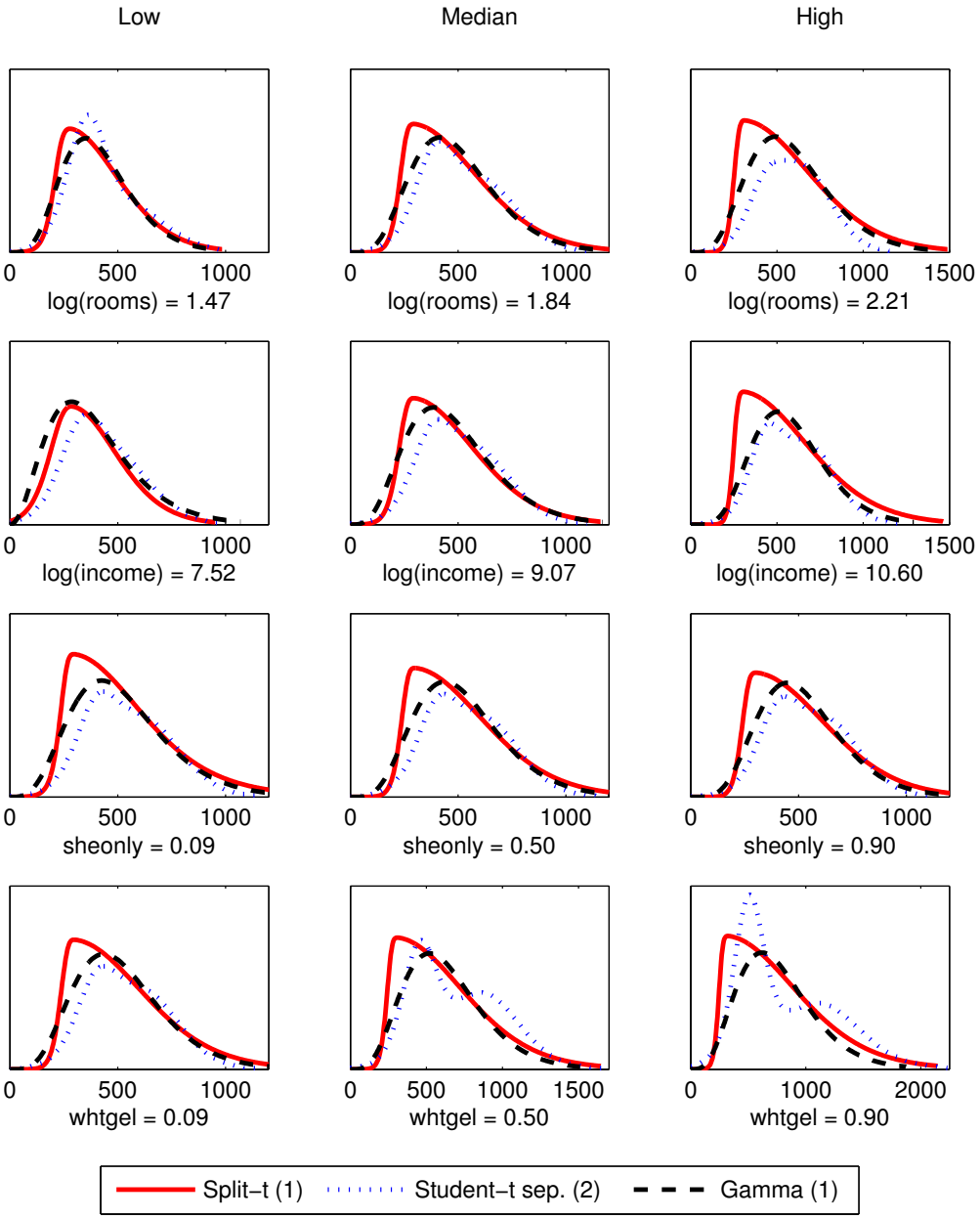

FIGURE 3. Conditional predictive densities for different values of the most important covariates. All other covariates are held fixed at their mean.

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

The general MCMC algorithm documented in Section 3 only requires the gradient and Hessian matrix of the conditional posteriors for each of the parameters in the components densities. The

gradient and Hessian for the split-t model is documented in Li et al. (2010). We now present the gradient and Hessian for the gamma model and log-normal model for completeness.

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

$$\begin{aligned}
\frac{\partial \ln p(y|\mu,\phi)}{\partial \mu} =& \frac{1}{\phi}\left(\mu + 2\mu\log\left(\frac{y\mu}{\phi}\right) - 2\mu\psi\left(\frac{\mu^2}{\phi}\right) - y\right)\\
\frac{\partial \ln p(y|\mu,\phi)}{\partial \phi} =& \frac{\mu}{\phi^2}\left(y - \mu - \mu\log\left(\frac{y\mu}{\phi}\right) + \mu\psi\left(\frac{\mu^2}{\phi}\right)\right)\\
\frac{\partial^2 \ln p(y|\mu,\phi)}{\partial \mu^2} =& \frac{1}{\phi}\left(3 + 2\log\left(\frac{y\mu}{\phi}\right)\right) - \frac{2}{\phi}\psi\left(\frac{\mu^2}{\phi}\right) - \frac{\mu^2}{\phi^2}\psi_1\left(\frac{\mu^2}{\phi}\right)\\
\frac{\partial^2 \ln p(y|\mu,\phi)}{\partial \phi^2} =& -\frac{\mu}{\phi^3}\left(2y - 3\mu - 2\mu\log\left(\frac{y\mu}{\phi}\right)\right) - \frac{2\mu^2}{\phi^3}\psi\left(\frac{\mu^2}{\phi}\right) - \frac{\mu^4}{\phi^4}\psi_1\left(\frac{\mu^2}{\phi}\right)
\end{aligned}$$

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

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

$$\begin{aligned}
\frac{\partial \log p(y|\mu,\phi)}{\partial \mu} =& \frac{\phi^2\left(3l^2 - 4h^2 + 4hl + 4l\right) + 2\mu^2\left(l^2 + 2hl\right)}{4\mu\left(\mu^2+\phi^2\right)l^2},\\
\frac{\partial \log p(y|\mu,\phi)}{\partial \phi} =& \frac{\phi\left(4h^2 - l^2 - 4l\right)}{4\left(\mu^2+\phi^2\right)l^2},\\
\frac{\partial^2 \log p(y|\mu,\phi)}{\partial \mu^2} =& -\frac{4\phi^4h^2}{\left(\mu^2+\phi^2\right)^2\mu^2l^3} + \frac{2\phi^4 + 4\left(\mu^2+\phi^2\right)\phi^2h + \left(3\mu^2+\phi^2\right)\phi^2h^2}{\left(\mu^2+\phi^2\right)^2\mu^2l^2}\\
& - \frac{2\phi^4 + \left(\mu^2+5\phi^2\right)\mu^2 + \left(\mu^2+\phi^2\right)^2h}{\left(\mu^2+\phi^2\right)^2\mu^2l} - \frac{\left(2\mu^2+\phi^2\right)\left(\mu^2+3\phi^2\right)}{4(\mu^2+\phi^2)^2\mu^2},\\
\frac{\partial^2 \log p(y|\mu,\phi)}{\partial \phi^2} =& -\frac{4\phi^2h^2}{\left(\mu^2+\phi^2\right)^2l^3} + \frac{2\phi^2 + (\mu^2-\phi^2)\left(h^2 + l^2/4 - l\right)}{\left(\mu^2+\phi^2\right)^2l^2}.
\end{aligned}$$

REFERENCES

BARTELS, R., FIEBIG, D. G. & PLUMB, M. H. (1996). Gas or electricity, which is cheaper? An econometric approach with application to Australian expenditure data. *The Energy Journal* **17**, 33–58.

CELEUX, G., HURN, M. & ROBERT, C. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association* **95**, 957.

DENISON, D., HOLMES, C. C., MALLICK, B. K. & SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Jone Wiley & Sons, Chichester.

DIEBOLT, J. & ROBERT, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 363–375.

ESCOBAR, M. & WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the american statistical association* **90**.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite mixture and Markov switching models*. Springer Verlag.

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

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

GEWEKE, J. & KEANE, M. (2007). Smoothly mixing regressions. *Journal of Econometrics* **138**, 252–290.

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

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

HOLST, U., HÖSSJER, O., BJÖRKLUND, C., RAGNARSON, P. & EDNER, H. (1996). Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of Lidar Measurements. *Environmetrics* **7**, 401–416.

JACOBS, R., JORDAN, M., NOWLAN, S. & HINTON, G. (1991). Adaptive mixtures of local experts. *Neural computation* **3**, 79–87.

JASRA, A., HOLMES, C. C. & STEPHENS, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20**, 50–67.

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

JIANG, W. & TANNER, M. A. (1999). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation* **11**, 1183–98.

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

JORDAN, M. & JACOBS, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation* **6**, 181–214.

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

KOHN, R., SMITH, M. & CHAN, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322.

LESLIE, D., KOHN, R. & NOTT, D. (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing* **17**, 131–146.

LI, F., VILLANI, M. & KOHN, R. (2010). Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference* **140**, 3638–3654.

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

NOTT, D. J. & LEONTE, D. (2004). Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics* **13**, 362–382.

NTZOUFRAS, I. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference* **111**, 165–180.

PENG, F., JACOBS, R. & TANNER, M. (1996). Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition. *Journal of the American Statistical Association* **91**.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **59**, 731–792.

RUPPERT, D., WAND, M. & CARROLL, R. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *The Annals of Statistics* **28**, 40–74.

VILLANI, M., KOHN, R. & GIORDANI, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* **153**, 155–173.

VILLANI, M., KOHN, R. & NOTT, D. (2010). A general approach to regression density estimation for discrete and continuous data. *Manuscript* .

WOOD, S., JIANG, W. & TANNER, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513.

ZEEVI, A. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks* .

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

Scandinavian Journal of Statistics
2013 (**in press**)

EFFICIENT BAYESIAN MULTIVARIATE SURFACE REGRESSION

FENG LI AND MATTIAS VILLANI

ABSTRACT. Methods for choosing a fixed set of knot locations in additive spline models are fairly well established in the statistical literature. The curse of dimensionality makes it non-trivial to extend these methods to non-additive surface models, especially when there are more than a couple of covariates. We propose a multivariate Gaussian surface regression model that combines both additive splines and interactive splines, and a highly efficient MCMC algorithm that updates all the knot locations jointly. We use shrinkage priors to avoid overfitting with different estimated shrinkage factors for the additive and surface part of the model, and also different shrinkage parameters for the different response variables. Simulated data and an application to firm leverage data show that the approach is computationally efficient, and that allowing for freely estimated knot locations can offer a substantial improvement in out-of-sample predictive performance.

KEYWORDS: Bayesian inference, free knots, Markov chain Monte Carlo, surface regression, splines.

1. INTRODUCTION

Flexible modeling of the regression function $\mathrm{E}(y|x)$ has been an active research field for decades, see e.g. Ruppert et al. (2003) for a recent textbook introduction and further references. Intensive research was initially devoted to kernel regression methods (Nadaraya, 1964; Watson, 1964; Gasser, 1979), and later followed by a large literature on spline regression modeling. A spline is a linear regression on a set of nonlinear basis functions of the original regressors. Each basis function is defined from a knot in regressor space and the knots determine the points of flexibility of the fitted regression function. This gives rise to a locally adaptable model with continuity at the knots.

The most widely used models assume additivity in the regressors, i.e. $\mathrm{E}(y|x_1,...,x_q) = \sum_{j=1}^{q} f_j(x_j)$, where $f_j(x_j)$ is a spline function for the jth regressor (Hastie & Tibshirani, 1986). Assuming additivity is clearly a very convenient simplification, but it is also somewhat unnatural to make such a strong assumption in an otherwise very flexible model. This has motivated research on surface models with interactions between regressors. One line of research extends the additive models by including higher-order interactions of the spline basis functions, see e.g. the structured ANOVA approach or the tensor product basis in Hastie et al. (2009). The multivariate adaptive regression splines introduced in Friedman (1991) is a version of the tensor product spline with interactions sequentially entering the model using a greedy algorithm. Regression trees (Breiman et al., 1984)

Li (corresponding author): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: feng.li@stat.su.se. Villani: Division of Statistics, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden. E-mail: mattias.villani@liu.se.

is another popular class of models, with the BART model in Chipman et al. (2010) as its most prominent Bayesian member. Our paper follows a recent strand of literature that models surfaces using radial basis functions splines, see e.g. Buhmann (2003). A radial basis function is defined in $\mathbb{R}^q$ and has a value that depends only on the distance from a covariate vector ($\boldsymbol{x}$) to its q-dimensional knot ($\boldsymbol{\xi}$), e.g. the cubic radial basis $\|\boldsymbol{x}-\boldsymbol{\xi}\|^3$, where $\boldsymbol{x}=(x_1,...,x_q)'$, $\boldsymbol{\xi}=(\xi_1,...,\xi_q)'$ and $\|\cdot\|$ is the Euclidean norm. The model is again linear in the basis expanded space.

The basic challenge in spline regression is the choice of knot locations. This problem is clearly much harder for a general surface than it is for additive models since any manageable set of q-dimensional knots are necessarily sparse in $\mathbb{R}^q$ when q is moderate or large, a manifestation of the curse of dimensionality. Most of the algorithms in the literature use a fixed set of knot locations, and the most prominent ones place the knots at the centroids from a clustering of the regressor observations. The response variables are not used in the clustering. To prevent overfitting, Bayesian variable selection methods are used to automatically remove or downweight the influence of the knots using Markov chain Monte Carlo (MCMC) techniques (Smith & Kohn, 1996). The reversible jump MCMC (RJMCMC) in for example Denison et al. (2002) treats the number of knots as unknown subject to an upper bound, but the location of the knots are still fixed throughout the analysis.

Using a fixed set of knot locations is impractical when estimating a surface with more than a few regressors. An algorithm that can move the knots rapidly over the regressor space is expected to be a clear improvement. All previous attempts have focused on efficient selection of fixed knots, and have paid little attention to moving the knots. The otherwise very elaborate RJMCMC approaches in Dimatteo et al. (2001), Denison et al. (1998), Gulam Razul et al. (2003) and Holmes & Mallick (2003) all include a very simple MCMC update where a single knot is re-located using a Metropolis random walk step with a proposal variance that is the same for all knots. There are typically strong dependencies between the knots, and local one-knot-at-a-time moves will lead to slow convergence of the algorithm and inability to escape from local modes, see Section 5.4 for some evidence.

The main contribution in this paper is a highly efficient MCMC algorithm for the Gaussian multivariate surface regression where the locations of all knots are updated jointly. Our joint updates of the knot locations are documented to dramatically increase the number of efficient MCMC draws for a fixed computing time. Rapid mixing of the knot locations is obtained from the following two features of our algorithm. First, the knots are simulated from a marginal posterior where the high-dimensional regression coefficients have been integrated out analytically. Second, the proposal distribution of the knots is tailored to the posterior distribution using the posterior gradient, which we derive in compact analytical form and evaluate efficiently by a careful use of sparsity.

Even a highly efficient MCMC algorithm is likely to have problems exploring the joint posterior of many surface knots in a high-dimensional covariate space. To deal with this, our model is decomposed into three parts: i) the original covariates entering in linear form, ii) additive spline basis functions and iii) radial basis functions for capturing the remaining part of the surface and interactions. The idea is to let the additive part of the model capture the bulk of the nonlinearities so that the radial basis functions can focus exclusively on modeling the interactions. This way

we can keep the number of knots in the interaction part of the model to a minimum, which is beneficial for MCMC convergence.

We use separate shrinkage priors for the three parts of the model. Moreover, we also allow for separate shrinkage parameters in each response equation. The shrinkage factors are treated as unknowns and estimated using a tailored joint MCMC updating step. This gives us an extremely flexible yet potentially parsimonious model where we can shrink out e.g. the surface part of the model in a subset of the response equations.

Our MCMC scheme is designed for a fixed number of knots, and we select the number of knots by Bayesian cross-validation of the log predictive score using parallel computing, see Section 3.3. This has the disadvantage of not accounting for the uncertainty regarding the number of knots as is done in RJMCMC schemes, but brings the advantages that the model choice is substantially more robust to variations in the prior, and that it is much easier to design an efficient MCMC algorithm. See Section 3.3 for a discussion.

We illustrate our algorithm on simulated and real data, and compare the predictive performance of the models using Bayesian cross-validation techniques. We find that the free knots model constantly outperforms the model with fixed knots. Additionally, we find it is easier to obtain better fitting result by combining additive knots and surface knots in the model.

2. BAYESIAN MULTIVARIATE SURFACE REGRESSION

2.1. **The model.** Our proposed model is a Gaussian multivariate regression with three sets of covariates:

$$\boldsymbol{Y} = \boldsymbol{X}_o\boldsymbol{B}_o + \boldsymbol{X}_a(\boldsymbol{\xi}_a)\boldsymbol{B}_a + \boldsymbol{X}_s(\boldsymbol{\xi}_s)\boldsymbol{B}_s + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{Y}(n \times p)$ contains n observations on p response variables, and the rows of $\boldsymbol{E}$ are error vectors assumed to be independent and identically distributed (*iid*) as $\mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The matrix $\boldsymbol{X}_o(n \times q_o)$ contains the original regressors (first column is a vector of ones for the intercept) and $\boldsymbol{B}_o$ holds the corresponding regression coefficients. The q_a columns of the matrix $\boldsymbol{X}_a(\boldsymbol{\xi}_a)$ are additive splines functions of the covariates in $\boldsymbol{X}_o$. Our notation makes it clear that $\boldsymbol{X}_a$ depends on the knots $\boldsymbol{\xi}_a$. Note that the knots in the additive part of the model are scalars, and that our model allows for unequal number of knots in the different covariates. Finally, $\boldsymbol{X}_s(\boldsymbol{\xi}_s)$ contains the surface, or interaction, part of the model. The knots in $\boldsymbol{\xi}_s$ are q_o-dimensional vectors. Note how this decomposition makes it possible for the additive part of the model to capture the main part of the nonlinearities so that the number of knots in $\boldsymbol{X}_s$ is kept to a minimum. We will refer to the three different parts of the model as the *linear component*, the *additive component* and the *surface component*, respectively. We will refer to $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_s$ as the additive and surface knots, respectively. Likewise, $\boldsymbol{B}_a$ and $\boldsymbol{B}_s$ are the additive and surface coefficients.

Denison et al. (2002) surveys the most commonly used spline bases. We use thin-plate splines for illustration, but our approach can be used with any basis with trivial changes, see Section 3 and Appendix A for computational details. The thin-plate spline basis in the surface case is of the form

$$\boldsymbol{x}_{sj}(\boldsymbol{\xi}_{sj}) = \|\boldsymbol{x}_o - \boldsymbol{\xi}_{sj}\|^2 \ln \|\boldsymbol{x}_o - \boldsymbol{\xi}_{sj}\|, \; j = 1, ..., q_s, \tag{2}$$

where $\boldsymbol{x}_o$ is one of the original data points and $\boldsymbol{\xi}_{sj}$ is the jth q_o-dimensional surface knot. The univariate thin-plate basis used in the additive part is a special case of the multivariate thin-plate in (2) where both the data point and the knot are one-dimensional.

For notational convenience, we sometimes write model (1) in compact form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{X} = [\boldsymbol{X}_o, \boldsymbol{X}_a, \boldsymbol{X}_s]$ is the $n \times q$ design matrix ($q = q_o + q_a + q_s$) and $\boldsymbol{B} = [\boldsymbol{B}_o', \boldsymbol{B}_a', \boldsymbol{B}_s']'$. Define also $\boldsymbol{b}_i = \text{vec}\boldsymbol{B}_i$ as the vectorization of the coefficients matrix $\boldsymbol{B}_i$, and $\boldsymbol{b} = [\boldsymbol{b}_o', \boldsymbol{b}_a', \boldsymbol{b}_s']'$.

For a given set of fixed knot locations, the model in (1) is linear in the regression coefficients $\boldsymbol{B}$. As explained in the Introduction, the great challenge with spline models is the choice of knot locations, especially in the surface case. We will treat the knot locations in $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_s$ as unknown parameters to be estimated and updated jointly in the MCMC. This is in principle straightforward from a Bayesian point of view, but great care is needed in the actual implementation of the posterior computations. We propose an efficient MCMC scheme for sampling from the joint posterior of the all knot locations and the regression coefficients, see Section 3 for details.

The model is clearly highly (over)parametrized and in need of some regularization of the parameters. The two main regularization techniques in Bayesian analysis are shrinkage priors and variable (knot) selection priors. Variable selection can in principle be incorporated in the analysis, but would be computationally demanding since the number of gradient evaluations needed in our MCMC algorithm would increase dramatically. This is important since evaluating the gradient with respect to the knots is time-consuming as the knot locations enter the likelihood in a very complicated nonlinear way; see Section 3.2 for details. Moreover, part of the attraction of variable selection is that they also provide interpretable measures of variable importance; this is much less interesting here since the covariates correspond to knot locations, which are not interesting in themselves. We have therefore chosen to achieving parsimony with shrinkage priors that pull the regression coefficients towards zero (or any other reference point if so desired), see Section 2.2 for details.

Allowing the knots to move freely in covariate space introduces a knot switching problem similar to the well-known label switching problem in mixture models. The likelihood is invariant to a switch of two knot locations and their regression coefficients. This lack of identification is not important if our aim is to model the regression surface $\text{E}(\boldsymbol{y}|\boldsymbol{x})$, without regard to the posterior of the individual knot locations (Geweke, 2007). Also, the MCMC draws of the knot locations can also be used to construct heat maps in covariate space to represent the density of knots in a certain regions, see Section 5. Such heat maps are clearly also immune to the knot switching problem.

2.2. **The prior.** We now introduce an easily specified shrinkage prior for the three sets of regression coefficients $\boldsymbol{B}_o$, $\boldsymbol{B}_a$ and $\boldsymbol{B}_s$ and the covariance matrix $\boldsymbol{\Sigma}$, conditional on the knots. The prior for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ are set as

$$\begin{aligned} \text{vec}\boldsymbol{B}_i|\boldsymbol{\Sigma},\ \boldsymbol{\lambda}_i &\sim \text{N}\left(\boldsymbol{\mu}_i,\ \boldsymbol{\Lambda}_i^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_i^{1/2} \otimes \boldsymbol{P}_i^{-1}\right),\ i \in \{o,a,s\}, \\ \boldsymbol{\Sigma} &\sim \text{IW}\left(n_0\boldsymbol{S}_0,\ n_0\right), \end{aligned}$$

with prior independence between the $\boldsymbol{B}_i$ conditionally on $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}_i$. The prior mean of $\text{vec}\boldsymbol{B}_i$ is $\boldsymbol{\mu}_i$, which we set to zero in our shrinkage prior. $\boldsymbol{\Lambda}_i = \text{diag}(\boldsymbol{\lambda}_i) = \text{diag}(\lambda_{i,1}, ..., \lambda_{i,p})$, $\boldsymbol{P}_i$ is a

positive definite symmetric matrix. IW($\cdot$) denotes the inverse Wishart distribution, with location matrix $\boldsymbol{S}_0$ and degrees of freedom n_0. $\boldsymbol{P}_i$ is typically either the identity matrix or $\boldsymbol{P}_i = \boldsymbol{X}_i'\boldsymbol{X}_i$. The latter choice has been termed a g-prior by Zellner (1986) and has the advantage of automatically adjusting for the different scales of the covariates. Setting $\lambda_i = n$ makes the information content of the prior equivalent to a single data point and is usually called the unit information prior. The choice of $\boldsymbol{P}_i = \boldsymbol{I}_{q_i}$ can prevent the design matrix from falling into singularity problem when some of the basis functions are highly correlated, which can easily happen with many spline knots. See also the discussion in Denison et al. (2002). Our default choice is therefore $\boldsymbol{P}_o = \boldsymbol{X}_o'\boldsymbol{X}_o$, $\boldsymbol{P}_a = \boldsymbol{I}_{q_a}$ and $\boldsymbol{P}_s = \boldsymbol{I}_{q_s}$. Other shrinkage priors on the regression coefficients can be used in our approach, for example the Laplace distribution leading to the popular Lasso (Tibshirani, 1996), but they will typically not allow us to integrate out the regression coefficients analytically, see Section 3.1. The optimal choice of shrinkage prior depends on the unknown data generating model (a normal prior is better when all coefficients have roughly the same magnitude; Lasso is better when many coefficients are close to zero, but some are really large etc).

We also estimate the shrinkage parameters, $\boldsymbol{\lambda}_o$, $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_s$ via a Bayesian approach. Note that our prior constructions for $\boldsymbol{B}$ allow for separate shrinkage of the linear, additive and surface components. This gives us automatic regularization/shrinkage of the regression coefficients and helps to avoid problems with overfitting. Our MCMC scheme in Section 3 allows for a user-specified prior on λ_{ij}, for $i \in \{o,a,s\}$ and $j = 1,2,...,p$ of essentially any functional form. However the default prior of λ_{ij} in this paper follows a log normal distribution with mean of $n/2$ and standard deviation of $n/2$ in order to ensure that both tight and flat shrinkages are attainable within one standard deviation in the prior. For computational convenience, we use a log link for λ_{ij} and make inference on $\log(\lambda_{ij})$. As a result the preceding prior on λ_{ij} yields a normal prior for $\log(\lambda_{ij})$ with mean $[\log(n) - 3/2 \cdot \log(2)]$ and variance $\log(2)$.

We use the same number of additive knots for each covariate in the simulations and the application in Section 4 and 5, but it should be clear that our approach also permits unequal number of knots in the different covariates. There is no particular requirements for the prior on the knots, but a vague prior should permit the knots to move freely in covariate space. Our default prior assumes independent knot locations following a normal distribution. The mean of the knots comes from the centers of a k-means clustering of the covariates. In the additive case, the prior variance of all the knots in the kth covariate is $c^2(\boldsymbol{a}'\boldsymbol{a})^{-1}$, where $\boldsymbol{a}$ is the kth column of $\boldsymbol{X}_o$. Similarly, the prior covariance matrix of a surface knot is $c^2(\boldsymbol{X}_o'\boldsymbol{X}_o)^{-1}$. We use $c^2 = n$ as the default setting.

The hyperparameter $\boldsymbol{S}_0$ in the IW prior for $\boldsymbol{\Sigma}$ should in principle be chosen subjectively, but in our application we set it equal to the estimated error covariance matrix from the fitted linear model $\hat{\boldsymbol{Y}} = \boldsymbol{X}_o\hat{\boldsymbol{B}}_o$, for simplicity. This is not a crucial choice since we use a relatively small degrees of freedom (n_0). We use $n_0 = 10$ as our default choice.

For notational convenience and further computational implementation, we write the prior for the regression coefficients in condensed form as $\boldsymbol{b}|\boldsymbol{\Sigma},\boldsymbol{\lambda} \sim \mathrm{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_{\boldsymbol{b}})$ where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_o', \boldsymbol{\lambda}_a', \boldsymbol{\lambda}_s')'$, $\boldsymbol{\mu}^* = (\boldsymbol{\mu}_o', \boldsymbol{\mu}_a', \boldsymbol{\mu}_s')'$, $\boldsymbol{\Sigma}_{\boldsymbol{b}} = (\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Sigma}_K\boldsymbol{\Lambda}^{1/2}) \circledast \boldsymbol{P}^{-1}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\Sigma}_K$ is a three-block diagonal matrix with $\boldsymbol{\Sigma}$ on each block, $\boldsymbol{P} = \mathrm{diag}(\boldsymbol{P}_o, \boldsymbol{P}_a, \boldsymbol{P}_s)$ is a block diagonal matrix and $\boldsymbol{A} \circledast \boldsymbol{C}$ denotes the Khatri-Rao product (Khatri & Rao, 1968) which is Kronecker product of the corresponding blocks of matrices $\boldsymbol{A}$ and $\boldsymbol{C}$. It will also be convenient to define $\boldsymbol{\beta} = \mathrm{vec}\boldsymbol{B}$. Note that $\boldsymbol{b}$ and $\boldsymbol{\beta}$ contain

the same elements with two different stacking orders. As a result, $\beta|\Sigma,\lambda \sim \mathrm{N}\left(\mu,\Sigma_\beta\right)$ where μ and Σ_β essentially have the same entries as μ^* and Σ_b have, respectively (Appendix A.3).

3. The posterior inference

3.1. **The posterior.** The posterior distribution can be decomposed as

$$p(B,\Sigma,\xi,\lambda|Y,X) = p(B|\xi,\lambda,\Sigma,Y,X)p(\xi,\lambda,\Sigma|Y,X),$$

where

$$\mathrm{vec}B|\xi,\lambda,\Sigma,Y,X \sim \mathrm{N}(\tilde{\beta},\ \Sigma_{\tilde{\beta}}),$$

$\Sigma_{\tilde{\beta}} = [\Sigma^{-1}\otimes X'X + \Sigma_\beta^{-1}]^{-1}$, $\tilde{\beta} = \mathrm{vec}\tilde{B} = \Sigma_{\tilde{\beta}}[\mathrm{vec}(X'Y\Sigma^{-1}) + \Sigma_\beta^{-1}\mu]$ (Zellner, 1971), and

$$\begin{aligned} p\left(\xi,\lambda,\Sigma|Y,X\right) = c\times p(\xi,\lambda)\times|\Sigma_\beta|^{-1/2}|\Sigma|^{-(n+n_0+p+1)/2}|\Sigma_{\tilde{\beta}}|^{-1/2} \\ \times\exp\left\{-\frac{1}{2}\left[\mathrm{tr}\Sigma^{-1}\left(n_0S_0+n\tilde{S}\right)+\left(\tilde{\beta}-\mu\right)'\Sigma_\beta^{-1}\left(\tilde{\beta}-\mu\right)\right]\right\} \end{aligned} \quad (3)$$

where we allow for separate shrinkage parameters for the linear, additive and surface parts of the model, and separate shrinkage parameters for the p responses within each of the three model parts. The shrinkage parameters are treated as unknowns and estimated, so that, for example, the surface part can be shrunk towards zero if this agrees with the data. $\tilde{S} = (Y-X\tilde{B})'(Y-X\tilde{B})/n$, $c = 2^{-(n_0+n+q)p/2}\pi^{-p(n+q)/2}\Gamma_p^{-1}(n_0/2)|n_0S_0|^{n_0/2}$, $\Gamma_p(a) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left[a+(1-j)/2\right]$ is the multivariate gamma function. It is important to note that it is in general not possible to integrate out Σ analytically in our model. This is a consequence of using different shrinkage factors for the different responses and on the original, additive and surface parts of the model (the prior covariance matrix of B does not have a Kronecker structure). Only in the special case with a univariate response ($p=1$) can we integrate out Σ analytically, since Σ is then a scalar. To obtain a uniform treatment of the models and their gradients, we have chosen to not integrate out Σ even for the case $p=1$. The next subsection proposes an MCMC algorithm for sampling from the joint posterior distribution of all parameters.

3.2. **The MCMC algorithm.** Our approach is to sample from $p\left(\xi,\lambda,\Sigma|Y,X\right)$ using a three-block Gibbs sampling algorithm with Metropolis-Hastings (MH) updating steps. Draws from $p(B|\xi,\lambda,\Sigma,Y,X)$ can subsequently be obtained by direct simulation. The updating steps of the Gibbs sampling algorithm are:

(1) Simulate Σ from $p(\Sigma|\xi,\lambda,Y,X)$.
(2) Simulate ξ from $p(\xi|\lambda,\Sigma,Y,X)$.
(3) Simulate λ from $p(\lambda|\xi,\Sigma,Y,X)$.

In the special case when $p=1$

$$\Sigma|\xi,\lambda,Y,X \sim \mathrm{IW}\left(n_0S_0+n\tilde{S}+\sum\nolimits_{i\in\{o,a,s\}}\Lambda_i^{-1/2}(\tilde{B}_i-M_i)'P_i(\tilde{B}_i-M_i)\Lambda_i^{-1/2},\ n_0+n\right) \quad (4)$$

where M_i and $\tilde{B}_i$ are the prior and posterior mean of B_i, respectively, and the IW density reduces to a scaled χ^2 distribution. When $p>1$, $p(\Sigma|\xi,\lambda,Y,X)$ is no longer IW, but the distribution in (4) is an excellent approximation and can be used as a very efficient MH proposal density.

The conditional posterior distributions for $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ in Steps (2) and (3) above are highly non-standard and we update these parameters using Metropolis-Hastings steps with a tailored proposal, which we now describe for a general parameter vector $\boldsymbol{\theta}$ with posterior $p(\boldsymbol{\theta}|\boldsymbol{Y})$, which could be a conditional posterior in a Metropolis-within-Gibbs algorithm (e.g. $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\boldsymbol{\Sigma},\boldsymbol{Y},\boldsymbol{X})$). This method was originally proposed by Gamerman (1997) and later extended by Nott & Leonte (2004) and Villani et al. (2012). All of these three articles are confined to a generalized linear model (GLM) or GLM-like context where the parameters enter the likelihood function through a scalar-valued link function. A contribution of our paper is to show that the algorithm can be extended to models with a less regular structure and that it retains its efficiency even when the parameters are high-dimensional and enter the model in a highly nonlinear way. The way the knot locations and the shrinkage parameters are buried deep in the marginal posterior (see Equation 3.1 above) makes the necessary gradients (see below) much more involved and numerically challenging (see Appendix A).

At any given MCMC iteration we use Newton's method to iterate R steps from the current point $\boldsymbol{\theta}_c$ in the MCMC sampling towards the mode of $p(\boldsymbol{\theta}|\boldsymbol{Y})$, to obtain $\hat{\boldsymbol{\theta}}$ and the Hessian at $\hat{\boldsymbol{\theta}}$. Note that $\hat{\boldsymbol{\theta}}$ may not be the mode but is typically close to it already after a few Newton iterations since the previously accepted $\boldsymbol{\theta}$ is used as the initial value; setting $R = 1, 2$ or 3 is therefore usually sufficient. This makes the algorithm very fast. Having obtained good approximations of the posterior mode and covariance matrix from the Newton iterations, the proposal $\boldsymbol{\theta}_p$ is now drawn from the multivariate $\boldsymbol{t}$-distribution with $\nu > 2$ degrees of freedom:

$$\boldsymbol{\theta}_p|\boldsymbol{\theta}_c \sim \boldsymbol{t}\left[\hat{\boldsymbol{\theta}},\ -\left(\frac{\partial^2 \ln p(\boldsymbol{\theta}|\boldsymbol{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)^{-1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},\ \nu\right],$$

where the second argument of the density is the covariance matrix and $\hat{\boldsymbol{\theta}}$ is the terminal point of the R Newton steps. The Metropolis-Hastings acceptance probability is

$$a(\boldsymbol{\theta}_c \to \boldsymbol{\theta}_p) = \min\left[1,\ \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_p)p(\boldsymbol{\theta}_p)g(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}{p(\boldsymbol{Y}|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c)g(\boldsymbol{\theta}_p|\boldsymbol{\theta}_c)}\right].$$

The proposal density at the current point $g(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)$ is a multivariate $\boldsymbol{t}$-density with mode $\tilde{\boldsymbol{\theta}}$ and covariance matrix equal to the negative inverse Hessian evaluated at $\tilde{\boldsymbol{\theta}}$, where $\tilde{\boldsymbol{\theta}}$ is the point obtained by iterating R steps with the Newton algorithm, *this time starting from* $\boldsymbol{\theta}_p$. The need to iterate backwards from $\boldsymbol{\theta}_p$ is clearly important to fulfill the reversibility of the Metropolis-Hastings algorithm. When the number of parameters in $\boldsymbol{\theta}$ is large one can successively apply the algorithm to smaller blocks of parameters in $\boldsymbol{\theta}$.

The tailored proposal distribution turns out to be hugely beneficial for MCMC efficiency, see Section 5.4 for some evidence, but a naive implementation can easily make the gradient and Hessian evaluations an insurmountable bottleneck in the computations, and a source of numerical instability. We have found the outer product of gradients approximation of the Hessian to work very well, so all we need to implement efficiently are the gradient vector of $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\boldsymbol{\Sigma},\boldsymbol{Y},\boldsymbol{X})$ and $p(\boldsymbol{\lambda}|\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{Y},\boldsymbol{X})$. Appendix A gives compact analytical expression for these two gradient vectors, and shows how to exploit sparsity to obtain fast and stable gradient evaluations. Our gradient evaluations can easily be orders of magnitudes faster than state-of-the-art numerical derivatives,

and substantially more stable numerically. For example, already in a relatively small-dimensional model in Section 5 with only four covariates, 20 surface knots and 4 additive knots, the analytical gradient for the knot parameters are more than 40 times faster compared to a numerical gradient with a tolerance of 10^{-3}. Since the gradient evaluations accounts for 70-90% of total computing time, this is clearly an important advantage.

3.3. **Model comparison.** The number of knots is determined via the D-fold out-of-sample log predictive density score (LPDS), defined as

$$\frac{1}{D}\sum_{d=1}^{D} \ln p(\tilde{\boldsymbol{Y}}_d|\tilde{\boldsymbol{Y}}_{-d},\boldsymbol{X}),$$

where $\tilde{\boldsymbol{Y}}_d$ is an $(n_d \times p)$-dimensional matrix containing the n_d observations in the dth testing sample and $\tilde{\boldsymbol{Y}}_{-d}$ denotes the training observations used for estimation. If we assume that the observations are independent conditional on $\boldsymbol{\theta}$, then

$$p(\tilde{\boldsymbol{Y}}_d|\tilde{\boldsymbol{Y}}_{-d},\boldsymbol{X}) = \int \prod_{i\in\tau_d} p(\boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{x}_i)p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})\mathrm{d}\boldsymbol{\theta},$$

where τ_d is the index set for the observations in $\tilde{\boldsymbol{Y}}_d$, and the LPDS is easily computed by averaging $\prod_{i\in\tau_d} p(\boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{x}_i)$ over the posterior draws from $p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})$. This requires sampling from each of the D posteriors $p(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{-d})$ for $d = 1,...,D$, but these MCMC runs can all be run in isolation from each other and are therefore ideal for straightforward parallel computing on widely available multi-core processors. The main advantage for choosing LPDS instead of the marginal likelihood (which underlies the inference from RJMCMC) is that the LPDS is not nearly as sensitive to the choice of prior as the marginal likelihood, see e.g. Kass (1993) and Richardson & Green (1997) for a general discussion. The reason is that the LPDS uses the training data, $\tilde{\boldsymbol{Y}}_{-d}$, to update the prior before evaluating the test data. The marginal likelihood can also lead to poor predictive inference when the true data generating process is not included in the class of compared models, see e.g. Geweke & Amisano (2011) for an illuminating perspective. The main disadvantage of using the LPDS for selecting the number of knots is that, unlike the marginal likelihood and RJMCMC, there is no rigorous way of including the uncertainty regarding the number of knots in the final inferences. The dataset is systematically partitioned into five folds in our firm leverage application in Section 5.

4. SIMULATIONS

We compare the performance of the traditional fixed knots approach to our approach with freely estimated knot locations using simulated data with different number of covariates and for varying degrees of nonlinearity in the true surface. We use shrinkage priors with estimated shrinkage both for the fixed and free knot models, but no variable selection.

4.1. **Simulation setup.** We consider data generating processes (DGP) with both univariate ($p = 1$) and bivariate ($p = 2$) responses, and datasets with $q_o = 10$ regressors and two sample sizes, $n = 200$ and $n = 1000$. We first generate the covariate matrix $\boldsymbol{X}_o$ from a mixture of multivariate normals with five components. The weight for the rth mixture component is $u_r/\sum_{l=1}^{5} u_l$, where $u_1,...,u_5$ are independent $\mathrm{U}(0,1)$ variables. The mean of each component is a draw from $\mathrm{U}(-1,1)$

and the components' variances are all 0.1. We randomly select five observations without replacement from $\boldsymbol{X}_o$ as the true surface knots $\boldsymbol{\xi}_s$, and then create the basis expanded design matrix $\boldsymbol{X}$ using the thin-plate radial basis surface spline, see Section 2.1. The coefficients matrix $\boldsymbol{B}$ is generated by repeating the sequence $\{-1,1\}$. The error term $\boldsymbol{E}$ is from multivariate normal distribution with mean zero, variance 0.1 and covariance 0.1. The average signal-to-noise ratio in the DGP is roughly three times larger than that in the real data used in Section 5.

Following Wood et al. (2002), we measure the degrees of nonlinearity (DNL) in the DGP by the distance between the true surface $f(\cdot)$ and the plane $\hat{g}(\cdot)$ fitted by ordinary least squares without any knots in the model, i.e.

$$\text{DNL} = \sqrt{n^{-1}\sum_{i=1}^{n}[f(\boldsymbol{x}_i) - \hat{g}(\boldsymbol{x}_i)]^2}. \tag{5}$$

A larger DNL indicates a DGP with stronger nonlinearity.

We generate 100 datasets and for each dataset we fit the fixed knots model with 5, 10, 15, 20, 25 and 50 surface knots, and also the free knots model with 5, 10, and 15 surface knots. All fitted models have only linear and surface components. The knot locations are determined by k-means clustering. We compare the models with respect to the mean squared loss

$$\text{Loss}(\tilde{f}_{q_s}) = \frac{1}{n^*}\sum_{i=1}^{n^*}[f(\boldsymbol{x}_i) - \tilde{f}_{q_s}(\boldsymbol{x}_i)]^2 \tag{6}$$

where $f(\cdot)$ is the true surface and $\tilde{f}_{q_s}(\cdot)$ is the posterior mean surface of a given model with q_s surface knots. The Loss in (6) is evaluated over a new sample of n^* covariate vectors, and it therefore measures out-of-sample performance of the posterior mean surface. We will here set $n^* = n$. Note that the shrinkages and the covariance matrix of the error terms are also estimated in both the fixed and free knots models.

4.2. **Results.** We present the results for $p = 2$ and $n = 200$. The results for $p = 1$ and $n \in \{200, 1000\}$, and $p = 2$ and $n = 1000$ are qualitatively similar and are available upon request. Appendix C documents the results for $p = 2$ and $n = 1000$ for a few different model configurations. Figure 1 displays boxplots for the log ratio of the mean squared loss in (6). The columns of the figure represents varying degrees of nonlinearity in the generated datasets according to the estimated DNL measure in equation (5). Each boxplot shows the relative performance of a fixed knots model with a certain number of knots compared to the free knots model with 5 (top row), 10 (middle row) and 15 (bottom row) surface knots, respectively. The short summary of Figure 1 is that the free knots model outperforms the fixed knots model in the large majority of the datasets. This is particularly true when the data are strongly nonlinear. The performance of the fixed knots model improves somewhat when we add more knots, but the improvement is not dramatic. Having more fixed knots clearly improves the chances of having knots close to the true ones, but more knots also increase the risk of overfitting.

The aggregate results in Figure 1 do not clearly show how strikingly different the fixed and free knots models can perform on a given dataset. We will now show that models with free rather than fixed knots are much more robust across different datasets. Figure 2 displays the Euclidean distance of the multivariate *out-of-sample* predictive residuals $\sqrt{\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}$ for a few selected datasets as a function of the distance between the covariate vector and the sample mean of the covariates.

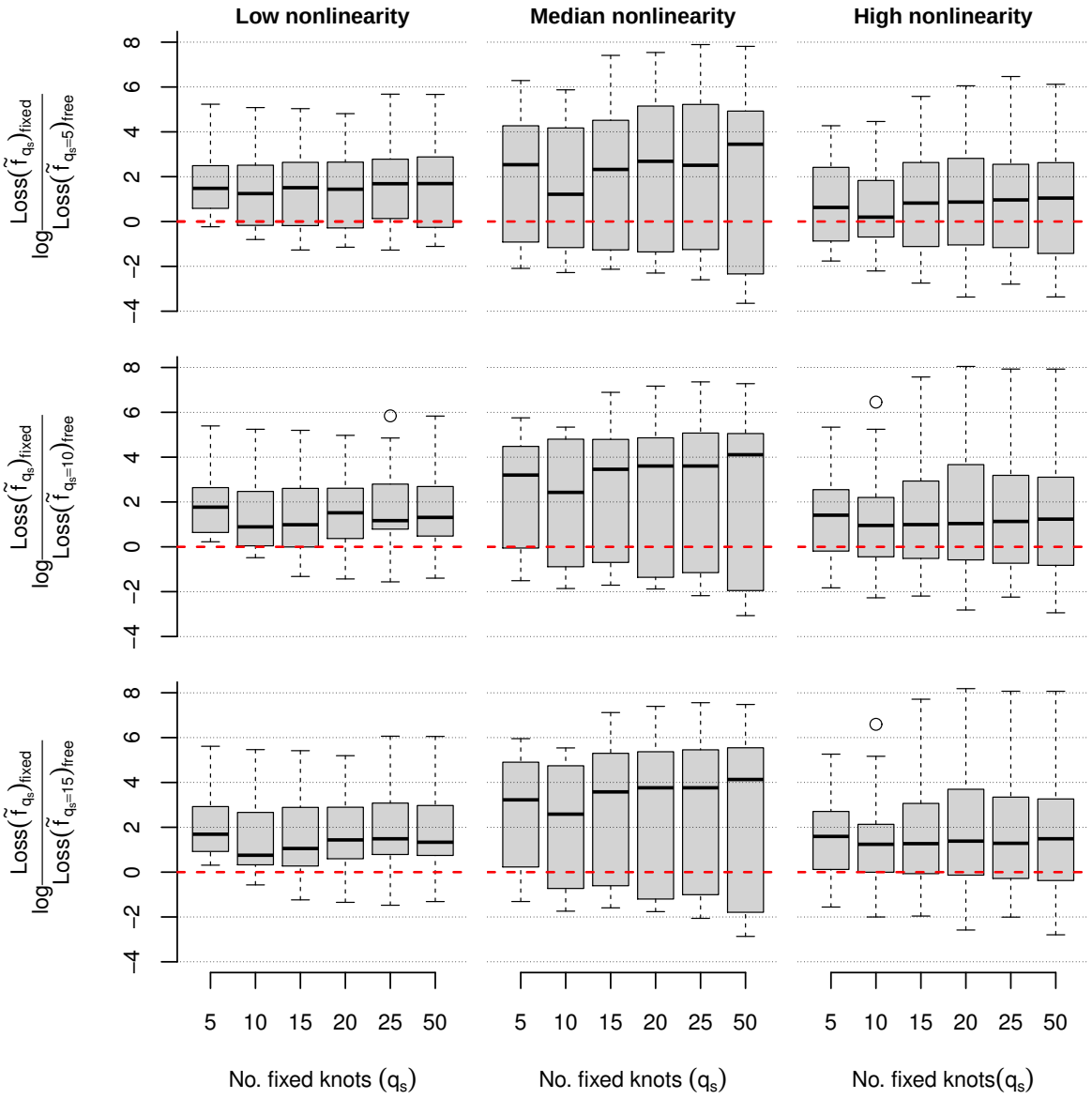

FIGURE 1. Boxplot of the log loss ratio comparing the performance of the fixed knots model with the free knots model for the DGP with $p = 2$ and $n = 200$. The three columns of the figure correspond to different degrees of nonlinearity of the realized datasets, as measured by estimated DNL in (5).

The normed residuals depicted in the leftmost column are from datasets chosen with respect to the ranking of the out-of-sample performance of the fixed knots model. For example, the upper left subplot shows the predictive residuals of both the model with 15 fixed knots (vertical bars above the zero line) and the model with 5 free knots (vertical bars below the zero line) on one of the datasets where the fixed knot models outperform the free knots model by largest margin

TABLE 1. Computing times (in minutes) for 5,000 iterations with a single dataset of 10 covariates.

	$n = 200$		$n = 1000$	
No. of free surface knots	$p = 1$	$p = 2$	$p = 1$	$p = 2$
2	9	9	16	17
5	13	14	23	26
10	17	18	42	45
15	24	27	61	75

(3rd best Loss in favor of fixed knots model). It is seen from this subplot that even in this very favorably situation for the fixed knots model, the free knots model is not generating much larger predictive residuals. Moving down to the last row in the left hand column of Figure 2, we see the performance of the two models when the fixed knots model performs very poorly (3rd worse Loss with respect to the fixed knots model). On this particular dataset, the free knots model does well while the fixed knots model is a complete disaster (note the different scales on the vertical axes of the subplots). The column to the right in Figure 2 shows the same analysis, but this time the datasets are chosen with respect to the ranking of the Loss of the free knots model. Overall, Figure 2 clearly illustrates the superior robustness of models with free knots: the free knots model never does much worse than the fixed knots model, but using fixed rather than free knots can lead to a dramatically inferior predictive performance on individual datasets.

4.3. **Computing time.** The program is written in native R code and all the simulations were performed on a Linux desktop with 2.8 GHz CPU and 4 GB RAM on single instance (without parallel computing). Table 1 shows the computing time in minutes for a single dataset. In general the computing time increases as the size of the design matrix increases, but it increases only marginally as we go from $p = 1$ to $p = 2$.

5. APPLICATION TO FIRM CAPITAL STRUCTURE DATA

5.1. **The data.** We illustrate our surface model in a finance application where a firm's leverage (fraction of external financing) is modeled as a function of the proportion of fixed assets, the firm's market value in relation to its book value, firm sales and profits. We use a similar data to the one in Rajan & Zingales (1995) which covers $4,405$ American non-financial firms with positive sales in 1992 and complete data records. See Table 2 for a definition of the variables in our dataset.

Figure 3 plots the response variable leverage in both original scale and logit scale ($\ln[y/(1-y)]$) against each of the four covariates. The relationships between leverage and the covariates are clearly highly nonlinear even when the logit transformation is used. There are also outliers which can be seen from the subplots with respect to covariates Market2Book and Profit. Strong nonlinearities seem to be a quite general feature of balance sheet data, but only a handful articles have suggested using nonlinear/nonparametric models, see e.g. Bastos & Ramalho (2010), Giordani et al. (in press) and Villani et al. (2012).

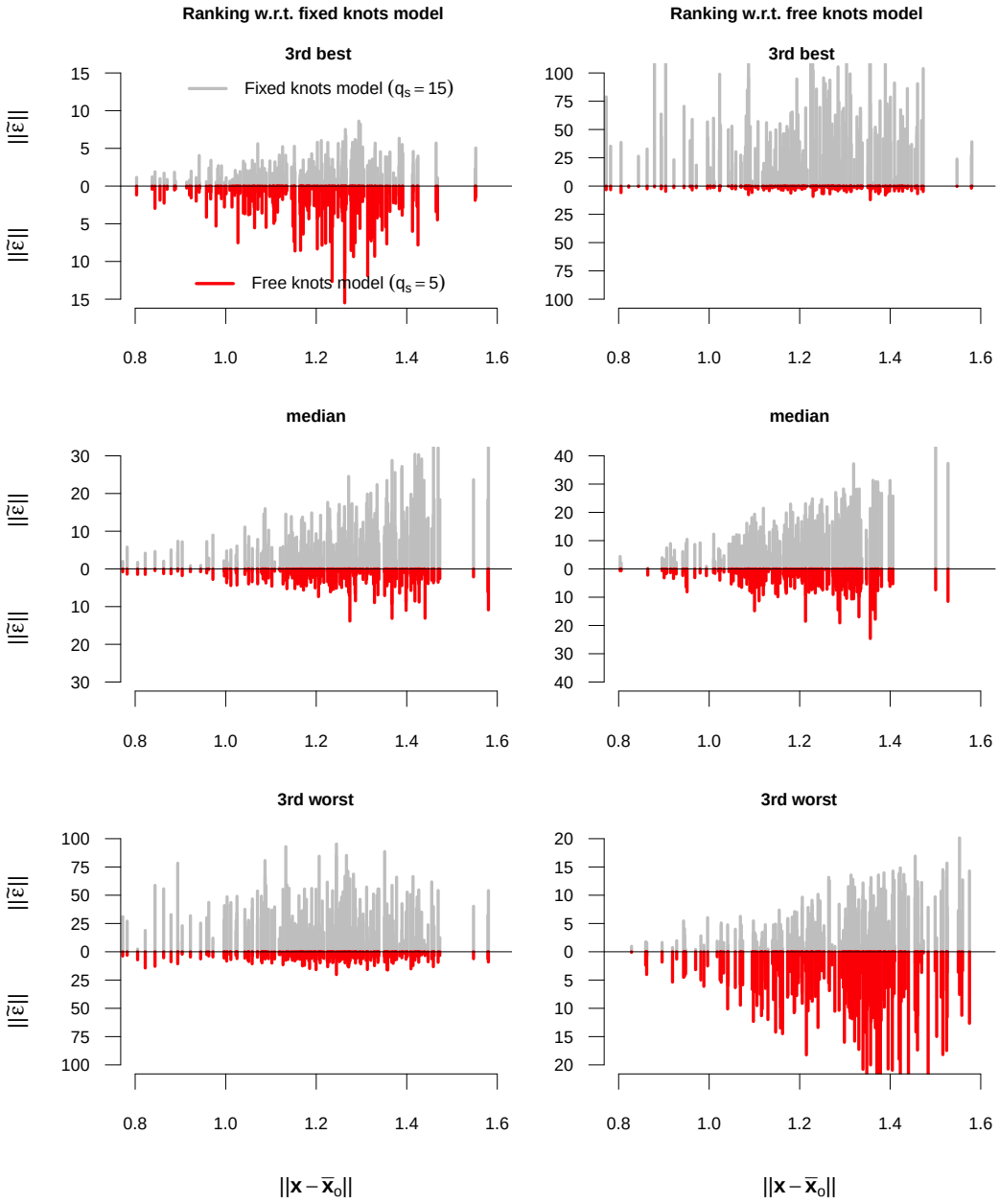

FIGURE 2. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 200$. The lines in each subplot are the normed residuals from the model with 15 fixed surface knots (vertical bars above the zero line), and the model with 5 free knots (vertical bars below the zero line). The column to the left shows the results for three datasets chosen when performance is ranked according to the fixed knots model, and the right column displays the results for three datasets chosen when performance is ranked according to the free knots model.

TABLE 2. Definitions of the variables in the firm capital structure data.

Variable name	Definition
Leverage	$\frac{\text{total debt}}{\text{total debt + book value of equity}}$
Tang	$\frac{\text{tangible assets}}{\text{book value of total assets}}$
Market2Book	$\frac{\text{book value of total assets - book value of equity + market value of equity}}{\text{book value of total assets}}$
LogSale	log of total sales
Profit	$\frac{\text{earnings before interest, taxes, depreciation, and amortization}}{\text{book value of total assets}}$

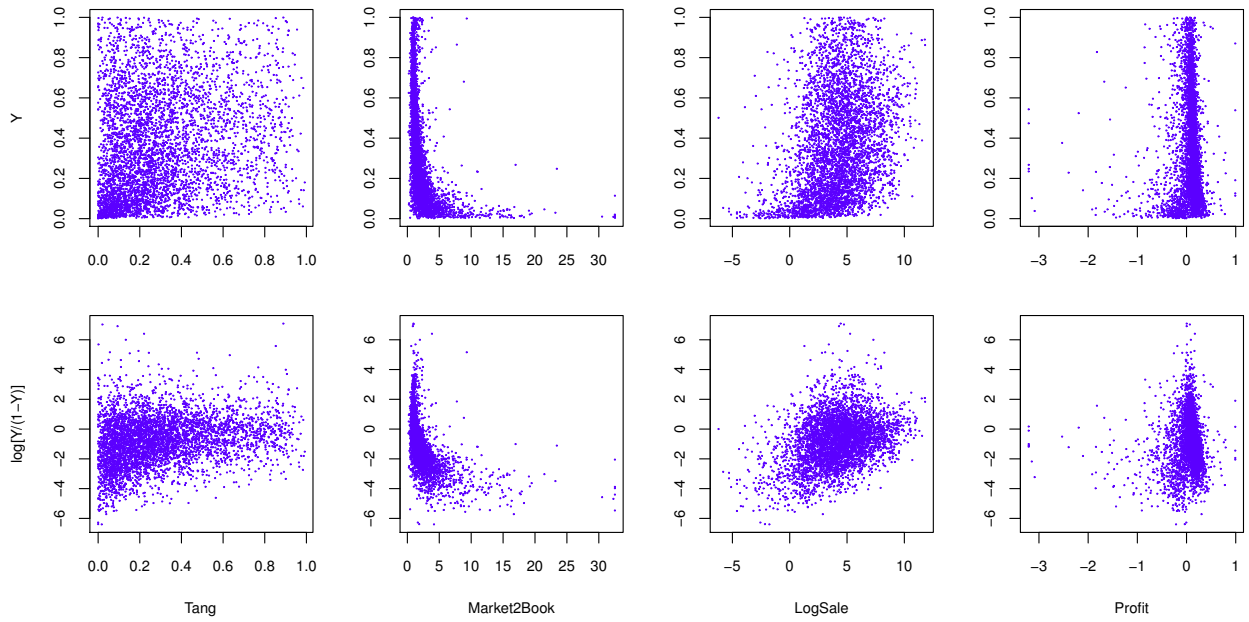

FIGURE 3. Scatter plots of the firm leverage data with leverage ($\boldsymbol{Y}$) on both original scale (top subplots) and logit transformed scale (bottom subplots) against each of the four covariates.

5.2. **Models with only surface or additive components.** We first fit models that either have only a surface component or only an additive component (both types of models also have a linear component). Note that the shrinkage parameters are also estimated in all cases. All four covariates are used in the estimation procedure and we use the logit transformation of the leverage, and standardize each covariate to have zero mean and unit variance.

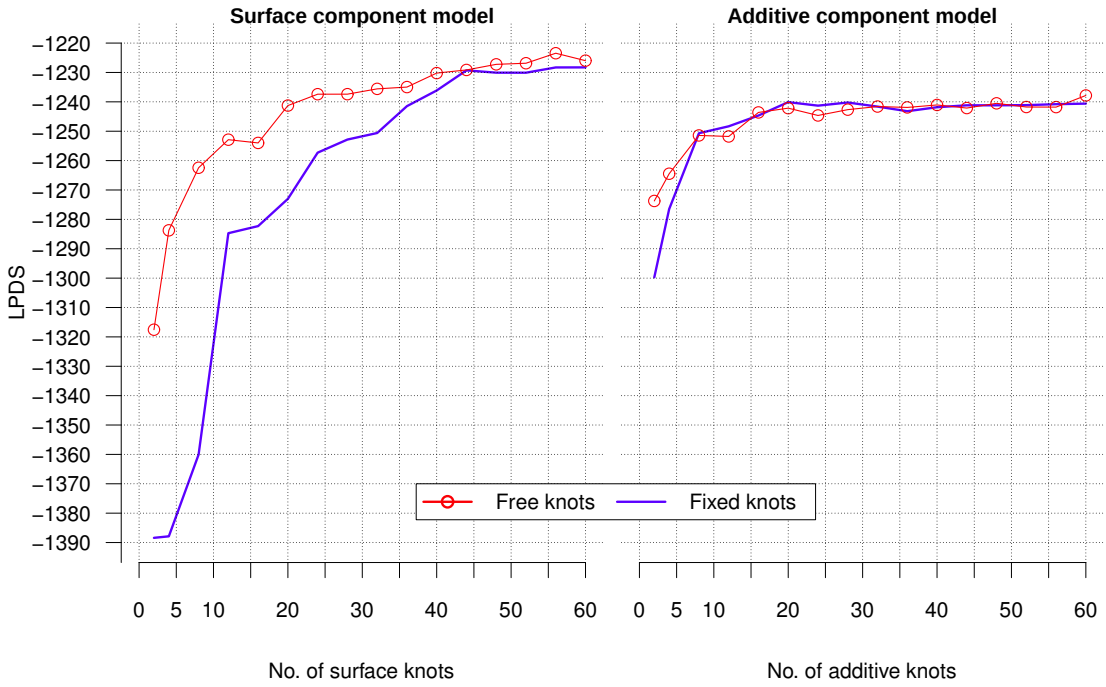

FIGURE 4. LPDS for the firm leverage data with surface component model (left) and additive component model (right). Note that the number of knots in additive model is the number of spline basis functions on each covariate.

Figure 4 depicts the LPDS for the surface component model and the additive component model for both the case of fixed and free knots. The LPDS generally improves as the number of knots increases for both the fixed and free knots models, but seems to eventually level off at large number of knots. The free knots model always outperforms the fixed knots model when only a surface component is used (left subplot). For example, the model with 12 free surface knots is roughly 32 LPDS units better than the fixed knots model with the same number of knots. This is a quite impressive improvement in out-of-sample performance considering that the fixed knot locations are chosen with state-of-the-art clustering methods for knot selection. The ability to move the knots clearly also helps to keep the number of knots to a minimum; it takes for example more than 30 fixed surface knots to obtain the same LPDS as a model with 12 free surface knots.

Turning to the strictly additive models in right subplot of Figure 4 we see that the additive models are in general inferior to the models with only surface knots, and that the differences in LPDS between the fixed and free knots approaches are much smaller here, at least for eight knots or more. The improvement in LPDS levels off at roughly 16 knots. It is important to note that the horizontal axis in Figure 4 displays the number of additive knots *in each covariate*, and the fact that we do not overfit bear testimony to the effectiveness of the shrinkage priors.

5.3. **Models with both additive and surface components.** We now consider models with both additive and surface components. It is worth mentioning that we draw from the joint posterior distribution of the surface and additive knots, see Appendix A for MCMC details.

Figure 5 shows that there are generally improvements from using both surface knots and additive knots in the same model. For example, the model with 4 free surface knots has an LPDS of $-1,284$. Adding two free additive knots increases the LPDS to $-1,270$ and adding another two additive knots gives a further increase of 14 LPDS units. Figure 5 also shows strong gains from estimating the knots' locations, but the improvement in LPDS from free knots tends to be less dramatic when more additive knots are used to complement the surface knots. There is little or no improvement in LPDS as the number of surface knots approaches 60. The results in Figure 5 reinforces the evidence in Figure 4 that the shrinkage prior is very effective in mitigating potential problems with overfitting.

To simplify the graphical presentation of the results, we choose to illustrate the posterior inference of the knot locations in a model with only the two covariates Market2Book and Profit. We use 20 surface knots and 4 additive knots in each covariate. The mean acceptance probabilities for the knot locations and the shrinkage parameters in Metropolis-Hastings algorithm are 0.73 and 0.64, respectively, which are exceptionally large considering that all $2 \times 20 + 2 \times 4 = 48$ knot location parameters are proposed jointly, as are all the shrinkage parameters. The acceptance probability in the updating step for $\boldsymbol{\Sigma}$ is 1 since we are proposing directly from the exact conditional posterior when $p = 1$. Because of the knot switching problem (see Section 3), it does not make much sense to display the posterior distribution of the knot locations directly. We instead choose to partition the covariate space into small rectangular regions, count the frequency of knots in each region over the MCMC iterations, and use heat maps to visualize the density of knots in different regions of covariate space. Figure 6 displays this knot density heat map. As expected, the estimated knot locations are mostly concentrated in the data dense regions, particularly in regions where the relation between the covariates and response in the data is most nonlinear, which is seen by comparing Figure 6 and Figure 3.

Finally, we present the posterior surface for the firm leverage data in Figure 7. To enhance the visual representation, the graphs zoom in on the region with the majority of the data observations. Figure 7 plots the mean (left) and the standard deviation (right) of the posterior surface. The latter object is for brevity sometimes referred to as the *posterior standard deviation surface*. Figure 7 (right) also displays the covariate observations to give a sense of where the data observations are located. The appendix to this article investigates the robustness of the posterior results to variations in both the prior mean and variance of the knot locations. The posterior heat map of the knot locations are affected by the fairly dramatic variations in the prior mean of the knots, and to a lesser extent by changes in the prior variance of the knot locations, but the posterior mean and standard deviation surfaces are robust to variations in the prior on the knots, especially in data dense regions. The appendix also shows that the posterior is robust to changes in the prior on the shrinkage factors.

5.4. **MCMC efficiency in the updating of the knot locations.** In order to study the efficiency of our algorithm for sampling the knot locations, we compare three types of MCMC updates of the knots: i) one-knot-at-a-time updates using a random walk Metropolis proposal with tuned variance (SRWM), ii) one-knot-at-a-time updates with the tailored Metropolis-Hastings step (SMH) in Section 3.2, and iii) full block updating of all knots using the tailored Metropolis-Hastings step (BMH) in Section 3.2. SRWM moves are used in state-of-the-art RJMCMC approaches such as

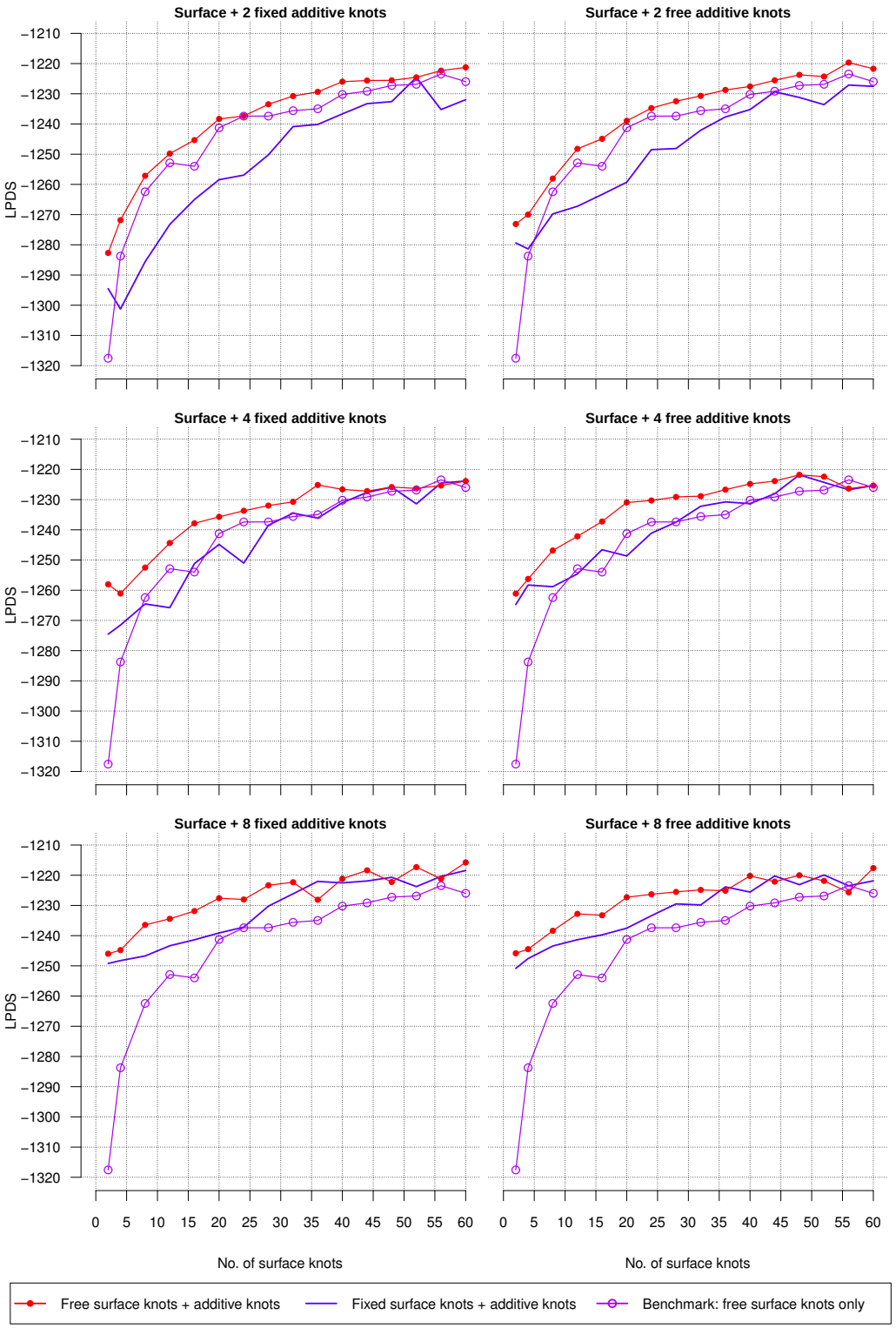

FIGURE 5. LPDS for the firm leverage data for the free and fixed knots models with varying number of surface and additive knots.

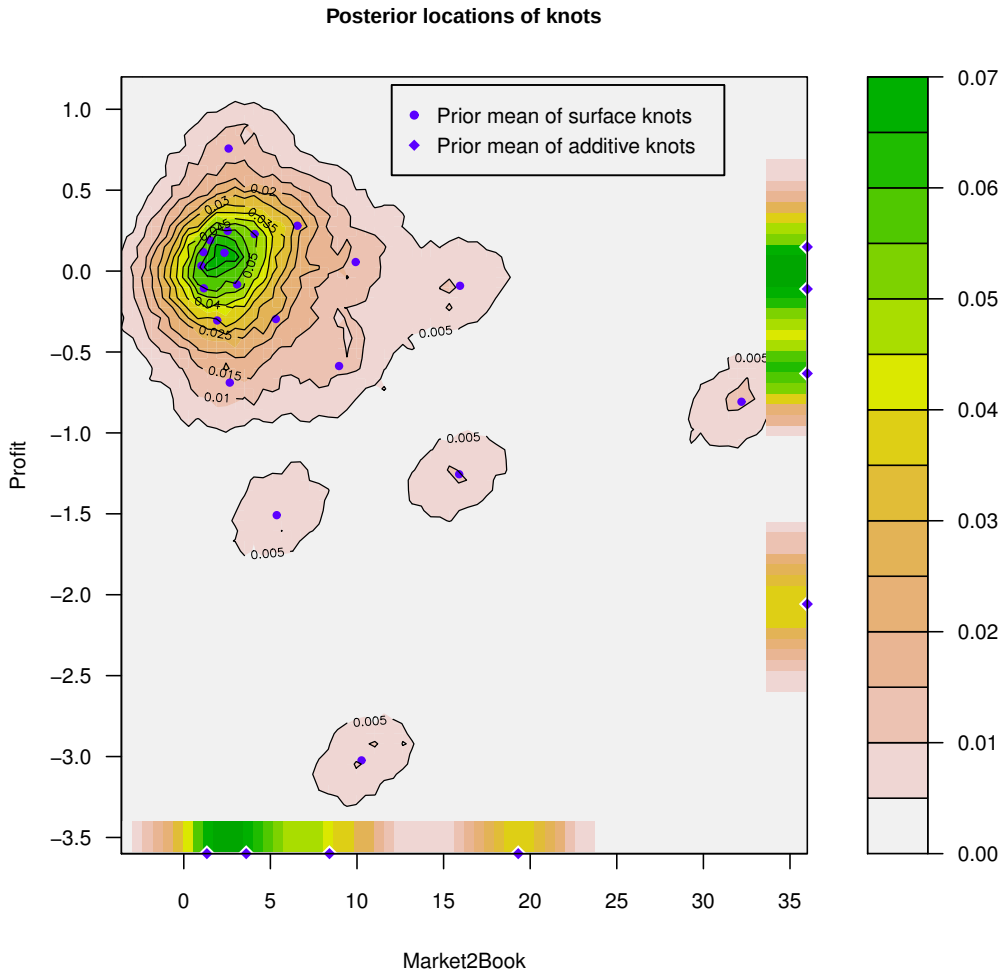

FIGURE 6. Heat map to visualize the posterior density of the knot locations in covariate space for model with 4 free additive knots and 20 free surface knots for the firm leverage dataset. The plot is constructed by partitioning the covariate space into 70×70 rectangular regions and counting the number of surface knots in each rectangle over the MCMC draws. The posterior density of the locations of the additive knots is constructed in a similar fashion and separate heat maps for the additive knots in each covariate are shown just above the horizontal axis and vertical axis, respectively.

Dimatteo et al. (2001) and Gulam Razul et al. (2003). Note that we are not studying the performance of a complete RJMCMC scheme; we are here interested in isolating this particular updating step and comparing it to our tailored proposal. We use the inefficiency factor (IF) (Geweke, 1992) to measure the efficiency of MCMC. The IF is a measure of the number of draws needed to obtain the equivalent of a single independent draw. It is defined as $\text{IF} = 1 + 2\sum_{i=1}^{\infty} \rho_i$ where ρ_i is the

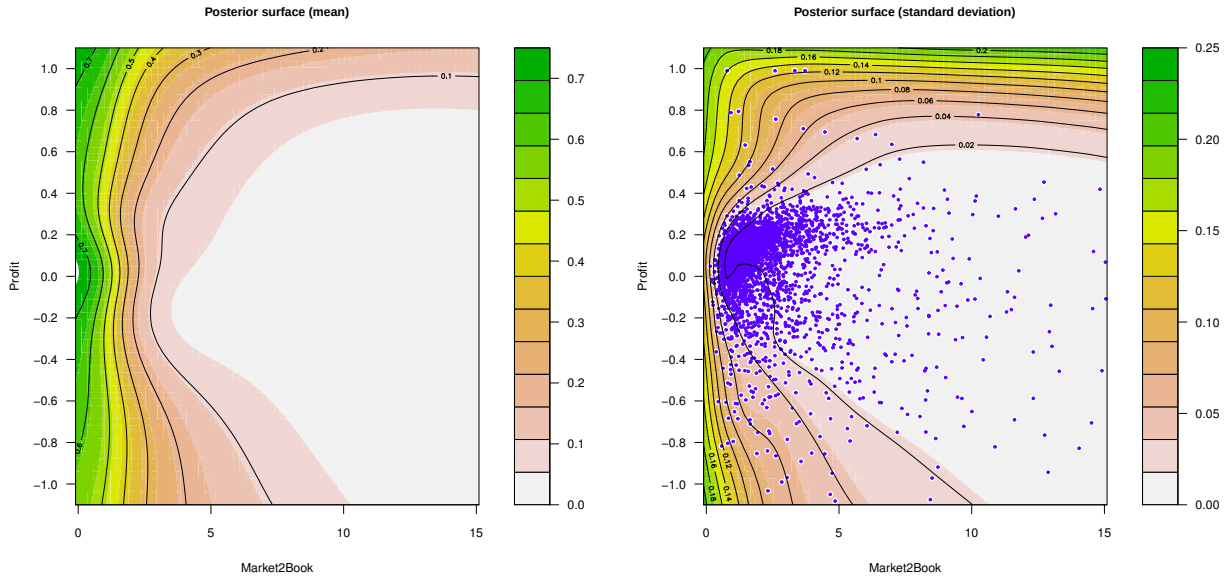

FIGURE 7. The posterior mean (left) and standard deviation (right) of the posterior surface for the model with 4 free additive knots and 20 free surface knots for firm leverage data. The subplot to the right also shows an overlay of the covariate observations.

TABLE 3. Comparison of algorithms for updating the knot locations in a model with 20 free surface knots and 4 additive knots in each covariate. Firm leverage data.

	SRWM	SMH	BMH
Mean IF for the posterior mean surface	29.63	2.70	1.16
Mean acceptance probability	0.26	0.62	0.88
Computing time (*min*)	388.21	1716.07	141.72
Effective sample size per minute	0.87	2.16	60.83

autocorrelation of the MCMC trajectory at lag i. We also document the effective sample size per minute, i.e. $(\text{number of MCMC draws})/(\text{IF} \times \text{computing time})$ to measure the overall efficiency of the MCMC.

Table 3 shows the efficiency of the three knot sampling algorithms in a model with 20 free surface knots and 4 additive knots in each covariate on the firm leverage data. The inefficiency factor in Table 3 is the average inefficiency of the posterior mean surface in 1000 random chosen points in covariate space. There is some gain from tailoring the proposal for each knot separately, but the really striking observation from Table 3 is the massive efficiency and speed gains from updating all the blocks jointly using a tailored proposal; the effective sample size per minute is roughly 70 times larger when our BMH algorithm is used instead of simple SRWM updates.

6. Concluding remarks

We have presented a general Bayesian approach for fitting a flexible surface model for a continuous multivariate response using a radial basis spline with freely estimated knot locations. Our approach uses shrinkage priors to avoid overfitting. The locations of the knots and the shrinkage parameters are treated as unknown parameters and we propose a highly efficient MCMC algorithm for these parameters with the coefficients of the multivariate spline integrated out analytically. An important feature of our algorithm is that all knot locations are sampled jointly using a Metropolis-Hastings proposal density tailored to the conditional posterior, rather than the one-knot-at-a-time random walk proposals used in previous literature. The same applies to the block of shrinkage parameters. Both a simulation study and a real application on firm leverage data show that models with free knots have a better out-of-sample predictive performance than models with fixed knots. Moreover, the free knots model is also more robust in the sense that it performs consistently well across different datasets. We also found that models that mix surface and additive spline basis functions in the same model perform better than models with only one of the two basis types.

Our approach can be directly used with other splines basis functions, other priors, and it is at least in principle straightforward to augment the model with Bayesian variable selection. We are currently working on removing the assumption of Gaussian error distribution by using a Dirichlet process mixture (DPM) prior on the model disturbances.

7. Acknowledgements

The authors are grateful to Paolo Giordani and Robert Kohn for stimulating discussions and constructive suggestions. The authors thank two anonymous referees for the helpful comments that improved the contents and presentation of the paper. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Appendix A. Details of the MCMC algorithm

In this section we briefly address the MCMC details and related computational issues. For details on matrix manipulations and derivatives, see e.g. Lütkepohl (1996). Our MCMC algorithm only requires the gradient of the conditional posteriors w.r.t. each parameter. Since users can always use their own prior on the knots and shrinkages, we will not document the gradient of any particular prior. In particular for the normal prior, one can directly find the results in e.g. Mardia & Kent (1979). We now present the full gradients for the knot locations and the shrinkage parameters.

A.1. **Gradient w.r.t. the knot locations.**

$$\begin{aligned}\frac{\partial \ln p\left(\boldsymbol{\xi}|\boldsymbol{\lambda},\Sigma,\boldsymbol{Y},\boldsymbol{X}\right)}{\partial \boldsymbol{\xi}'} &= \frac{\partial \log p(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}'} - \frac{p}{2}\sum_{i\in\{o,a,s\}}(\mathrm{vec}\boldsymbol{P}_i)'\frac{\partial \mathrm{vec}\boldsymbol{P}_i}{\partial \boldsymbol{\xi}'} \\ &- \left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)'\Sigma_{\boldsymbol{\beta}}^{-1}\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\xi}'} - \frac{1}{2}\left(\mathrm{vec}\Sigma_{\tilde{\boldsymbol{\beta}}}\right)'\frac{\partial \mathrm{vec}[\Sigma^{-1}\otimes \boldsymbol{X}'\boldsymbol{X}]}{\partial \boldsymbol{\xi}'} \\ &- \frac{1}{2}\left(\mathrm{vec}\Sigma^{-1}\right)'(\boldsymbol{I}_p+\boldsymbol{K}_{p,p})\left\{(\boldsymbol{I}_p\otimes\tilde{\boldsymbol{E}}'\boldsymbol{X})\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\xi}'}+(\tilde{\boldsymbol{B}}'\otimes\tilde{\boldsymbol{E}}')\frac{\partial \mathrm{vec}\boldsymbol{X}}{\partial \boldsymbol{\xi}'}\right\} \\ &- \frac{1}{2}\left\{\mathrm{vec}\left[\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)'+\Sigma_{\tilde{\boldsymbol{\beta}}}\right]\right\}'\frac{\partial \mathrm{vec}\Sigma_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\xi}'},\end{aligned}$$

where $\tilde{\boldsymbol{E}}=\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{B}}$, $\boldsymbol{I}_p$ is the identity matrix, $\boldsymbol{K}_{p,p}$ is the commutation matrix and

$$\frac{\partial \mathrm{vec}[\Sigma^{-1}\otimes \boldsymbol{X}'\boldsymbol{X}]}{\partial \boldsymbol{\xi}'} = \left(\boldsymbol{I}_p\otimes\boldsymbol{K}_{q,p}\otimes\boldsymbol{I}_q\right)\left(\mathrm{vec}\Sigma^{-1}\otimes\boldsymbol{I}_{q^2}\right)\left(\boldsymbol{I}_{q^2}+\boldsymbol{K}_{q,q}\right)\left(\boldsymbol{I}_q\otimes\boldsymbol{X}'\right)\frac{\partial \mathrm{vec}\boldsymbol{X}}{\partial \boldsymbol{\xi}'},$$

$$\begin{aligned}\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\xi}'} &= \Sigma_{\tilde{\boldsymbol{\beta}}}\left[\left[\Sigma^{-1}\boldsymbol{Y}'\otimes\boldsymbol{I}_q\right]\boldsymbol{K}_{n,q}\frac{\partial \mathrm{vec}\boldsymbol{X}}{\partial \boldsymbol{\xi}'}+\left(\boldsymbol{\mu}'\otimes\boldsymbol{I}_{pq}\right)\frac{\partial \mathrm{vec}\Sigma_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\xi}'}\right] \\ &- \left[\left\{\left[\mathrm{vec}\left(\boldsymbol{X}'\boldsymbol{Y}\Sigma^{-1}\right)+\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}\right]'\Sigma_{\tilde{\boldsymbol{\beta}}}\right\}\otimes\Sigma_{\tilde{\boldsymbol{\beta}}}\right]\left[\frac{\partial \mathrm{vec}[\Sigma^{-1}\otimes\boldsymbol{X}'\boldsymbol{X}]}{\partial \boldsymbol{\xi}'}+\frac{\partial \mathrm{vec}\Sigma_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\xi}'}\right].\end{aligned}$$

We can decompose the gradient for the design matrix w.r.t the knots as

$$\frac{\partial \mathrm{vec}\boldsymbol{X}}{\partial \boldsymbol{\xi}'} = \begin{bmatrix} \boldsymbol{0}_{(nq_o\times l_s)} & \boldsymbol{0}_{(nq_o\times l_a)} \\ \partial \mathrm{vec}\boldsymbol{X}_s/\partial \mathrm{vec}(\boldsymbol{\xi}_s')' & \boldsymbol{0}_{(nq_s\times l_a)} \\ \boldsymbol{0}_{(nq_a\times l_s)} & \partial \mathrm{vec}\boldsymbol{X}_a/\partial {\boldsymbol{\xi}_a}' \end{bmatrix}$$

where l_s and l_a are numbers of parameters in the knots locations for surface and additive component, respectively. This decomposition makes user-specified basis functions for different components possible and one may update the locations in a parallel mode (efficient for small models) or batched mode (for models with many parameters). In particular for the thin-plate spline, we have

$$\frac{\partial \mathrm{vec}\boldsymbol{X}_i}{\partial {\boldsymbol{\xi}_i}'} = -\begin{bmatrix} (1+2\ln\|\boldsymbol{x}_i-\boldsymbol{\xi}_{ij}\|)(\boldsymbol{x}_i-\boldsymbol{\xi}_{ij}) & & \\ & \ddots & \\ & & (1+2\ln\|\boldsymbol{x}_i-\boldsymbol{\xi}_{ij}\|)(\boldsymbol{x}_i-\boldsymbol{\xi}_{ij}) \end{bmatrix}_{\substack{i\in\{a,s\},\\ j\in\{1,\ldots,q_i\}.}}$$

Note that the gradient can be obtained efficiently by applying Lemma 1 and Algorithm 2 in Section A.3 below whenever $\partial \mathrm{vec}\Sigma_{\boldsymbol{\beta}}^{-1}/\partial \boldsymbol{\xi}'$ and the commutation matrix appear.

A.2. **Gradient w.r.t. the shrinkage parameters.**

$$\begin{aligned}\frac{\partial \ln p(\boldsymbol{\lambda}|\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{Y},\boldsymbol{X})}{\partial \boldsymbol{\lambda}'} = \frac{\partial \log p(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}'} - \frac{1}{2}\left[q_o\boldsymbol{\lambda}_o{}', q_s\boldsymbol{\lambda}_s{}', q_a\boldsymbol{\lambda}_a{}'\right] - \left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} \\ - \frac{1}{2}\left(\mathrm{vec}\boldsymbol{\Sigma}^{-1}\right)'\left(\boldsymbol{I}_p + \boldsymbol{K}_{p,p}\right)\left(\boldsymbol{I}_p \otimes \tilde{\boldsymbol{E}}'\boldsymbol{X}\right)\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} \\ - \frac{1}{2}\mathrm{vec}\left[\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)' + \Sigma_{\tilde{\boldsymbol{\beta}}}\right]'\frac{\partial \mathrm{vec}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\lambda}'},\end{aligned}$$

where

$$\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\lambda}'} = \left\{\left[\left(\mathrm{vec}\left(\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\right) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right] \otimes \boldsymbol{\Sigma}_{\boldsymbol{\beta}} - \boldsymbol{\mu}' \otimes \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}\right\}\frac{\partial \mathrm{vec}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\lambda}'},$$

and $\partial \mathrm{vec}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}/\partial \boldsymbol{\lambda}'$ can be obtained efficiently by applying Lemma 1 in Section A.3 and by

$$\begin{aligned}\frac{\partial \mathrm{vec}[(\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i^{-1/2}) \otimes \boldsymbol{P}_i]}{\partial \boldsymbol{\lambda}_i{}'} = &\left(\boldsymbol{I}_p \otimes \boldsymbol{K}_{q_i,p} \otimes \boldsymbol{I}_{q_i}\right)\left(\boldsymbol{I}_{p^2} \otimes \mathrm{vec}\boldsymbol{P}_i\right)\left(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{p,p}\right) \\ &\times \left(\boldsymbol{I}_p \otimes [\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\Sigma}^{-1}]\right)\frac{\partial \mathrm{vec}\boldsymbol{\Lambda}_i^{-1/2}}{\partial \boldsymbol{\lambda}_i'},\ i \in \{a,s\}.\end{aligned}$$

where $\partial \mathrm{vec}\boldsymbol{\Lambda}_i/\partial \boldsymbol{\lambda}_i'$ is $p^2 \times p$ matrix with elements $\nabla_{j(p+1)-p,\ j} = -1/2\lambda_{i,j}^{-3/2}$ for $j = 1,...,p$ and zero elsewhere.

A.3. **Computational remarks.** The computational implementation of gradients in Section A.1 and Section A.2 is straightforward but the sparsity of some of the matrices can be exploited in moderate to large datasets. We now present a lemma and an algorithm that can dramatically speed up the computations. It is convenient to define $\boldsymbol{A}(\boldsymbol{i},:)$ and $\boldsymbol{A}(:,\boldsymbol{j})$ as matrix operations that reorders the rows and columns of matrix $\boldsymbol{A}$ with indices $\boldsymbol{i}$ and $\boldsymbol{j}$. Therefore, $\boldsymbol{\beta} = \boldsymbol{b}(\boldsymbol{c},:)$, $\boldsymbol{\mu} = \boldsymbol{\mu}^*(\boldsymbol{c},:)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{b}}(\boldsymbol{c},\boldsymbol{c})$ for proper indices $\boldsymbol{c}$, and $|\boldsymbol{\Sigma}_{\boldsymbol{b}}| = |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|$ since permuting two rows or columns changes the sign but not the magnitude of the determinant.

Lemma 1. *Given matrix $\boldsymbol{C}$ and the indexing vector $\boldsymbol{z}$ such that $(\mathrm{vec}\boldsymbol{\Sigma}_{\boldsymbol{b}})(\boldsymbol{z},:) = \mathrm{vec}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ holds, we can decompose the following gradient as*

$$\boldsymbol{C}\frac{\partial \mathrm{vec}[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}'} = \left[\boldsymbol{C}_s\frac{\partial \mathrm{vec}[(\boldsymbol{\Lambda}_s^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_s^{-1/2}) \otimes \boldsymbol{P}_s]}{\partial \boldsymbol{\theta}_s{}'},\ \boldsymbol{C}_a\frac{\partial \mathrm{vec}[(\boldsymbol{\Lambda}_a^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_a^{-1/2}) \otimes \boldsymbol{P}_a]}{\partial \boldsymbol{\theta}_a{}'}\right]$$

where $\boldsymbol{\theta}$ is any parameter vector of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{C}_s = \{[\boldsymbol{C}(:,\boldsymbol{z})](:,\boldsymbol{h}_s)\}(:,\boldsymbol{z}_s \neq 0)$, $\boldsymbol{h}_s = [(p^2qq_o+1),(p^2qq_o+2),...,p^2q(q_o+q_s)]'$, $\boldsymbol{z}_s = \mathrm{vec}([\boldsymbol{0}_{pq_s\times pq_o},\ \boldsymbol{1}_{pq_s\times pq_s},\ \boldsymbol{0}_{pq_s\times pq_a}]')$, $\boldsymbol{C}_a = \{[\boldsymbol{C}(:,\boldsymbol{z})](:,\boldsymbol{h}_a)\}(:,\boldsymbol{z}_a \neq 0)$, $\boldsymbol{h}_a = [(p^2q(q_o+q_s)+1),(p^2q(q_o+q_s)+2),...,p^2q^2]'$ and $\boldsymbol{z}_a = \mathrm{vec}([\boldsymbol{0}_{pq_a\times p(q_o+q_s)},\ \boldsymbol{1}_{pq_a\times pq_a}]')$.

Algorithm 2. *An efficient algorithm to calculate $\boldsymbol{K}_{m,n}\boldsymbol{Q}$ (or $\boldsymbol{Q}\boldsymbol{K}_{m,n}$) where $\boldsymbol{K}_{m,n}$ is the commutation matrix and $\boldsymbol{Q}$ is any dense matrix that is conformable to $\boldsymbol{K}_{m,n}$.*

(1) *Create an $m \times n$ (or $n \times m$) matrix $\boldsymbol{T}$ and fill it by columns with the sequence $\{1,2,...,nm\}$.*

(2) *Obtain the indexing vector* $\boldsymbol{t} = \text{vec}(\boldsymbol{T}')$.
(3) *Return* $\boldsymbol{Q}(\boldsymbol{t},:)$ *(or* $\boldsymbol{Q}(:,\boldsymbol{t})$*).*

APPENDIX B. PRIOR ROBUSTNESS

This section explores the sensitivity of the posterior inferences with respect to variations in the prior. There are clearly many aspects of the prior to explore, but we will here focus on the sensitivity with respect to the prior on the shrinkage factors and the prior on the knot locations, which are the most influential priors for the model. Since our model is very flexible and richly parametrized it is natural to expect, or even desireable, that the posterior responds to variations in the prior hyperparameters. But since the prior in complex models is always hard to specify, it is hoped that moderate changes in the prior should at least not overturn the posterior inferences.

B.1. **The prior on the shrinkage parameters.** Figure 8 displays the posterior sensitivity of the knot locations, the posterior mean and standard deviation surfaces to changes in the prior variance on the shrinkage factors. The posterior and predictive results are clearly very robust to changes in this particular aspect of the prior.

B.2. **The prior on the knot locations.** Figure 9 displays the effect on the posterior knot density from changes in both the mean (columns) and the variance (rows) of prior on the knot locations. While there are some differences in the posterior knot densities when the prior variance changes (changes across rows), there is much larger difference in the posterior of the knots when the prior mean of the knot locations change. This is partly explained by fact that the differences between the three ways of placing the prior means are rather dramatic, but it is clear that the prior mean of the knot locations are affecting where the knots are located a posteriori. Considering the complexity in the inference on the knot locations and the fact that many of the knots probably correspond to regression coefficients that are close to zero, this is perhaps not too surprising.

The posterior inference of the knot locations is typically not of interest. What matters is the inferences on the conditional predictive distribution $p(\mathbf{y}|\mathbf{x})$. Figure 10 and 11 investigate the sensitivity of the posterior mean and standard deviation surfaces to changes in the prior mean and variance of the knot locations. Here the robustness to variations in the prior is much larger. Both the predictive mean and the predictive standard deviation remain largely unchanged, considering the magnitude of the changes in the prior. The main differences in the prior mean surface occur in points of covariate space where the uncertainty in the predictive mean is large.

APPENDIX C. FURTHER SIMULATION RESULTS

C.1. **Additional results from the simulation study in Section 4 of the paper.** This section documents the simulation results for the simulation setup with $p = 2$ and $n = 1000$. The simulation setup is identical to the one in Section 4 of the paper with the exception that number of data points is increased from $n = 200$ to $n = 1000$. Figure 12 compares the estimation loss from using fixed and free knots, respectively. Figure 13 compares the out-of-sample predictive residuals from a models with 15 fixed surface knots to a model with 5 free knots, and Figure 14 does the same type of comparison for a model with 20 fixed surface knots to a model with 10 free knots.

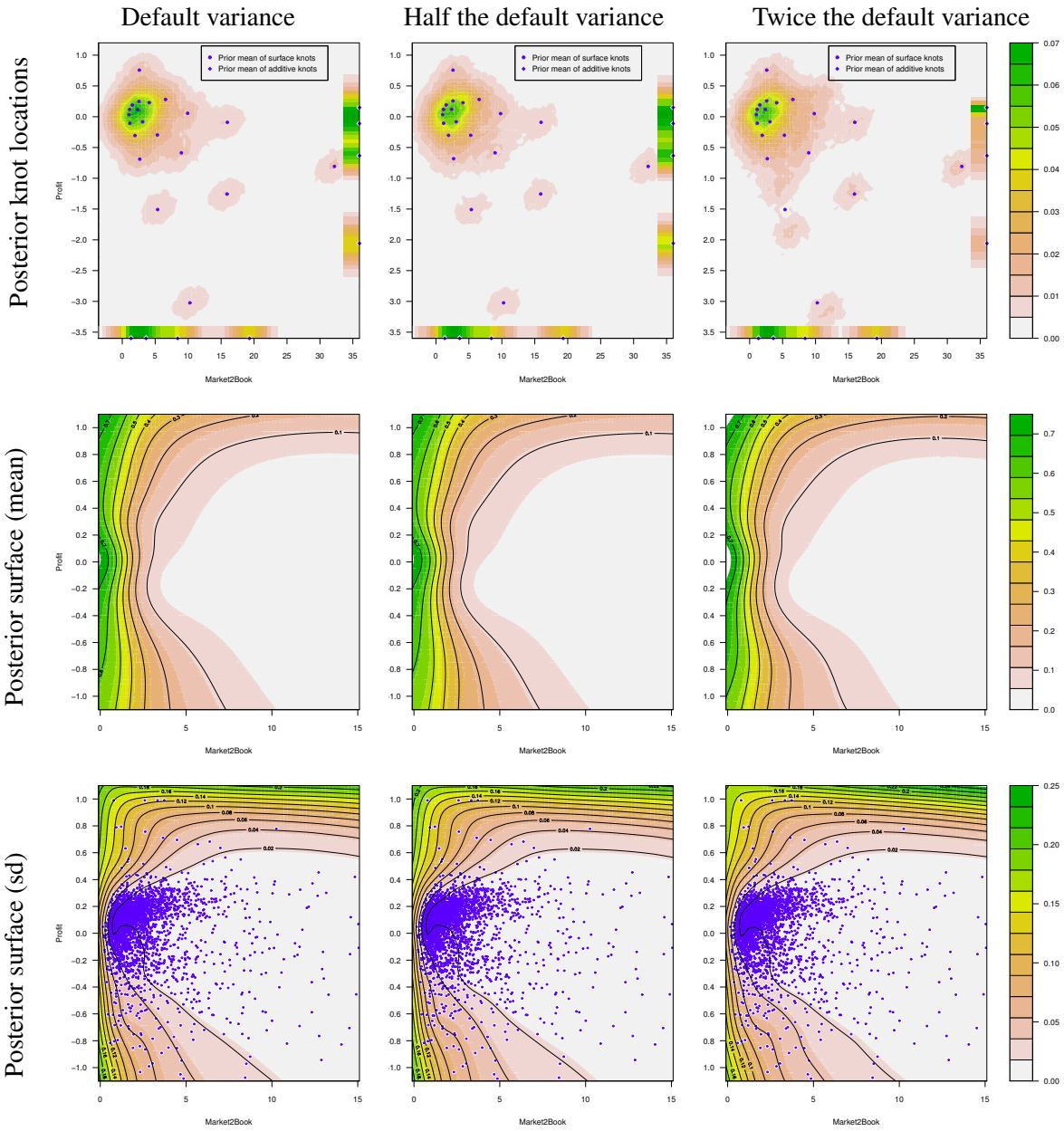

FIGURE 8. Posterior sensitivity with respect to changes in the prior variance of the shrinkage factors. The first column shows the posterior of the knot density (top row), the posterior mean surface (middle row) and the posterior standard deviation surface (bottom row) using the default prior in the paper. The second and third columns demonstrates the effect on the posterior inferences when the prior variance is half of the default value (second column) and twice the default value (third column).

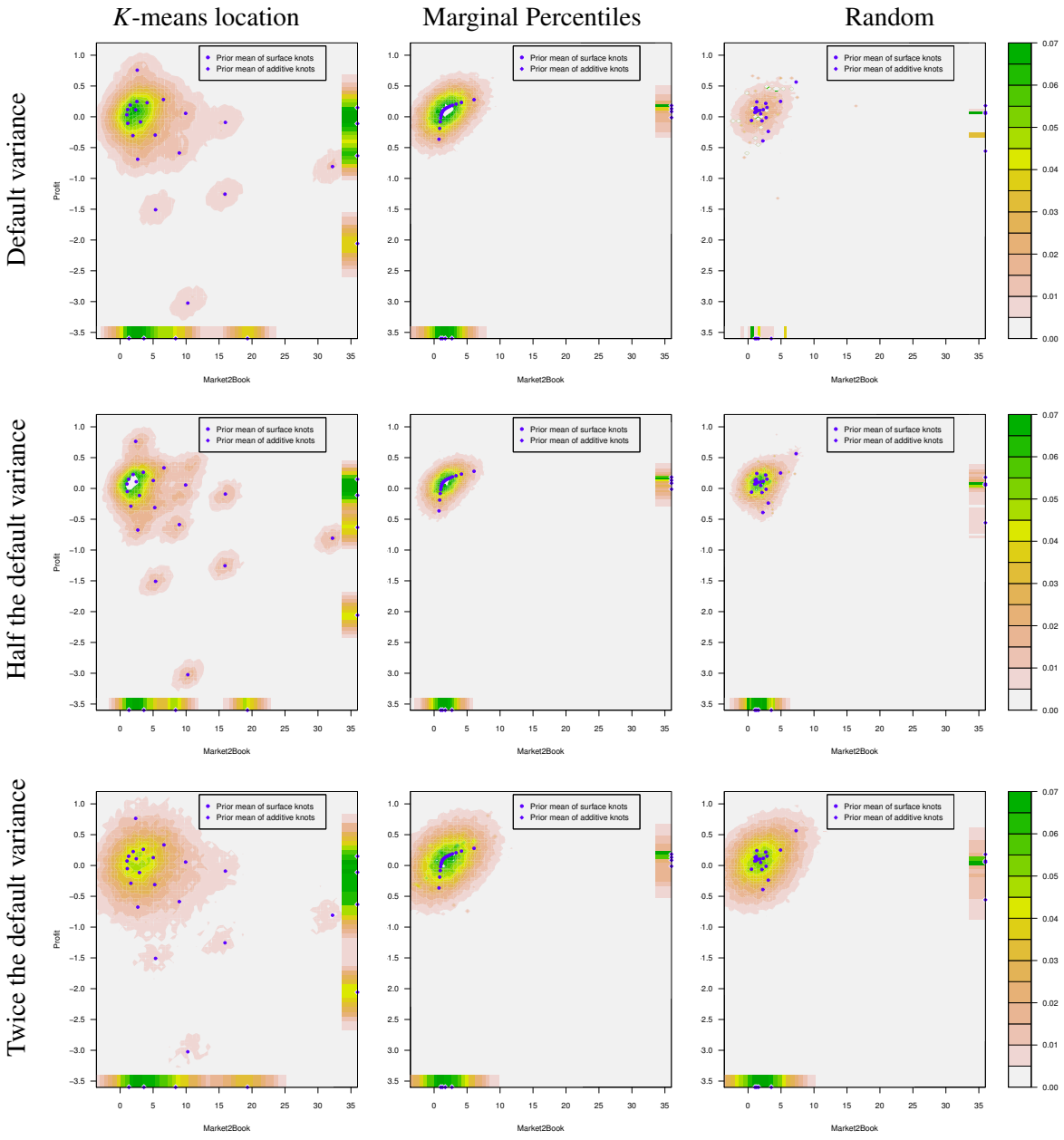

FIGURE 9. Sensitivity of the posterior knot density with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

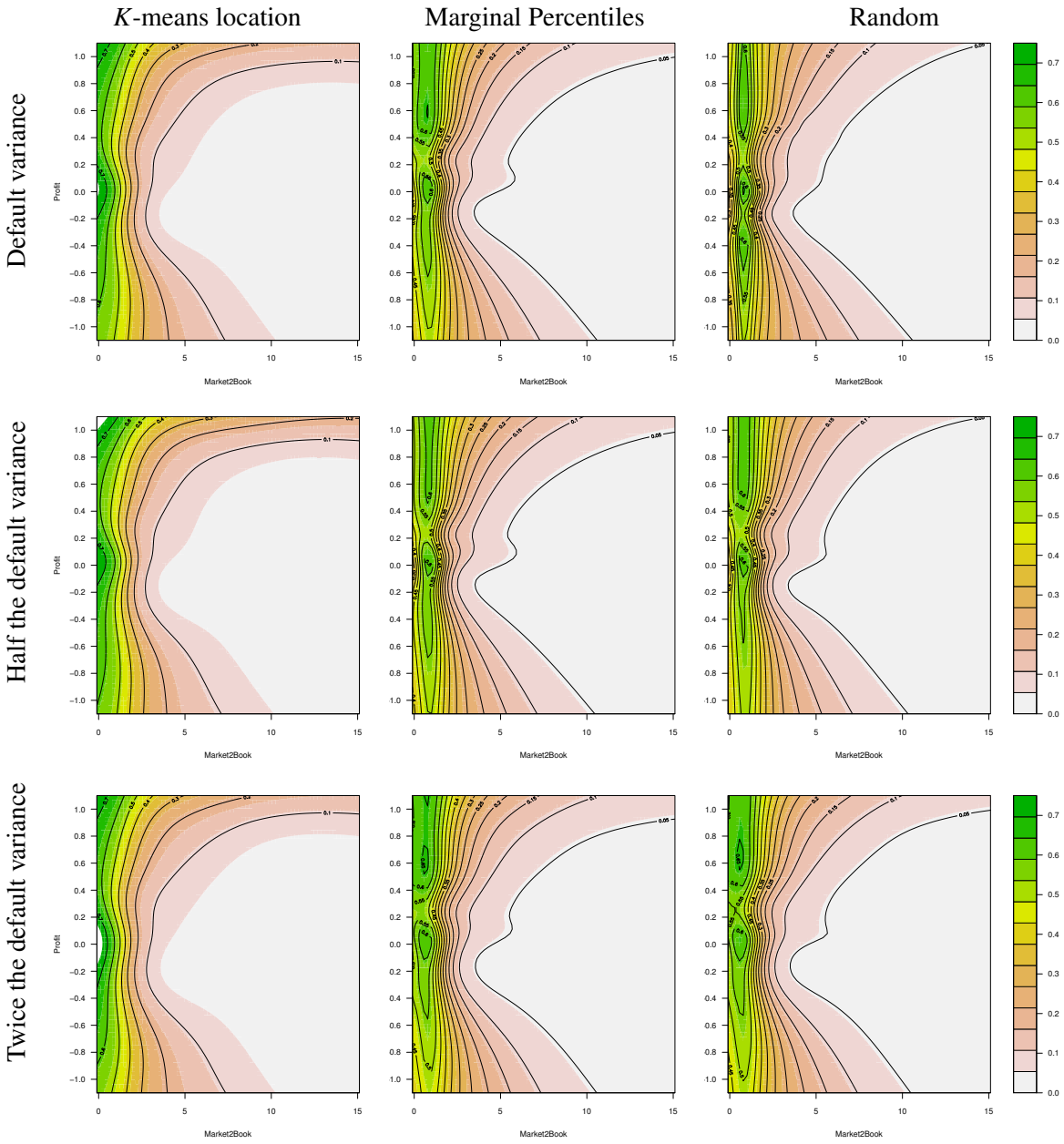

FIGURE 10. Sensitivity of the posterior mean surface with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

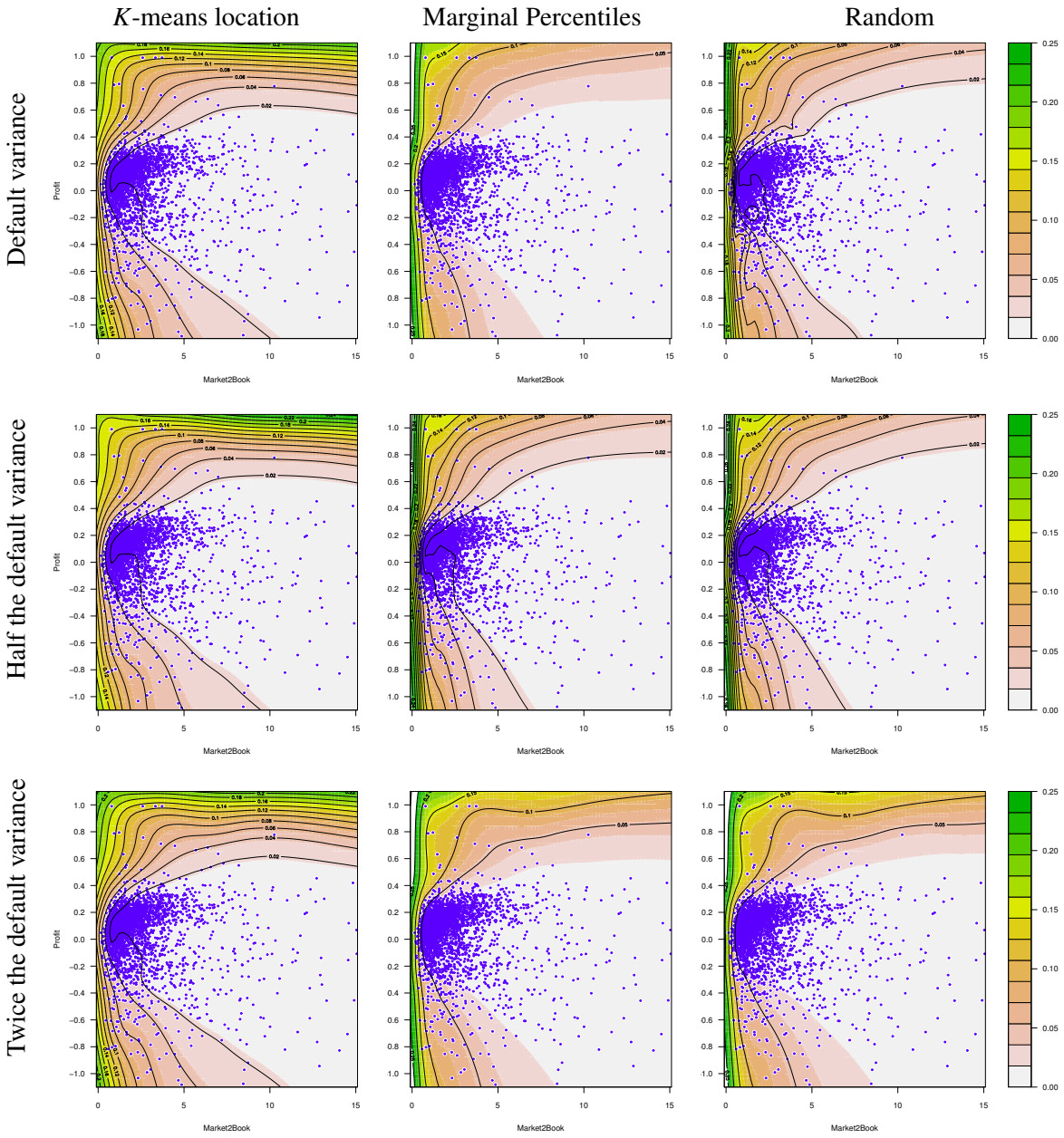

FIGURE 11. Sensitivity of the posterior standard deviation surface with respect to changes in the mean (columns) and variance (rows) in the prior distribution of the knot locations. The prior mean of the locations in the second columns is chosen from the empirical marginal distribution of each covariate, and the prior mean in the third column are random draws without replacement among the data points.

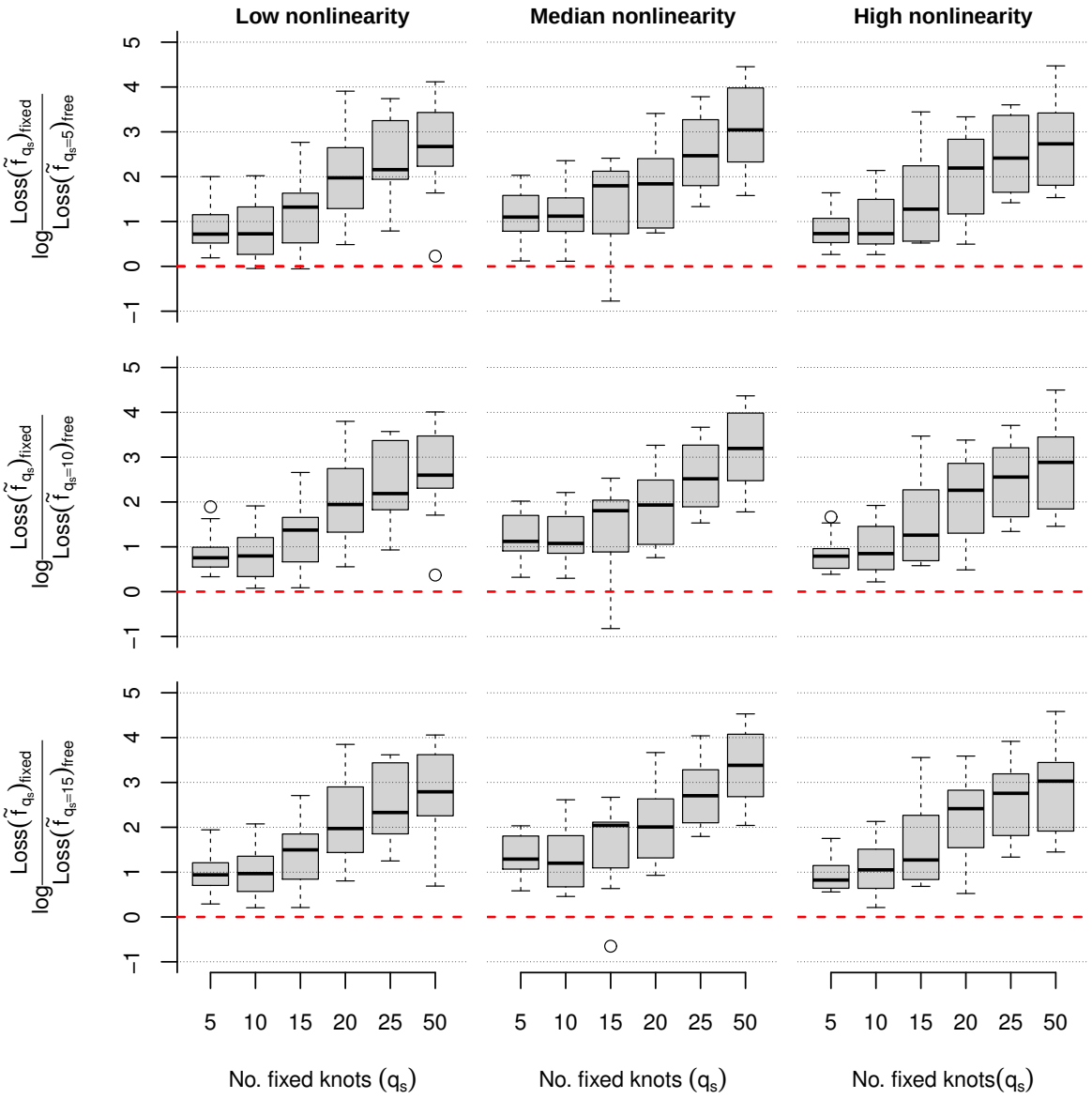

FIGURE 12. Boxplot of the log loss ratio comparing the performance of the fixed knots model with the free knots model for the DGP with $p = 2$ and $n = 1000$. The three columns of the figure correspond to different degrees of nonlinearity of the realized datasets, as measured by estimated DNL in (5) in the paper.

C.2. **Simulation results from a situation where none of the two models are correct.** In this section, we briefly describe a simple simulation example where the true model is not nested in any

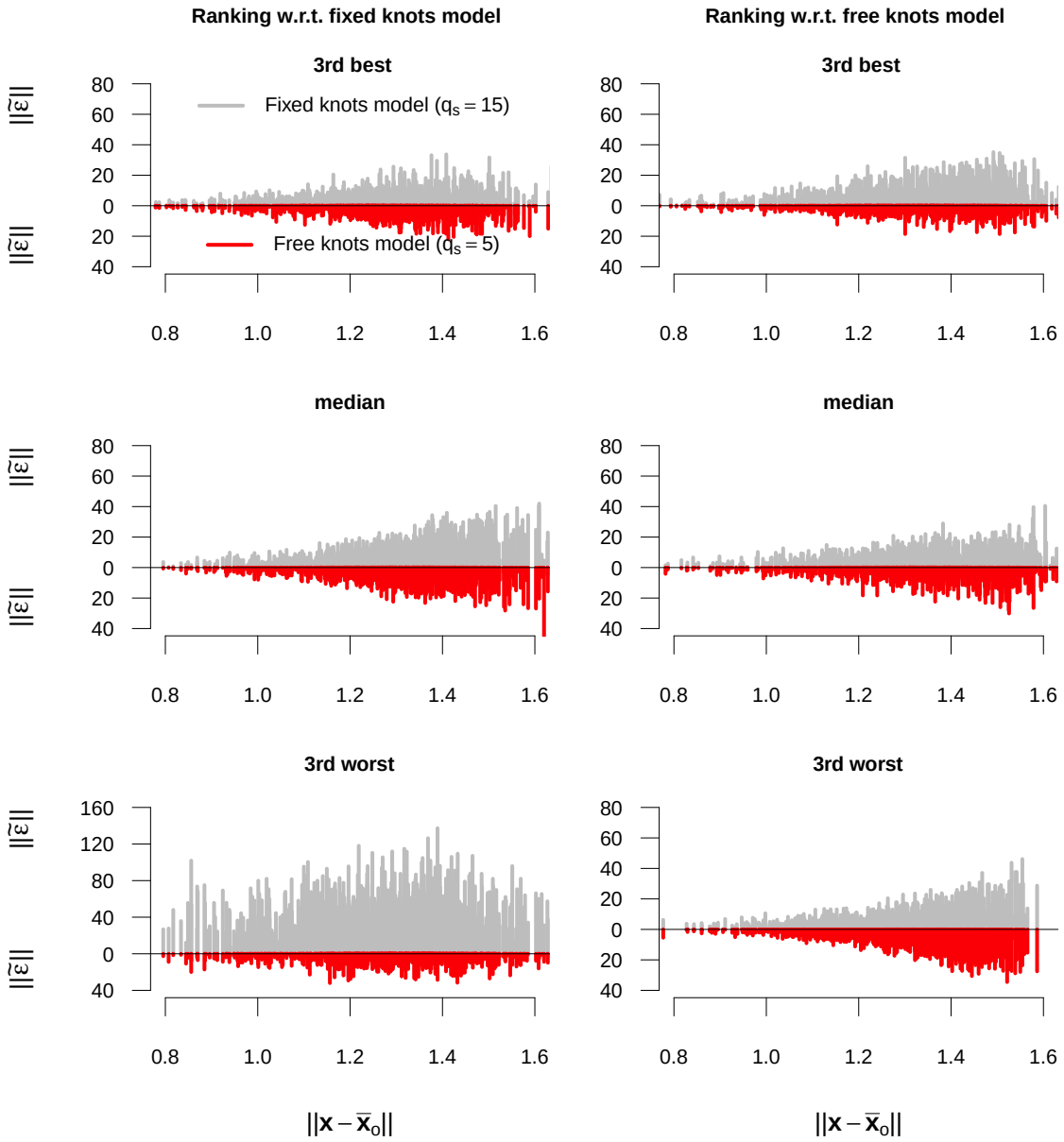

FIGURE 13. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 1000$. The lines in each subplot are the normed residuals from the model with 15 fixed surface knots (vertical bars above the zero line), and the model with 5 free knots (vertical bars below the zero line). The column to the left shows the results for three datasets chosen when performance is ranked according to the fixed knots model, and the right column displays the results for three datasets chosen when performance is ranked according to the free knots model.

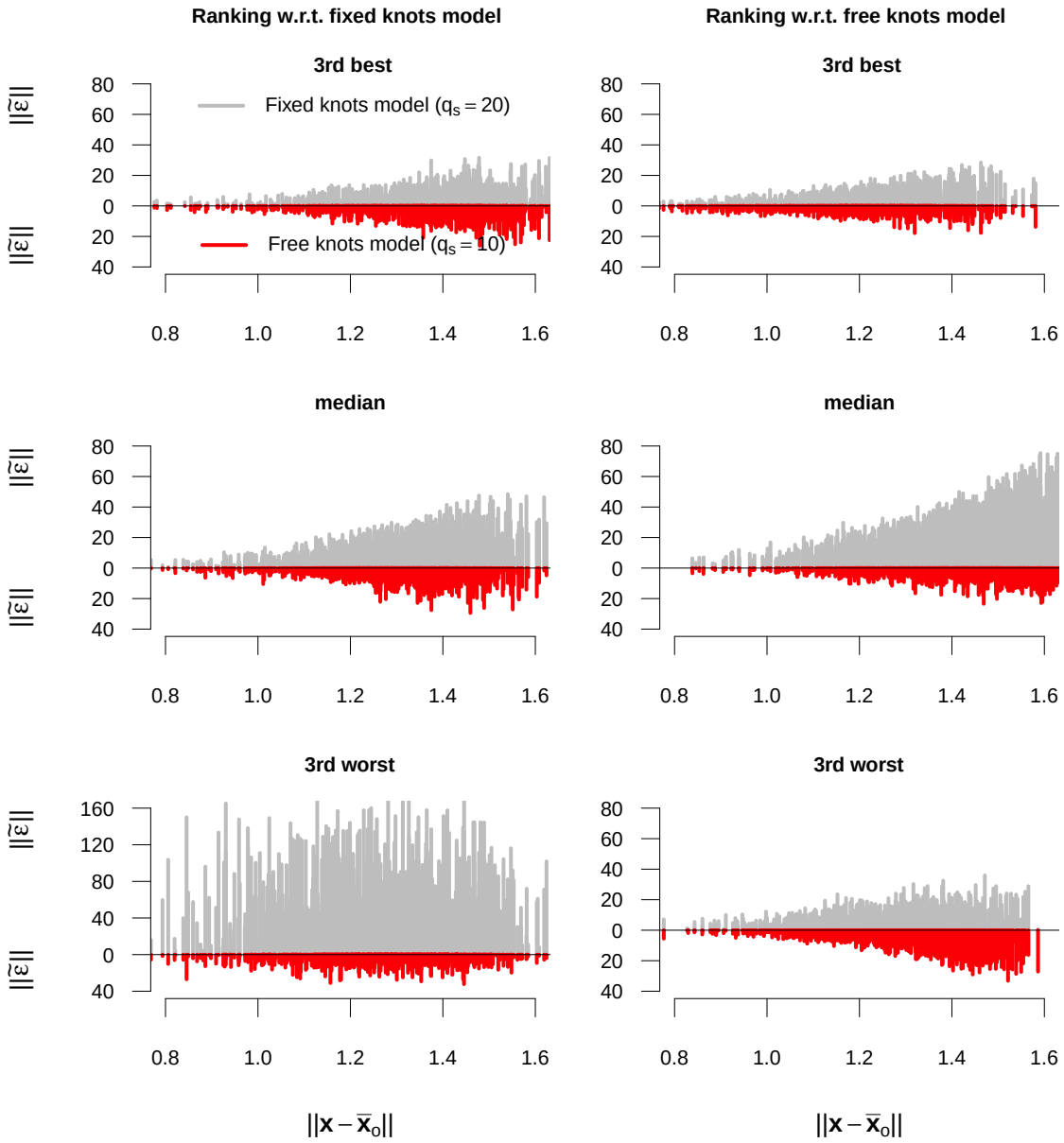

FIGURE 14. Plotting the norm of the predictive multivariate residuals as a function of the distance between the covariate vector and its sample mean. The results are for the DGP with $p = 2$ and $n = 1000$. The lines in each subplot are the normed residuals from the model with 20 fixed surface knots (vertical bars above the zero line), and the model with 10 free knots (vertical bars below the zero line). The column to the left shows the results for three datasets chosen when performance is ranked according to the fixed knots model, and the right column displays the results for three datasets chosen when performance is ranked according to the free knots model.

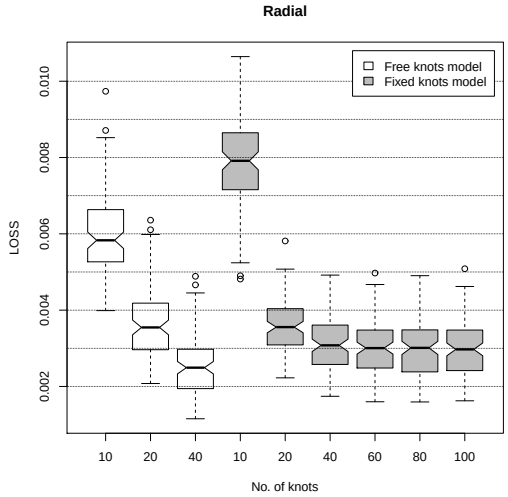

FIGURE 15. Boxplots of the loss in the simulations for the radial mean function.

of the two estimated models. We generate Gaussian data around the following mean surface

$$\begin{aligned} y = 42.659[0.1 + (x_1 - 0.5)(0.05 + (x_1 - 0.5)^4 \\ - 10(x_1 - 0.5)^2(x_2 - 0.5)^2 + 5(x_2 - 0.5)^2)] \end{aligned} \tag{7}$$

The function in Equation (7) is called a radial function. We generate 100 datasets using $N(0, 0.1)$ disturbances around the mean. The number of observations are 1000 in each dataset. We use linear and surface components. The number of knots used in the free knot models is 10, 20, and 40, and the number of knots used in the fixed knots model is 10, 20, 40, 60, 80 and 100.

Figure 15 displays boxplots of the losses for the different number of knots in each model. The free knots model outperforms the fixed knots model, but the improvement from using free knots are not that large here since the covariate space is only two-dimensional, which is small enough for the fixed knots to have a decent coverage.

REFERENCES

BASTOS, J. & RAMALHO, J. (2010). Nonparametric models of financial leverage decisions. *CEMAPRE Working Papers* Available at: `http://cemapre.iseg.utl.pt/archive/preprints/426.pdf`.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. & STONE, C. (1984). *Classification and regression trees*. Chapman and Hall/CRC, New York.

BUHMANN, M. D. (2003). *Radial basis functions: theory and implementations*. Cambridge University Press, Cambridge.

CHIPMAN, H. A., GEORGE, E. I. & MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298.

DENISON, D., HOLMES, C. C., MALLICK, B. K. & SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Jone Wiley & Sons, Chichester.

DENISON, D., MALLICK, B. & SMITH, A. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **60**, 333–350.

DIMATTEO, I., GENOVESE, C. & KASS, R. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055.

FRIEDMAN, J. (1991). Multivariate adaptive regression splines. *The annals of statistics* **19**, 1–67.

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

GASSER, T. (1979). Kernel estimation of regression functions. *Smoothing techniques for curve estimation* , 23–68.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. David & A. F. M. Smith, eds. Oxford University Press, Oxford, pp. 169–193.

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

GEWEKE, J. & AMISANO, G. (2011). Optimal prediction pools. *Journal of Econometrics* **164**, 130–141.

GIORDANI, P., JACOBSON, T., VILLANI, M. & VON SCHEDVIN, E. (in press). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis* .

GULAM RAZUL, S., FITZGERALD, W. & ANDRIEU, C. (2003). Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **497**, 492–510.

HASTIE, T. & TIBSHIRANI, R. (1986). Generalized additive models. *Statistical science* , 297–310.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

HOLMES, C. C. & MALLICK, B. K. (2003). Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines. *Journal of the American Statistical Association* **98**, 352–368.

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

KHATRI, C. & RAO, C. (1968). Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: The Indian Journal of Statistics, Series A* **30**, 167–180.

LÜTKEPOHL, H. (1996). *Handbook of matrices*. John Wiley & Sons, Chichester.

MARDIA, K. & KENT, J. (1979). *Multivariate analysis*. Academic Press, London.

NADARAYA, E. (1964). On estimating regression. *Theory of Probability & Its Applications* **9**, 141–142.

NOTT, D. J. & LEONTE, D. (2004). Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics* **13**, 362–382.

RAJAN, R. & ZINGALES, L. (1995). What do we know about capital structure? Some evidence from international data. *The Journal of Finance* **50**, 1421–1460.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **59**, 731–792.

RUPPERT, D., WAND, M. & CARROLL, R. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

SMITH, M. & KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* , 267–288.

VILLANI, M., KOHN, R. & NOTT, D. J. (2012). Generalized smooth finite mixtures. *Journal of Econometrics* **171**, 121–133.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyà: The Indian Journal of Statistics, Series A* **26**, 359–372.

WOOD, S., JIANG, W. & TANNER, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513.

ZELLNER, A. (1971). *An introduction to Bayesian inference in econometrics*. John Wiley & Sons, New York.

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

Manuscript (2013)

MODELING COVARIATE-CONTINGENT CORRELATION AND TAIL-DEPENDENCE WITH COPULAS

FENG LI

ABSTRACT. Copulas provide an attractive approach for constructing multivariate densities with flexible marginal distributions and different forms of dependence. Of particular importance in many areas is the possibility of explicitly modeling tail-dependence. Most of the available approaches estimate tail-dependence and correlations via nuisance parameters, yielding results that are neither tractable nor interpretable for practitioners. We propose a general Bayesian approach for directly modeling tail-dependence and correlations as explicit functions of covariates. Our method allows for variable selection among the covariates in the marginal models and in the copula parameters. Posterior inference is carried out using a novel and efficient MCMC simulation method.

KEYWORDS: Covariate-dependent copula; Bayesian variable selection; tail-dependence; Kendall's τ; MCMC.

1. INTRODUCTION

Copula modeling has been an active research area dating back to Sklar's theorem (Sklar, 1959) in which he proves that a multivariate cumulative distribution function $F(x_1,...,x_M)$ can be written in terms of univariate marginal distributions and a copula function $C(u_1,...,u_M)$, where $u_i = F_i(x_i)$ is the i:th marginal CDF. Various properties and applications of copulas have thereafter been studied, see e.g. Nelsen (2006) for an introduction to copulas, Joe (1997) for dependence and extreme value distribution with copulas, and Dorota (2010) for constructions of multivariate dependences using bivariate copulas.

Copula models have been widely used in financial applications due to its ability in modeling tail-dependence and correlations, see Patton (2012b) for a recent survey. Jaworski et al. (2010) reviews the state of the art approaches in copula estimation, including pair-copula constructions (Czado, 2010). An important concept in copula modeling is the tail-dependence. In bivariate copulas, the tail-dependence describes the dependence of random variables in the tail: $\lim_{u\to 0^+} p(X_1 < F_1^{-1}(u)|X_2 < F_2^{-1}(u))$ is called the lower tail-dependence and $\lim_{u\to 1^-} p(X_1 > F_1^{-1}(u)|X_2 > F_2^{-1}(u))$ is the upper tail-dependence for the two random variables X_1 and X_2. Dobrić & Schmid (2005) use nonparametric methods to estimate the lower tail-dependence in bivariate copulas with continuous marginals. Schmidt & Stadtmuller (2006) explore the tail-dependence estimators for the *tail copula* where the lower tail copula and upper tail copula for a bivariate copula C are defined as $\lim_{t\to\infty} tC(x/t,y/t)$ and $\lim_{t\to\infty}(x+y-t+tC(x/t,y/t))$ respectively, if the limits exist. Another special case in the tail-dependence is the asymptotically independent: x_1 and x_2 are asymptotically independent if $F(x_1,x_2) = \lim_{x_1\to\infty} F(x_1,x_2)\lim_{x_2\to\infty} F(x_1,x_2)$. We do

Feng Li (feng.li@stat.su.se): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.

not further explore this possibility but see Draisma et al. (2004) for hypothesis testing to detect the dependence of extreme events when they are asymptotically independent, and for the study of asymptotic dependence case, see e.g. Ledford & Tawn (1997).

Among the vast and extensive literature in copula modeling, few articles have investigated the underlying causes of the dependence structures. This is partially because the computational complexity is still a challenge in many situations, which explains the relatively simplistic models used for the copula. Moreover, most copula approaches for modeling rank correlation and tail-dependence for existing copulas require firstly modeling intermediate parameters and obtaining the correlation and tail-dependence in the end. Thus the results are neither tractable nor interpretable for practitioners. In this paper we present a general Bayesian approach for copula modeling with explanatory variables entering both the tail-dependence and correlation parameters in the copula function. This construction allows us to explore the drivers of the different forms of dependence between the variables. We propose an efficient MCMC simulation method for the posterior inference which also allows for variable selection in both the marginal models and the copula *features*. In this paper a feature stands for a characteristic in the copula function, e.g. Kendall's τ and tail-dependence are two features of a copula.

The outline of the paper is as follows. In Section 2 we introduce the Bayesian covariate-dependent copula model and Section 2.3 introduces the reparametrized Joe-Clayton copula with some new properties. We discuss the prior specification for the model and present the general form for copula posterior in Section 3. Section 4 presents the details of the MCMC scheme. In Section 5 we apply the model to the daily returns from the S&P100 and S&P600 stock market indices. Section 6 concludes the paper and discusses potential directions for further research. The appendix of the paper documents the necessary analytical computation used in the MCMC.

2. THE COPULA MODEL

A copula is a multivariate distribution that separates the univariate marginals and the multivariate dependence structure. The correspondence between a multivariate distribution $F(x_1,...,x_M)$ and a copula function $C(u_1,...,u_M)$ can be expressed as

$$F(x_1,...,x_M) = F(F_1^{-1}(u_1),...,F_M^{-1}(u_M)) = C(u_1,...,u_M) = C(F_1(x_1),...,F_M(x_M))$$

where the correspondence is one to one with continuous marginal distributions. Copulas provide a general approach to constructing more flexible multivariate densities. For example, the bivariate Gaussian copula $C(u_1,u_2|\rho) = \Psi(F_1^{-1}(u_1),F_2^{-1}(u_2)|\rho)$ with the correlation parameter ρ, is a relaxed Gaussian density in the sense that $F_1(\cdot)$ and $F_2(\cdot)$ do not need to be normal, see e.g. Pitt et al. (2006). New classes of multivariate densities are possible to construct in terms of copulas, see Joe (1997) for details. A key feature in copula models is that the multivariate dependence structure does not depend on the marginal densities. Thus multivariate modeling with copulas consists of two parts: i) separate modeling of each marginal distribution and, ii) modeling the multivariate dependence.

2.1. **Marginal models.**

In this paper we use *margins* as synonym of marginal models. In principle, the copula approach can be used with any margins, but we will assume the margins to be split-t distributions (Li et al., 2010) in our application in Section 5. The split-t is a flexible four-parameter

distribution with the student's t, the asymmetric normal and symmetric normal distributions as its special cases; see Li et al. (2010) for some properties of the split-t distribution.

Following Li et al. (2010) we allow the location parameter μ, the scale parameter ϕ, the degrees of freedom ν and the skewness parameter κ in the split-t density in the margins to be linked to covariates as

(1) $\mu_{ij} = x'_{ij}\beta_{\mu_j}$
(2) $\phi_{ij} = \exp(x'_{ij}\beta_{\phi_j})$
(3) $\nu_{ij} = \exp(x'_{ij}\beta_{\nu_j})$
(4) $\kappa_{ij} = \exp(x'_{ij}\beta_{\lambda_j})$, for $i = 1,...,n$, $j = 1,...,M$

where x_{ij} is the covariate vector for the i:th observation in the j:th margin.

One may also consider using mixture models in the margins. Li et al. (2010) show in an application to S&P500 data that the one-component split-t model with all parameters linked to covariates does well in comparison with mixtures of split-t components. In this paper, we will therefore use the one-component split-t model for demonstration purposes. Note that it is possible to also specify different margins. Our inference procedure can be generally applied.

2.2. **Dependence concepts.** Modeling the multivariate dependence typically involves quantifying two important quantities: correlation and tail-dependence. In copula models, the correlation between two variables are usually measured with rank correlations like Kendall's τ

$$\tau = 4\int\int F(x_1,x_2)\mathrm{d}F(x_1,x_2) - 1 = 4\int\int C(u_1,u_2)\mathrm{d}C(u_1,u_2) - 1.$$

In this paper we focus on modeling Kendall's τ, but our methodology also applies to other correlations like Spearman's ρ. The tail-dependence describes the concordance between extreme values of random variables X_1 and X_2. The lower tail-dependence λ_L and the upper tail-dependence λ_L can also be expressed in terms bivariate copulas

$$\lambda_L = \lim_{u\to 0^+} p(X_1 < F_1^{-1}(u) | X_2 < F_2^{-1}(u)) = \lim_{u\to 0^+} \frac{C(u,u)}{u},$$
$$\lambda_U = \lim_{u\to 1^-} p(X_1 > F_1^{-1}(u) | X_2 > F_2^{-1}(u)) = \lim_{u\to 1^-} \frac{1-C(u,u)}{1-u}.$$

In principle, correlations and tail-dependencies can attain all values in the domain $[-1,1]$, but not all copulas can specify them in the whole interval. For example the Clayton copula and Gumbel copula (Joe, 1997) can only have positive correlations and tail-dependence. Gumbel exhibits strong upper tail-dependence and relatively weak lower tail-dependence. The Fréchet–Hoeffding bounds can help us to select a proper copula that describes the dependence correctly. The copula function satisfies the inequalities

$$\sum_{i=1}^{M} u_i - M + 1 \le C(u_1,...,u_M) \le \min\{u_1,...,u_M\}$$

where the left and right bounds are Fréchet–Hoeffding lower and upper bounds, respectively. In the bivariate case, if the copula is close to the upper bound, it shows strong positive dependence

and if the copula is close to the lower bound, it shows strong negative dependence, see Nelsen (2006) for the proof.

2.3. **The reparametrized Joe-Clayton copula.** In this paper we focus on the bivariate copula modeling with the two-parameter Joe-Clayton copula, which is commonly used in the literature. Our copula modeling approach is general and can be applied to modeling any copulas with any feature of interest. The popular Vine copula construction (Aas et al., 2009; Czado et al., 2012) can also be used to extend our bivariate copula modeling to higher dimensions.

2.3.1. *The copula density.* The Joe-Clayton copula, also known as the BB7 copula, was introduced by Joe (1997) and is of the form

$$\begin{aligned} C(u,v|\theta,\delta) =& \eta(\eta^{-1}(u)+\eta^{-1}(v)) \\ =& 1-\left[1-\left\{\left(1-\bar{u}^{\theta}\right)^{-\delta}+\left(1-\bar{v}^{\theta}\right)^{-\delta}-1\right\}^{-1/\delta}\right]^{1/\theta} \end{aligned}$$

where $\eta(s)=1-[1-(1+s)^{-1/\delta}]^{1/\theta}$, $\theta\geq 1$, $\delta>0$, $\bar{u}=1-u$, $\bar{v}=1-v$ with lower tail-dependence parameter $\lambda_L=2^{-1/\delta}$ and upper tail-dependence parameter $\lambda_U=2-2^{1/\theta}$. The copula density function for the Joe-Clayton copula is then

$$\begin{aligned} c(u,v|\theta,\delta)=\frac{\partial^2 C(u,v,\theta,\delta)}{\partial u\partial v} =& [T_1(u)T_1(v)]^{-1-\delta}\,T_2(u)T_2(v)L_1^{-2(1+\delta)/\delta} \\ & \times(1-L_1^{-1/\delta})^{1/\theta-2}\left[(1+\delta)\theta L_1^{1/\delta}-\theta\delta-1\right] \end{aligned} \tag{1}$$

where $T_1(s)=1-(1-s)^{\theta}$, $T_2(s)=(1-s)^{\theta-1}$ and $L_1=T_1(v)^{-\delta}+T_1(u)^{-\delta}-1$.

The Joe-Clayton copula has been widely used in modeling tail-dependence. Patton (2006) uses a symmetric version of Joe-Clayton copula to model time-varying dependence with its autoregressive terms for daily return of the Deutsche mark with the U.S. dollar and the Japanese yen with U.S. dollar. Aas et al. (2009) and Czado et al. (2012) among others develop a flexible class of multivariate copulas allowing multivariate dependence via a vine structure based on bivariate copulas, including the Joe-Clayton copula. Bouyé & Salmon (2009) apply the Joe-Clayton copula to a quantile regression that allows both positive and negative slopes for the quantile curves. But none of these works have a model that can explain the driving forces behind the dependence structures.

2.3.2. *Kendall's τ.* Kendall's τ of the Joe-Clayton copula for the case $1\leq\theta<2$ can be found in e.g, Smith & Khaled (2012). We now present the full expression for all $\theta\geq 1$.

$$\tau(\theta,\delta)=\begin{cases} 1-2/[\delta(2-\theta)]+4B(\delta+2,2/\theta-1)/(\theta^2\delta), & 1\leq\theta<2; \\ 1-[\psi(2+\delta)-\psi(1)-1]/\delta, & \theta=2; \\ 1-2/[\delta(2-\theta)]-4\pi/\left[\theta^2\delta(2+\delta)\sin(2\pi/\theta)B(1+\delta+2/\theta,2-2/\theta)\right], & \theta>2, \end{cases}$$

where $B(\cdot)$ is the beta function and $\psi(\cdot)$ is the digamma function. It is easy verified that Kendall's τ is continuous for all $\theta\geq 1$.

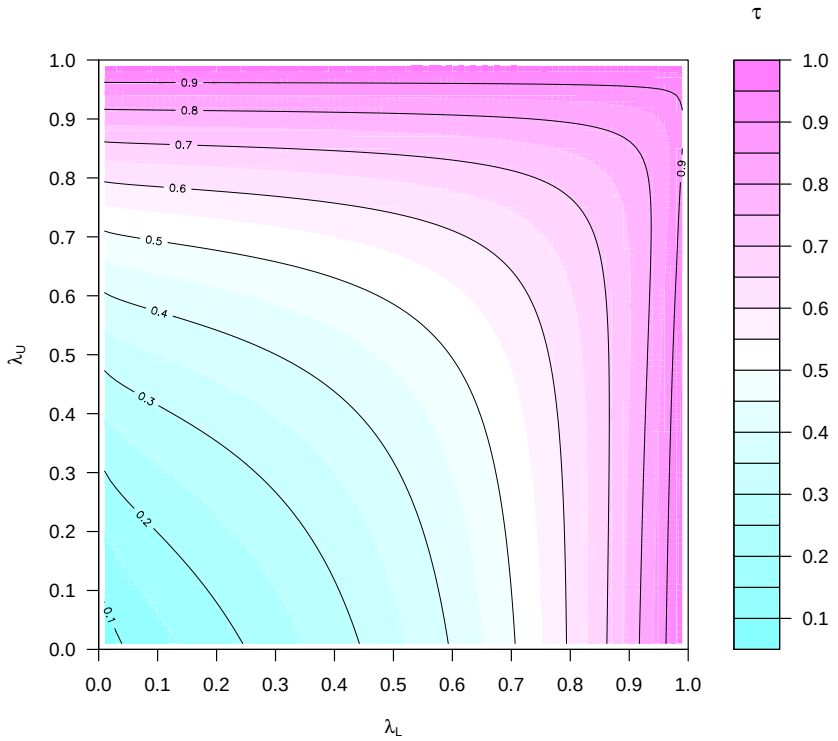

FIGURE 1. Contour plot of the Kendall's τ with respect to lower tail-dependence (λ_L) and upper tail-dependence (λ_U) for the Joe-Clayton copula.

If we employ the equations $\delta = -\log 2/\log \lambda_L$ and $\theta = \log 2/\log(2-\lambda_U)$ to the preceding result, we can rewrite the Kendall's τ in terms of lower and upper tail-dependence as $\tau(\lambda_L, \lambda_U)$, see Figure 1 for a contour plot of these relationships.

The Joe-Clayton copula can only determine positive correlations. If the relationship between two variables is negative, we just need to rotate the axes of the copula and the estimation procedure remains the same. For example, for copulas rotated by 90 degrees, u has to be set to $1-u$; for 270 degrees let v be $1-v$ and for 180 degrees set u and v to $1-u$ and $1-v$, respectively. See Durrleman et al. (2000) for the proof that after this transformation it is still a copula and for other possible transformations to extend current bivariate copula with desired properties. In the financial application in Section 5, the correlations are known to be positive, but modeling negative correlations requires no extra work also for other copulas than the Joe-Clayton copula.

2.3.3. *Some properties.* The Joe-Clayton copula has some unique attributes. The upper tail-dependence and lower tail-dependence are not functionally dependent. The Clayton copula (Clayton, 1978) and the B5 copula (Joe, 1993) are special cases of the Joe-Clayton copula. All of them belong to a more general class of Archimedean copulas. Furthermore, we also find the following new properties.

(1) The inequality holds $0 \leq \lambda_L \leq 2^{1/2-1/(2\tau)}$ when the lower tail-dependence is not extremely high. We say that λ_L and τ are *variationally dependent*. The proof is non-trivial, but we have verified the inequality numerically in a very careful way. We discover the bound of the inequality through the limit of $\tau(\lambda_L, \lambda_U)$ when $\lambda_U \to 0$, see Figure 1 for an illustrative plot.

(2) When $\lambda_L \to 0$ (i.e. $\delta \to 0$), we have

$$\tau \to 1 - \frac{2H(2/\theta) - 2}{2-\theta} = 1 - \frac{2H(2\log(2-\lambda_u)/\log 2) - 2}{2 - \log 2/\log(2-\lambda_u)}$$

and

$$\frac{\partial \tau}{\partial \theta} \to \frac{2(1-H(2/\theta))}{(\theta-2)^2} - \frac{4\psi_1(2/\theta+1)}{(\theta-2)\theta^2}$$

where $H(\cdot)$ is the Harmonic number. A special case is when $\theta \to 2$ (i.e. $\lambda_U \to 2-\sqrt{2} \approx 0.59$), we have $\tau \to 2 - \pi^2/6 \approx 0.36$ and $\partial\tau/\partial\theta \to \pi^2/12 - Zeta(3)/2 \approx 0.22$. where $Zeta(\cdot)$ is the Riemann zeta function.

(3) Furthermore, we also have derived the analytical gradients for the copula density with respect to Kendall's τ and tail-dependence of Joe-Clayton copula in Appendix A.2, which will be used to construct efficient proposal distributions for MCMC.

2.3.4. *Reparametrization.* The parameters in most copula functions do not directly represent the copula features, see e.g. the tail-dependence parameters and Kendall's τ in the Joe-Clayton copula. In this section we describe how to reparameterize the copula function so that the parameters are the copula features of interest.

To simplify the interpretation of the copula model, we parameterize it in terms of lower tail-dependency parameter λ_L and Kendall's τ,

$$\begin{aligned} C(u,v|\lambda_L,\tau) =& 1 - \left[1 - \left\{ \left[1 - \bar{u}^{\log 2/\log(2-\tau^{-1}(\lambda_L))}\right]^{\log 2/\log\lambda_L} \right.\right. \\ & \left.\left. + \left[1 - \bar{v}^{\log 2/\log(2-\tau^{-1}(\lambda_L))}\right]^{\log 2/\log\lambda_L} - 1 \right\}^{\log\lambda_L/\log 2} \right]^{\log(2-\tau^{-1}(\lambda_L))/\log 2} \end{aligned}$$

where $\tau^{-1}(\lambda_L)$ is the inverse function of Kendall's τ given λ_L, i.e. the upper tail-dependence λ_U. And the related reparametrized copula density is obtained by substituting $\delta = -\log 2/\log\lambda_L$ and $\theta = \log 2/\log(2-\tau^{-1}(\lambda_L))$ from (1).

An attractive property of Kendall's τ is that it is invariant with respect strictly monotonic transforms. Other types of correlation like Spearman's rank correlation can be equally well modeled with our method. For measuring the dependence in trivariate distributions, one may consider using a three-dimensional version of correlations, see e.g. García et al. (2013). Correlations in higher dimensions are usually estimated pairwise.

Our parameterization has two main advantages. Firstly, it reduces the efforts for specifying the prior information in our Bayesian approach in Section 3. Secondly, and most importantly, this parameterization make it possible to directly link correlations and tail-dependence to covariates, see Section 2.4 for details.

Modeling the upper tail-dependence is also important in financial applications. There are also alternative reparametrization schemes that allow modeling upper tail dependence parameter λ_L directly. A simple way to achieve the same effect is to rotate the copula for 180 degrees in our parameterization.

2.4. **Covariate-dependent copula parameters.** Letting the Kendall's τ and tail-dependence parameters in copula modeling be fixed numbers is very restrictive. This is particularly true in financial time-series applications where the tail-dependence has been shown to vary with time (Patton, 2006). We introduce a covariate-dependent copula model that allows the copula features to be linked to observed covariate information. A prominent example is covariate-dependent correlation and tail-dependence:

(1) $\tau = l_\tau^{-1}(\boldsymbol{x}'\boldsymbol{\beta}_\tau)$
(2) $\lambda = l_\lambda^{-1}(\boldsymbol{x}'\boldsymbol{\beta}_\lambda)$

where λ without subscripts represents the dependence parameter in the lower and/or upper tail, τ is Kendall's τ, and $\boldsymbol{x}$ is the set of covariates used in the margins. Furthermore $l_\tau(\cdot)$ and $l_\lambda(\cdot)$ are suitable link functions that connect λ and τ with $\boldsymbol{x}$. Other copula parameters can be linked to covariates in the same way.

Patton (2006) allows ARMA-like variation in the dependence parameter. Our approach makes it possible to use all marginal information to model the dependence parameters. This approach not only leads to more interesting interpretations of the features, but generates more accurate forecasts (see Section 5.3) and allows for heteroscedasticity in the dependence parameters. We also use variable selection to select meaningful covariates that influence the dependence. Furthermore, variable selection also reduces the model complexity and prevents overfitting, see Section 4 for details.

3. THE PRIOR AND POSTERIOR

We use the same technique to specify the priors for the marginal parameters and the prior in the copula features. We omit the subscript that indicates the functionality of the parameter in this section for convenience. We will first assume that the model parameters are independent *a priori*, and then turn to a more general situation with dependent model parameters in Section 3.3.

Let $\mathscr{I}$ be the variable selection indicator for a given covariate

$$\mathscr{I}_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0, \end{cases}$$

where β_j is the jth covariate in the model. More informally, this can be expressed as

$$\mathscr{I}_j = \begin{cases} 1 & \text{if variable } j \text{ enters the model} \\ 0 & \text{otherwise.} \end{cases}$$

We standardize each covariate to have zero mean and unit variance and assume prior independence between the intercept β_0 and the slope $\boldsymbol{\beta}$. We can decompose the joint prior as

$$p(\beta_0, \boldsymbol{\beta}, \mathscr{I}) = p(\beta_0)p(\boldsymbol{\beta}, \mathscr{I}) = p(\beta_0)p(\boldsymbol{\beta}|\mathscr{I})p(\mathscr{I}).$$

We will use normal priors for both β_0 and β. We also assume that the intercept is always included in each parameter, so the variable selection indicator for β_0 is always one.

3.1. **Prior for the intercept .** We set the prior for the intercept by following the strategy in Villani et al. (2012) that firstly puts prior information on the model parameters (e.g. τ and λ_L in the Joe-Clayton copula), and then derive the implied prior on the intercept β_0 under the assumption that the covariates are at their means. The technique can be applied to the following two situations directly. When the link is the identity, setting the implied prior on the model parameter is trivially the same as on the intercept. When the link is the log function, assuming a log-normal distribution on the model parameter with mean m and variance σ^2 yields a normal prior with mean $\log(m) - \log[\sigma^2/m^2+1]/2$ and variance $\log[\sigma^2/m^2+1]$ in the intercept.

We now generalize this to a general situation that can be applied to any link function with any distribution. We take the tail-dependence parameter λ as an example, where $0<\lambda<1$ in some copula functions, e.g. the Joe-Clayton copula in Section 2.3. It is natural to consider using the logit link to connect the tail-dependence λ with covariates. When there is only an intercept in the covariates, we have $\lambda = 1/(1+\exp(-\beta_0))$. Therefore, if we assume λ to have a beta distribution $Beta(m,\sigma^2)$ with mean m and variance σ^2, we have the mean and variance for β_0 as

$$E(\beta_0) = \int_0^1 \log(\frac{x}{1-x}) Beta(x,m,\sigma^2)\mathrm{d}x = \psi(\alpha_1) - \psi(\alpha_2),$$
$$V(\beta_0) = \int_0^1 (\log(\frac{x}{1-x}))^2 Beta(x,m,\sigma^2)\mathrm{d}x - E^2(\beta_0) = \psi_1(\alpha_1) + \psi_1(\alpha_2)$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ are the digamma and trigamma functions respectively, $\alpha_1 = -m(m^2-m+\sigma^2)/\sigma^2$ and $\alpha_2 = -1+m+(m-1)^2 m/\sigma^2$. Higher order moments are also possible to obtain either analytically or numerically. We can now set the prior on the intercept β_0 based on the derived mean and variance information. The prior for the intercept in the Kendall's τ parameter can be elicited in the same fashion but the integration domain and link function should be changed accordingly.

3.2. **Prior for the slope and variable selection indicators.** We first consider the case without variable selection. We assume that the slopes are normally distributed with mean 0 and covariance matrix Σ. The extension to a non-zero mean is trivial. The covariance matrix is defined as $\Sigma = c^2 \cdot P^{-1}$ where P is a positive definite symmetric matrix and c is a scaling factor. In the application, P is the identity matrix. Using the inverse Fisher information for P as in Villani et al. (2012) is also possibility.

We now consider the case with variable selection. Conditional on the variable selection indicators, the slopes are still normal distributed with mean $\mu_{\mathscr{J}} + \Sigma_{21}\Sigma_{\mathscr{J}^c}^{-1}(\beta_{\mathscr{J}^c} - \mu_{\mathscr{J}^c})$ and covariance matrix becomes $\Sigma_{\mathscr{J}} - \Sigma_{21}\Sigma_{\mathscr{J}^c}^{-1}\Sigma_{12}$ (Mardia & Kent, 1979), with obvious notations. The prior for each variable selection indicator is identically Bernoulli distributed with probability of p.

A shrinkage prior is often used as an alternative method for reducing model complexity. In our experience, the choice between variable selection and shrinkage estimator depends on the context of the application. Variable selection is usually used to select meaningful variables, which is of great interest here as we are exploring which variables that explain or drive the dependence among

variables. See also Vach et al. (2001) for a comparison for the two approaches in some commonly used models.

3.3. **Priors when the parameters are dependent.** In this section we consider a special case when two or more model parameters are dependent *a priori*. When we reparametrize the original density function in terms of other parameters, it is common to introduce a variational dependence between the new parameters in the sense that the outcome of one parameter puts a restriction on the domain of the other parameter. In our model, the original Joe-Clayton copula has two parameters θ and δ which are variationally independent. When we reparametrize it in terms of lower tail-dependence and Kendall's τ, Section 2.3.3 shows the inequality between the two parameters (see Figure 1 for a visualization of the relations between the parameters).

As before, our aim is to elicit a prior on $\beta_{\tau_0,}, \beta_\tau, \beta_{\lambda_0}$ and β_λ via an elicited joint distribution on τ and λ. When the parameters are variationally dependent and we can no longer assume prior independence and instead we decompose the joint prior for the model parameters as

$$p(\tau,\lambda) = p(\tau|\lambda)p(\lambda). \tag{2}$$

The marginal priors for $\beta_{\lambda_0}, \beta_\lambda$ and its variable selection indicators $\mathscr{I}_\lambda$ are the same as in Section 3.1 and Section 3.2. We will now document the prior for $\beta_{\tau_0}, \beta_\tau$ and its variable selection indicators $\mathscr{I}_\tau$ conditional on λ.

We first introduce the generalized beta function and the generalized logit link function.

Definition 1. ***The generalized beta distribution***. *Let $gBeta(x,a,b,m,\sigma)$ be the generalized beta distribution with mean m, standard deviation σ where $a < x < b$. Then $(x-a)/(b-a)$ follows the beta distribution with mean $(m-a)/(b-a)$ and standard deviation $\sigma/(b-a)$.*

Definition 2. ***The generalized logit function.*** *The generalized logit function that extends the logit function with two parameters a and b as*

$$glogit(x,a,b) = a + \frac{b-a}{1+\exp(-x)},$$

where $a < x < b$.

The generalized beta distribution and the generalized logit link function now both have two boundary parameters a and b. Furthermore, when $a = 0$ and $b = 1$, they reduce to their usual form.

Based on the decomposition in (2) we now assume that τ in Section 2 follows a generalized beta distribution $gBeta(x,a,b,m,\sigma)$ conditional on λ with the generalized logit link $glogit(X_\tau\beta_\tau,a,b)$ where $a = \log(2)/(\log(2)-\log(\lambda))$ and $b = 1$. We can now elicit the prior on the intercept, slopes and variable selection indicators for τ conditional on λ by following Section 3.1 and Section 3.2.

Furthermore, for conditional dependence with more than two parameters, we can always decompose the joint distribution with pairwise conditional distributions and apply the technique thereafter. It is shown in our application that the conditional link function used in the prior also makes the MCMC algorithm more robust and gives higher acceptance probability in Metropolis-Hastings algorithm compared to the case where the prior is simply truncated to the region of allowed (τ,λ) pairs and all proposal draws outside this region are rejected in the MCMC.

3.4. **The joint posterior.** The posterior in the copula model can be written in terms of the likelihoods from the marginal distributions, the copula likelihood and the prior for parameters in the copula and marginal distributions as

$$\begin{aligned}\log p(\{\boldsymbol{\beta},\mathscr{I}\}|\boldsymbol{y},\boldsymbol{x}) =& \text{constant} + \sum_{j=1}^{M} \log p(\boldsymbol{y}_{\cdot j}|\{\boldsymbol{\beta},\mathscr{I}\}_j,\boldsymbol{x}_j) \\ &+ \log \mathscr{L}_C(\boldsymbol{u}|\{\boldsymbol{\beta},\mathscr{I}\}_C,\boldsymbol{y},\boldsymbol{x}) + \log p(\{\boldsymbol{\beta},\mathscr{I}\})\end{aligned}$$

where $\log p(\boldsymbol{y}_{\cdot j}|\{\boldsymbol{\beta},\mathscr{I}\}_j,\boldsymbol{x}_j)$ is the log likelihood in j:th margin, the sets $\{\beta,\mathscr{I}\}_j$ are the parameter blocks in the j:th margin. Furthermore, $\boldsymbol{u} = (\mathbf{u}_1,...,\mathbf{u}_M)$, where $\mathbf{u}_j = \left(u_{1j},...,u_{nj}\right)'$ and $u_{ij} = F_j(y_{ij})$, and $F_j(\cdot)$ is the CDF of the j:th marginal distribution and $\mathscr{L}_C$ is the likelihood for the copula function. In our application, we have $M = 2$ and we use the reparametrized Joe-Clayton copula defined in Section 2.3.4.

4. The general MCMC scheme

We update the copula component together with the marginal components jointly. The joint posterior is not tractable and we use the Metropolis–Hastings within Gibbs sampler, i.e. a Gibbs sampler is used for updating the joint parameter components, with each conditional parameter block $\{\boldsymbol{\beta},\mathscr{I}\}$ updated by the Metropolis–Hastings algorithm. The complete updating scheme is as follows.

4.1. **Metropolis–Hastings within Gibbs.** The updating order in the Gibbs sampler is given in Table 1. We jointly update the coefficients and variable selection indicators $\{\beta,\mathscr{I}\}$ in each parameter block using an efficient tailored Metropolis–Hastings algorithm with integrated finite-step Newton proposals. The acceptance probability for a proposed draw $\{\beta^{(p)},\mathscr{I}^{(p)}\}$ conditional on current value of the parameters $\{\beta^{(c)},\mathscr{I}^{(c)}\}$ is

$$\min\left[1, \frac{p(\{\beta^{(p)},\mathscr{I}^{(p)}\}|\{\beta^{(c)},\mathscr{I}^{(c)}\},Y,X)g(\{\beta^{(c)},\mathscr{I}^{(c)}\}|\{\beta^{(p)},\mathscr{I}^{(p)}\})}{p(\{\beta^{(c)},\mathscr{I}^{(c)}\}|\{\beta^{(p)},\mathscr{I}^{(p)}\},Y,X)g(\{\beta^{(p)},\mathscr{I}^{(p)}\}|\{\beta^{(c)},\mathscr{I}^{(c)}\})}\right] \tag{3}$$

where $g(\cdot)$ is the jumping rule in the Metropolis-Hastings for $\{\beta,\mathscr{I}\}$. Note that it is convenient to decompose $g(\{\beta^{(p)},\mathscr{I}^{(p)}\}|\{\beta^{(c)},\mathscr{I}^{(c)}\} = g_1(\beta^{(p)}|\{\beta^{(c)},\mathscr{I}^{(p)}\}_i)g_2(\mathscr{I}^{(p)}|\{\beta^{(c)},\mathscr{I}^{(c)}\})$. And $g_1(\cdot)$ is the proposal distribution where the proposal mode is from finite-step Newton approximation of the posterior distribution (usually smaller than three steps) starting on current draw and the proposal covariance matrix is from the negative inverse Hessian matrix. In our application, $g_1(\cdot)$ is multivariate t distribution with six degrees of freedom. The distribution of $g_2(\cdot)$ is such that we always propose a change of $\mathscr{I}$: a currently excluded variable is proposed to enter the model, and vice versa. We do not allow all indicator to change in a given iteration, each indicator is proposed to change with probability p_{prop}. This simple scheme works well in the copula model. For alternative types of variable selection schemes, see e.g. Nott & Kohn (2005).

The updating scheme is used in e.g. Villani et al. (2009) and Villani et al. (2012) where it is shown that Metropolis–Hastings with finite-step Newton proposals increases the convergence rate rapidly. The algorithm only requires the gradient for the marginal distribution and copula model with respect to their the (low-dimensional) parameters. Appendix A.1 documents the details for

TABLE 1. The Gibbs sampler for covariate-dependent copula. The notation $\{\beta_\mu, \mathscr{I}_\mu\}_m$ denotes the covariates coefficients and variable selection indicators in copula component m for parameter feature μ. And the notation $\{\beta_\mu, \mathscr{I}_\mu\}_{-m}$ indicates all other parameters in the model except $\{\beta_\mu, \mathscr{I}_\mu\}_m$. The updating order is column-wise from left to right. If dependent link functions are used, the updating should be ordered accordingly.

Margin component (1)	...	Margin component (M)	Copula component (C)
(1.1) $\{\beta_\mu, \mathscr{I}_\mu\}_1 \vert \{\beta_\mu, \mathscr{I}_\mu\}_{-1}$	...	(M.1) $\{\beta_\mu, \mathscr{I}_\mu\}_M \vert \{\beta_\mu, \mathscr{I}_\mu\}_{-M}$	(C.1) $\{\beta_\lambda, \mathscr{I}_\lambda\}_C \vert \{\beta_\lambda, \mathscr{I}_\lambda\}_{-C}$
(1.2) $\{\beta_\phi, \mathscr{I}_\phi\}_1 \vert \{\beta_\phi, \mathscr{I}_\phi\}_{-1}$	...	(M.2) $\{\beta_\phi, \mathscr{I}_\phi\}_M \vert \{\beta_\phi, \mathscr{I}_\phi\}_{-M}$	(C.2) $\{\beta_\tau, \mathscr{I}_\tau\}_C \vert \{\beta_\tau, \mathscr{I}_\tau\}_{-C}$
(1.3) $\{\beta_\nu, \mathscr{I}_\nu\}_1 \vert \{\beta_\nu, \mathscr{I}_\nu\}_{-1}$	...	(M.3) $\{\beta_\nu, \mathscr{I}_\nu\}_M \vert \{\beta_\nu, \mathscr{I}_\nu\}_{-M}$	
(1.4) $\{\beta_\kappa, \mathscr{I}_\kappa\}_1 \vert \{\beta_\kappa, \mathscr{I}_\kappa\}_{-1}$	...	(M.4) $\{\beta_\kappa, \mathscr{I}_\kappa\}_M \vert \{\beta_\kappa, \mathscr{I}_\kappa\}_{-M}$	

calculating the gradient with respect to copula features for reparametrized copulas in the MCMC implementation with both independent and dependent link functions.

An alternative approach is the two-stage estimation method which first independently estimate the margins and then estimates the copula likelihood conditional on the estimated margins, see e.g. Xu (1996) and Joe (1997). The two-stage estimation method is widely used as it reduces the computational difficulty in maximizing the likelihood for high-dimensional copula models. Joe (2005) shows that the asymptotic relative efficiency of the two-stage estimation procedure depends on how close the copula is to the Fréchet bounds. The initial values for the MCMC are obtained by numerical optimization of the posterior distribution. Alternatively, one can use the estimates from a two-stage approach as initial values. An R package for estimating the covariate-dependent copula model with our MCMC scheme is available upon request.

5. APPLICATION TO FINANCIAL DATA

In order to illustrate our method, we use an financial application with daily stock returns. The copula model is the reparametrized Joe-Clayton copula with split-t distributions on the continuous margins. For the discrete case, see e.g. the approach by latent variables for the Gaussian copula in Pitt et al. (2006) and the extension to a general copula in Smith & Khaled (2012).

5.1. **The S&P100 and S&P600 data.**

Our data are daily returns from the S&P100 and S&P600 daily stock market indices during the period from September 15, 1995 to January 16, 2013. The S&P100 index includes the largest and most established companies in the U.S. which is a subset of the well-known S&P500 index. The S&P600 index covers the small capitalization companies which present the possibility of greater capital appreciation, but at greater risk. The S&P600 index covers roughly three percent of the total US equities market.

Patton (2012a) uses hypothesis tests to show that there is significant time-varying dependence between S&P100 and S&P600. Both parametric and nonparametric methods are used to estimate the tail-dependence coefficient in different copula models in Patton (2012a). Nevertheless, little effort has been devote to interpreting the dependence, in particular from using covariate information.

TABLE 2. Description of variables in the S&P100 and S&P600 data.

Variable	Description
Return	Daily return $y_t = 100\log(p_t/p_{t-1})$ where p_t is the closing price.
RM1	Return of last day.
RM5	Return of last week.
RM20	Return of last month.
CloseAbs95	Geometrically decaying average of absolute returns $(1-\rho)\sum_{s=0}^{\infty}\rho^s\lvert y_{t-2-s}\rvert$ with $\rho = 0.95$.
CloseAbs80	Geometrically decaying average of past absolute returns with $\rho = 0.80$.
MaxMin95	Measure of volatility $(1-\rho)\sum_{s=0}^{\infty}\rho^s(\log(p^h_{t-1-s}) - \log(p^l_{t-1-s}))$ with $\rho = 0.95$, where p^h and p^l are the highest and lowest prices.
MaxMin80	Measure of volatility with $\rho = 0.80$.
CloseSqr95	Geometrically decaying average of returns $((1-\rho)\sum_{s=0}^{\infty}\rho^s y^2_{t-2-s})^{1/2}$ with $\rho = 0.95$.
CloseSqr80	Geometrically decaying average of returns with $\rho = 0.80$.

The covariates we used in the margins and in the copula function are described in Table 2. Villani et al. (2009) and Li et al. (2010) apply similar covariates in univariate response regression density estimation on S&P500 data using mixtures of Gaussian and asymmetric student's t densities.

Figure 2 shows the time series of the daily returns. It is seen that there is huge volatility in the returns for both S&P100 and S&P600 during the 2008 financial crisis. Figure 3 depicts the empirical copula $\hat{C}_n$ for Return estimated as proposed by Dobrić & Schmid (2005) by assuming independent observations

$$\hat{C}_n(\frac{i}{n},\frac{j}{n}) = \frac{1}{n}\sum_{k=1}^{n} 1(Y_{1k} \le Y_{1(i)}, Y_{2k} \le Y_{2(j)})$$

for $i, j = 1,...,n$, where $Y_{1(1)} \le,...,\le Y_{1(n)}$ and $Y_{2(1)} \le,...,\le Y_{2(n)}$ are the ordered values of Return for S&P100 and S&P600, respectively. Figure 3 shows that the empirical copula is very close to the Fréchet-Hoeffding upper bound copula, which means extreme positive dependence between data S&P100 and S&P600 (Nelsen, 2006). Figure 3 also suggests that the positive dependence restriction of the Joe-Clayton copula is appropriate for modeling the data without rotating the scale. Furthermore, Joe (2005) shows that the usual two-stage approach for copula estimation is not efficient for extreme dependence near the Fréchet bounds, which is the case in our application.

5.2. **Posterior summary.** We present the posterior summary for the model in Table 3. The conditional link function is used for the dependent Kendall's τ and lower tail-dependence. The efficiency of the MCMC is monitored via the inefficiency factor $\text{IF} = 1 + 2\sum_{i=1}^{\infty}\rho_i$, where ρ_i is the autocorrelation at lag i in the MCMC iterations.

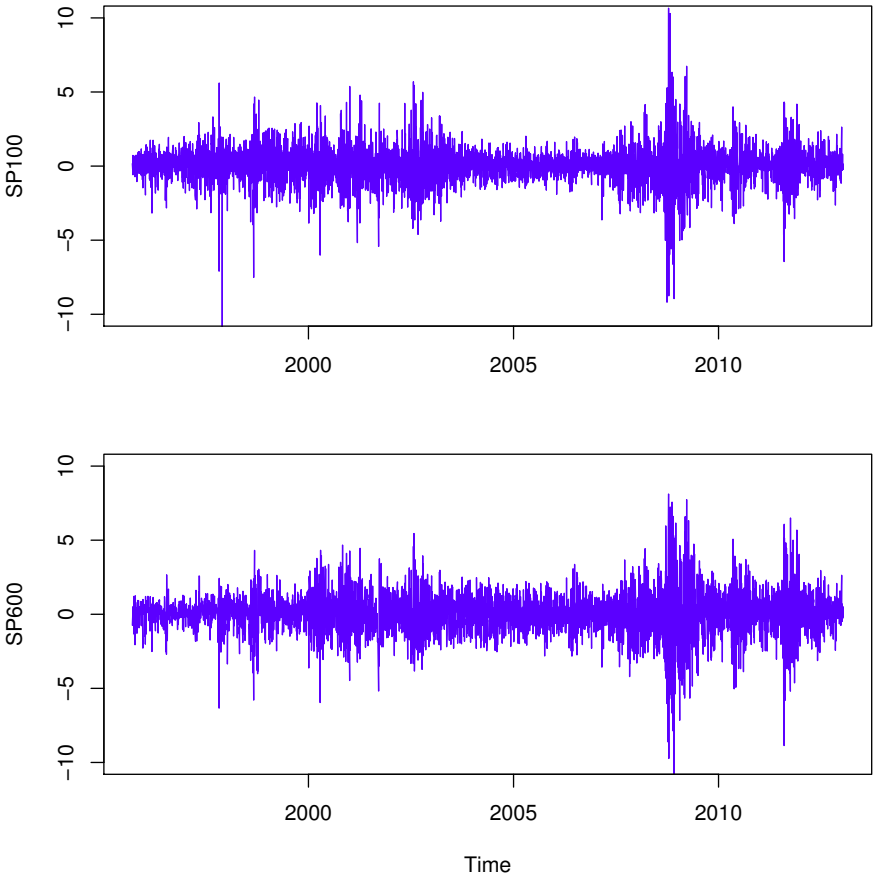

FIGURE 2. Daily return of the S&P100 and S&P600 indices from September 15, 1995 to January 16, 2013.

Note that our marginal models are similar to the model in Li et al. (2010) except that the location parameter of the split-t is fixed in Li et al. (2010). We also estimated the model with the independent link for comparison. The mean posterior acceptance probability is 71% in comparison to 50% when the independent link function is used.

We focus on explaining the results in the copula component and refer to Li et al. (2010) for a detailed interpretation of the marginal models. The variable selection results show that important variables for Kendall's τ are RM1, RM5, CloseAbs80 and MaxMin95 in the S&P600 margin and the variables CloseAbs80, MaxMin95 , MaxMin80 and CloseSqr95 in the S&P100 margin. Variables with large posterior inclusion probabilities in the lower-tail dependence part are: CloseSqr95, RM1 and RM5 in S&P600 margin and RM1, RM20 and CloseAbs80 from the variables in the S&P100 margin. Multicollinearity may occur when the same variables are used in a margin and in the copula parameters. Figure 4 shows that the same covariate in S&P100 and in S&P600 tend to be highly correlated. Since covariates in both margins are used in the copula, there is a risk of covariate duplication with associated problems with non-identification. Table 3 shows that when a covariate in one margin is selected in the copula feature, the same covariate in the

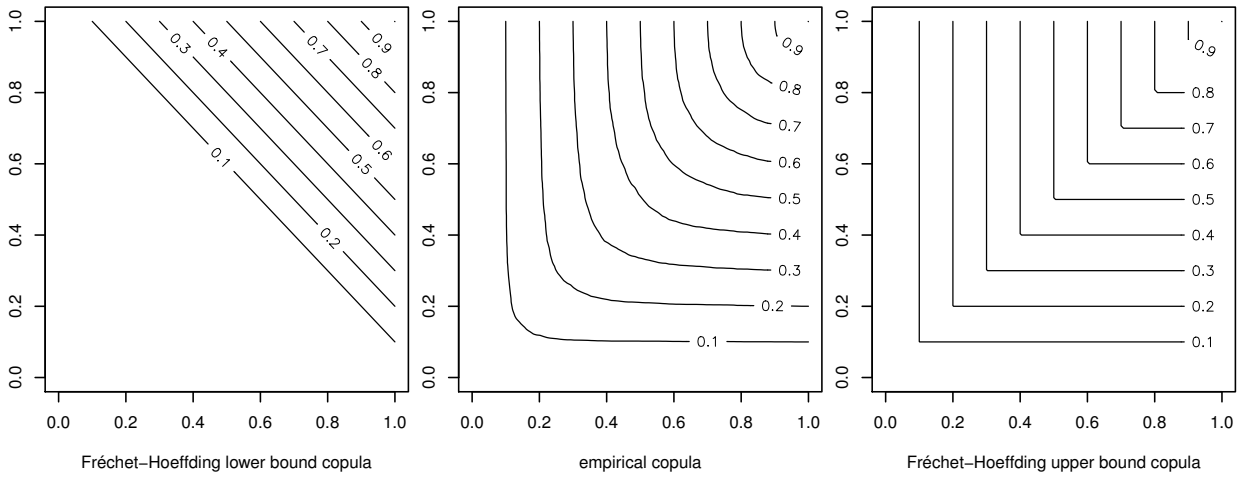

FIGURE 3. The empirical copula for daily return of S&P100 and S&P600 indices from September 15, 1995 to January 16, 2013 (middle) and the Fréchet-Hoeffding lower bound copula (left) and Fréchet-Hoeffding upper bound copula (right).

other margin does not appear in the copula, which indicates that our variable selection algorithm is efficient in removing superfluous covariates.

Figure 5 shows a correlation between S&P100 and S&P600 indices with the posterior mean of Kendall's τ being rather stable around 0.3 over time (bottom left subgraph). The tail-dependence is not so strong during normal time, but there is significant variation over time (bottom right subgraph). The variation in λ_L over time is larger than the variation in Kendall's τ and there is a very high dependence in the tail even though the overall correlation is relatively small.

Figure 6 depicts the posterior contour plot for the Joe-Clayton copula model for some random dates before and during the 2008 financial crisis. By comparing with the Fréchet-Hoeffding upper bound in Figure 3, we can see that the lower tail-dependence is higher during the crisis in comparison to the dates before the crisis. Also note that the empirical copula overestimates the dependence during normal time and underestimates the dependence during the 2008 financial crisis because it assumes temporally independent observations. Our covariate-dependent copula model not only captures but also estimates the dynamic dependence.

5.3. **Model comparison.** We evaluating the model performance based on out-of-sample prediction. In our time series application, we estimate the model based on the 80% of historical data and then predict the last 20% data. We evaluate the quality of the one-step-ahead predictions using the log predictive score (LPS)

$$\text{LPS} = \log p(D_{(T+1):(T+p)}|D_{1:T}) = \sum_{i=1}^{p} \log \int p(D_{T+i}|\theta, D_{1:(T+i-1)}) p(\theta|D_{1:(T+i-1)}) \mathrm{d}\theta$$

where $D_{a:b}$ is the dataset from time a to b and θ are the model parameters. This calculation of the LPS is usually computationally costly because every prediction needs a new posterior sample from the posterior based on the data available at the time of the forecast. We approximate the LPS by assuming that the posterior does not change much as we add a few data points to the estimation

TABLE 3. Posterior summary of copula model with S&P100 and S&P600 data. In the copula component part, the first row and second row for β and $\mathscr{I}$ are the results for the combined covariates that are used in the first and second marginal model, respectively. The intercept are always included in the model.

	Intercept	RM1	RM5	RM20	CloseAbs95	CloseAbs80	MaxMin95	MaxMin80	CloseSqr95	CloseSqr80
					Marginal component (1)					
β_μ	0.222	−0.076	−0.104	0.096	0.205	−0.385	0.090	0.173	−0.316	0.372
$\mathscr{I}_\mu$	1	0.15	0.09	0.08	0.06	0.11	0.11	0.10	0.03	0.05
β_ϕ	−0.001	−0.019	−0.118	**−0.050**	0.143	−0.283	**0.021**	0.005	0.070	0.352
$\mathscr{I}_\phi$	1	0.01	0.091	**0.89**	0.01	0.00	**0.92**	0.07	0.00	0.00
β_ν	0.900	−0.088	−0.401	0.263	−0.500	−0.832	0.851	−0.053	−0.802	−0.592
$\mathscr{I}_\nu$	1	0.15	0.00	0.01	0.16	0.12	0.05	0.05	0.11	0.02
β_κ	−0.272	0.007	0.009	−0.061	−0.451	0.512	0.003	−0.357	0.445	−0.241
$\mathscr{I}_\kappa$	1	0.02	0.16	0.19	0.02	0.13	0.10	0.19	0.11	0.15
					Marginal component (2)					
β_μ	0.260	0.150	−0.089	0.044	0.173	−0.079	0.306	−0.124	−0.037	0.164
$\mathscr{I}_\mu$	1	0.20	0.17	0.10	0.05	0.03	0.11	0.05	0.14	0.14
β_ϕ	0.208	0.025	**−0.127**	−0.042	−0.002	−0.300	0.145	0.165	0.066	0.291
$\mathscr{I}_\phi$	1	.10	**1.00**	0.09	0.03	0.00	0.01	0.11	0.03	0.07
β_ν	2.843	0.091	−0.435	−0.612	0.417	−0.388	0.299	−0.270	0.338	−0.467
$\mathscr{I}_\nu$	1.00	0.10	0.13	0.03	0.04	0.10	0.12	0.18	0.08	0.12
β_κ	−0.265	−0.104	0.059	−0.055	0.389	−0.197	−0.708	0.401	0.253	−0.110
$\mathscr{I}_\kappa$	1	0.06	0.02	0.05	0.00	0.00	0.09	0.00	0.06	0.05
					Copula component (C)					
β_{λ_L}	−8.165	**−0.555**	1.793	**0.005**	−0.170	**0.110**	−0.667	−1.448	−0.636	0.050
		1.463	**0.405**	0.934	−2.138	−1.288	−1.954	−1.577	**−1.873**	−1.805
$\mathscr{I}_{\lambda_L}$	1.00	**0.98**	0.37	**0.63**	0.02	**0.61**	0.36	0.35	0.39	0.29
		1.00	**1.00**	0.00	0.30	0.35	0.40	0.00	**0.61**	0.34
β_τ	−1.726	0.181	−0.217	−0.304	−0.107	**0.115**	**0.005**	**−0.257**	**1.068**	0.037
		−0.191	**0.170**	0.274	0.144	**−0.051**	**−0.671**	0.059	−0.209	−0.181
$\mathscr{I}_\tau$	1.00	0.00	0.00	0.00	0.00	**0.90**	**0.99**	**1.00**	**0.85**	0.00
		1.00	**1.00**	0.00	0.00	**1.00**	**1.00**	0.00	0.00	0.00

The inefficiency factors for the parameters are all bellow 25.

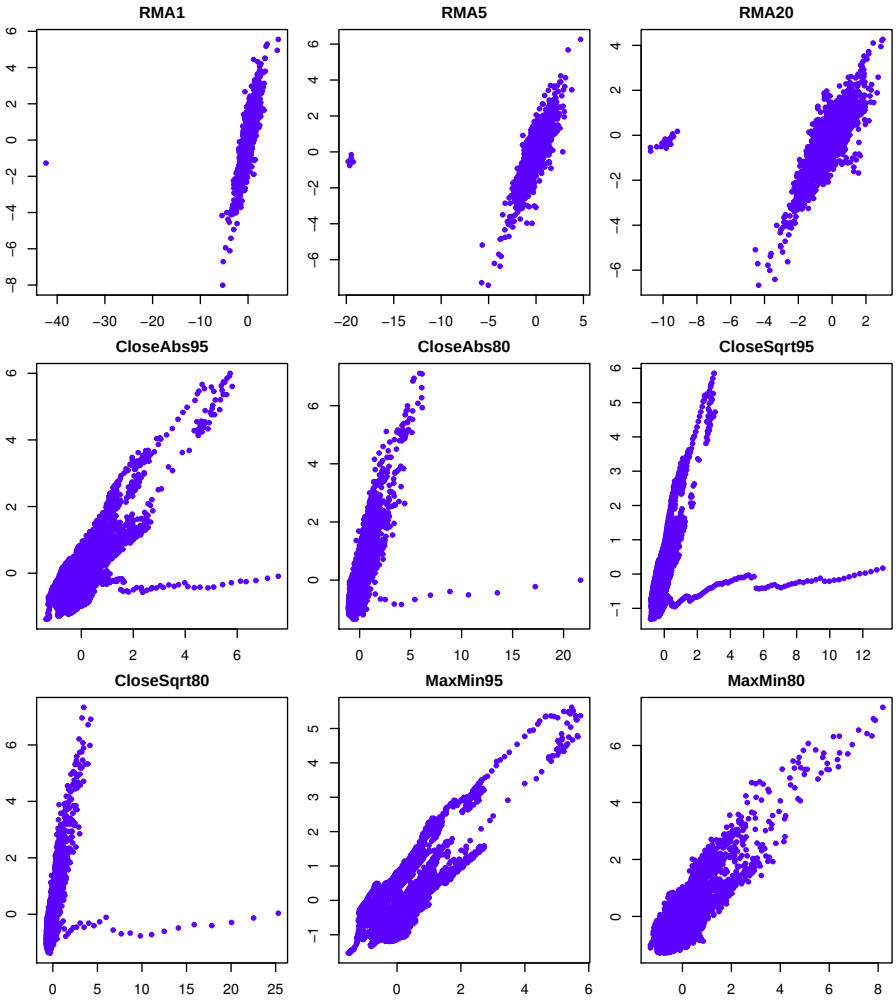

FIGURE 4. The pairwise scatter plot for the same covariate in S&P100 (in x-axis) and S&P600 (in y-axis) margins that also enters in the copula function.

sample. This approximation has the advantage that we can parallelize the LPS evaluation on multiple processors.Villani et al. (2009) document that this type of approximation is accurate in an application of smooth mixture of Gaussians for density predictions of the S&P500 data.

Table 4 shows the out-of-sample comparison of models. Models with covariates in the copula features outperforms the commonly used model without covariates in the copula. Moreover, variable selection also enhances the model's predictive performance.

6. CONCLUDING REMARKS

We have proposed a general approach for modeling a covariate-dependent copula. The copula parameters as well as the parameters in the margins are linked to covariates. We use an efficient Bayesian MCMC method to sample the posterior distribution and to simultaneously perform variable selection in all parts of the model. An application to the daily returns of S&P100 and S&P600

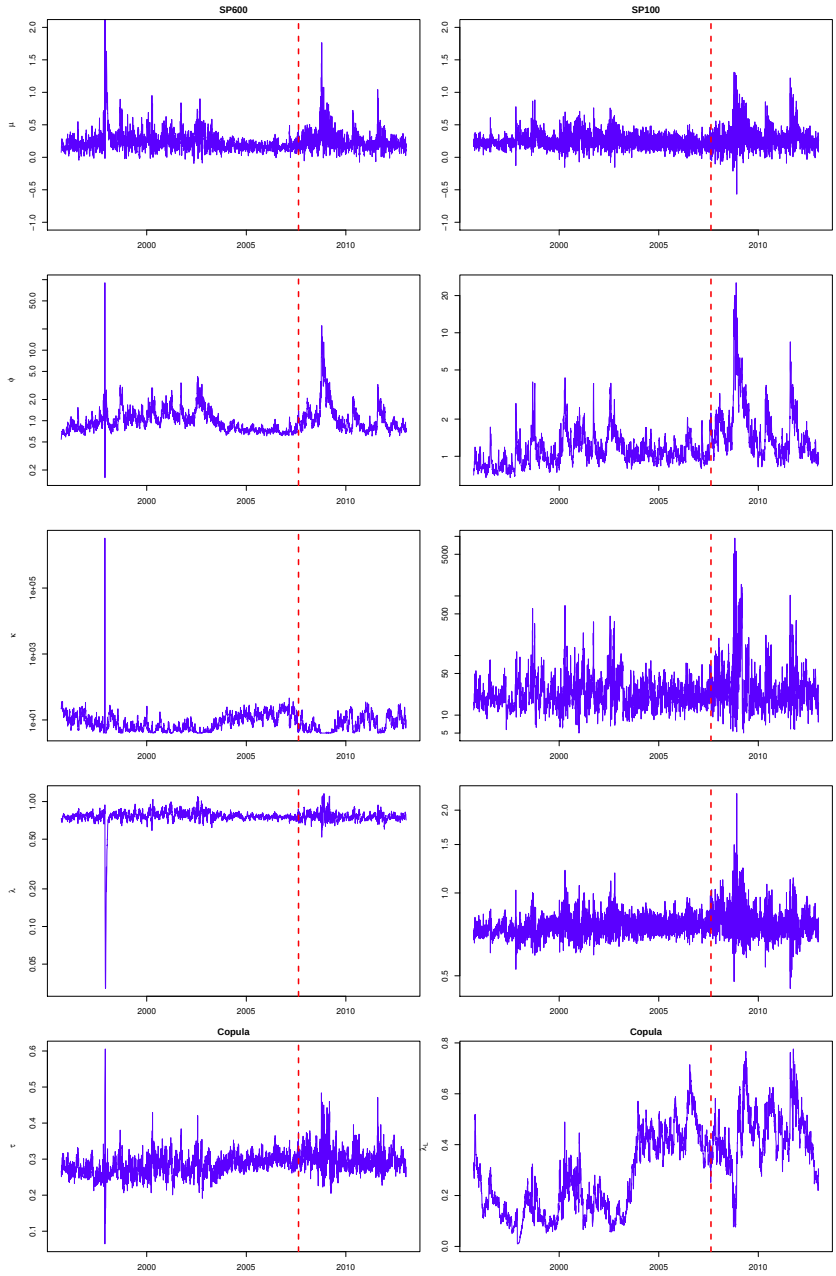

FIGURE 5. The posterior mean for the model parameters with S&P100 and S&P600 data from September 15, 1995 to January 16, 2013. The first four rows of subplots show time series plots of the location, scale, degrees of freedom and skewness parameters in margin S&P100 (left) and S&P600 (right). The subplots on the last row are the time series plot for Kendall's τ (bottom-left) and the lower tail-dependence (bottom-right). The dashed vertical bars indicate the beginning of the 2008 financial crisis.

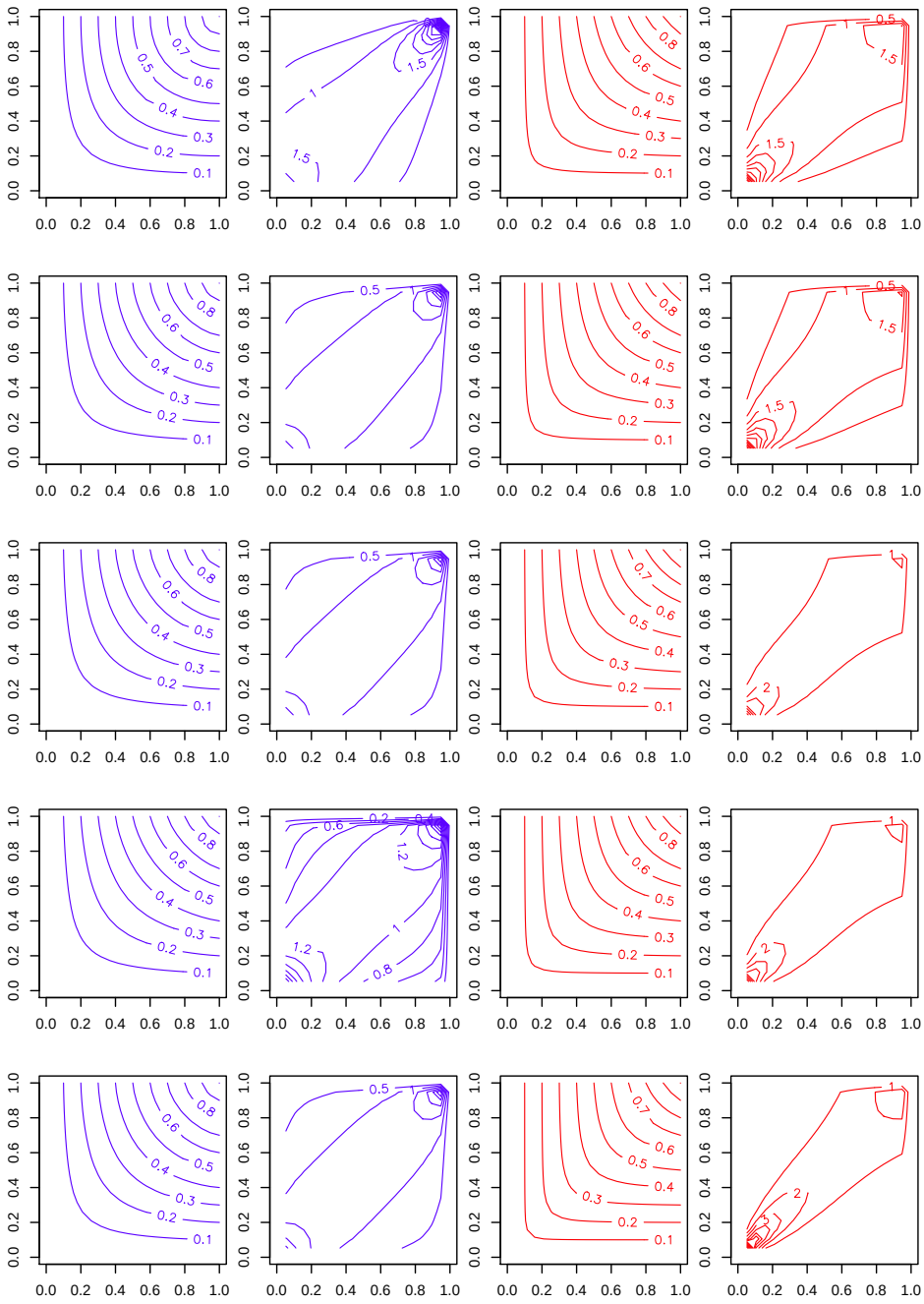

FIGURE 6. Contour plot for the posterior copula. The two leftmost columns show contour plots of the estimated copula and its density at five random dates before the 2008 financial crisis. The contour plots in each column are ordered with respect to the lower tail-dependence. The two rightmost columns show the same plots but for five dates during the 2008 financial crisis.

TABLE 4. Model comparison with LPS. Note that variable selection is used in all situations in the margins.

	LPS	Numerical standard error
No-covariates in τ and λ	-1324	0.5
Covariates in τ and λ - no variable selection	-1250	1.2
Covariates in τ and λ - variable selection	-1238	1.1

indices shows the advantages of this approach in terms of understanding the copula dependence via covariates and improved out-of-sample prediction performance. Covariate-dependent copula modeling with discrete margins is also possible using the data augmentation in Pitt et al. (2006) and Smith & Khaled (2012).

7. ACKNOWLEDGEMENTS

This work was initiated during the author's visit to Division of Statistics, Department of Computing and Information Science at Linköping University in the winter 2011.

APPENDIX A. THE MCMC DETAILS

In this section, we briefly present the MCMC details. The MCMC implementation is straightforward, but requires great care of the proposal distribution in the Metropolis–Hastings algorithm.

A.1. **The chain rule.** We use the finite-step Newton method embedded in the Metropolis-Hastings algorithm that requires the analytical gradient for the posterior with respect to the parameters of interest in marginal and copula components. The chain rule of gradient conveniently modularizes the copula model and reduces the complexity of the the gradient calculation.

A.1.1. *The chain rule for copula parameters.*

$$\begin{aligned}
\frac{\partial \log c(u_1,...,u_M,\lambda_L,\tau)}{\partial \lambda_L} &= \frac{\partial \log c(u_1,...,u_M,\theta,\delta)}{\partial \delta} \times \left(\frac{\partial \lambda_L}{\partial \delta}\right)^{-1} \\
\frac{\partial \log c(u_1,...,u_m,\lambda_L,\tau)}{\partial \tau} &= \frac{\partial \log c(u_1,...,u_m,\theta,\delta)}{\partial \theta} \times \left(\frac{\partial \tau(\theta,\delta)}{\partial \theta}\right)^{-1} \\
\frac{\partial \log c(u_1,...,u_M,\lambda_L,\tau)}{\partial \varphi_m} &= \frac{\partial \log c(u_1,...,u_M,\theta,\delta)}{\partial u_m} \times \frac{\partial u_m}{\partial \varphi}
\end{aligned}$$

where φ_m is any parameter in the m:th margin and u_m is the CDF function of its marginal density. The parameters θ and δ are the intermediate parameters that link the dependence and correlations with the traditional parametrization for copulas.

The MCMC algorithm requires evaluating $\tau_{\lambda_L}^{-1}$ excessively. It can be evaluated numerically or through a dictionary-lookup method. In practice, we have found that the dictionary-lookup method is particularly fast and robust. Modeling the upper tail-dependence can be done in the same manner.

Our model in Section 2 is covariate-dependent. Let $l(\varphi) = x'\beta$ be the link function where φ is the parameter of interest. The gradient expression can be written as

$$\frac{\partial \log c(u_1,...,u_M,\varphi)}{\partial \beta} = \frac{\partial \log c(u_1,...,u_M,\varphi)}{\partial \varphi} \times \left(\frac{\partial l(\varphi)}{\partial \varphi}\right)^{-1} \times \frac{\partial x'\beta}{\partial \beta}.$$

When the conditional link function is used, e.g. τ depending on λ_L in our model in the link function $l(\tau|\lambda_L) = x'\beta$, the gradient for λ_L is slightly complicated. One needs to write τ as a function of λ_L with the link function and substitute it into the copula density. The gradient for λ_L is obtained thereafter. The details are straightforward, but lengthy, and will be omitted here.

A.2. **Gradients for parameters in Joe-Clayton copula.** The gradient for the Joe-Clayton copula w.r.t lower tail-dependence parameters λ_L can be decomposed as

$$\begin{aligned}\frac{\partial \log c(u,v,\theta,\delta)}{\partial \delta} =& -\log T_1(u) - \log T_1(v) - \frac{2(1+\delta)\Delta_1}{\delta L_1} - (\frac{1}{\theta} - 2)\frac{\log L_1 - \delta\Delta_1/L_1}{\delta^2(L_1^{1/\delta} - 1)} \\ &+ \frac{2\log L_1}{\delta^2} + \frac{L_1^{1/\delta} - (1+\delta)L_1^{1/\delta}(\log L_1 - \delta\Delta_1/L_1)/\delta^2 - 1}{(1+\delta)L_1^{1/\delta} - \delta - 1/\theta},\end{aligned}$$

where $\Delta_1 = \partial L_1/\partial\delta = -T_1(u)^{-\delta}\log T_1(u) - T_1(v)^{-\delta}\log T_1(v)$. Furthermore, $\partial\lambda_L/\partial\delta = 2^{-1/\delta}\log 2/\delta^2$.

The gradient for Joe-Clayton copula with respect to the Kendall's τ is decomposed as

$$\begin{aligned}\frac{\partial \log c(u,v,\theta,\delta)}{\partial \theta} =& -(1+\delta)\Delta_2(0) + \Delta_3(u) + \Delta_3(v) + 2(1+\delta)\Delta_2(-\delta)/L_1 \\ &+ \frac{(1-2\theta)\Delta_2(1/\delta)}{(1-L_1^{-1/\delta})\theta\delta} - \frac{\log(1-L_1^{-1/\delta})}{\theta^2} \\ &+ \frac{(1+\delta)L_1^{1/\delta} + \theta(1+\delta)/\delta L_1^{1/\delta-1}\Delta_2(-1/\delta) - \delta}{(1+\delta)\theta L_1^{1/\delta} - \theta\delta - 1}\end{aligned}$$

where $\Delta_2(d) = -T_1(u)^{d-1}(1-u)^{\theta}\log(1-u) - T_1(v)^{d-1}(1-v)^{\theta}\log(1-v)$ and $\Delta_3(s) = \partial\log T_2(s)/\partial\theta = (1-s)^{\theta-1}\log(1-s)/T_2(s)$. Furthermore,

$$\frac{\partial\tau(\theta,\delta)}{\partial\theta} = \begin{cases} -2/[(\theta-2)^2\delta] - 8B(2+\delta,2/\theta-1)\left[\theta + \psi(2/\theta-1) - \psi(2/\theta+\delta+1)\right]/(\theta^4\delta), \\ \qquad\qquad 1 \le \theta < 2; \\ -[12 + 24\psi(1) + 6\psi^2(1) + \pi^2 - 12(2+\psi(1))\psi(2+\delta) + 6\psi^2(2+\delta) - 6\psi_1(2+\delta)]/(24\delta), \\ \qquad\qquad \theta = 2; \\ \dfrac{\begin{pmatrix} -2(2+\delta)\theta^4 B(1+\delta+2/\theta,2-2/\theta) - 8\pi^2(\theta-2)^2\cos(2\pi/\theta)/\sin^2(2\pi/\theta) \\ -8\pi(\theta-2)^2[\psi(1+\delta+2/\theta) - \psi(2-2/\theta) - \theta]/\sin(2\pi/\theta) \end{pmatrix}}{\delta(2+\delta)(\theta-2)^2\theta^4 B(1+\delta+2/\theta,2-2/\theta)}, \\ \qquad\qquad \theta > 2. \end{cases}$$

where $\psi_1(\cdot)$ is the trigamma function. then gradient for the case $\theta = 2$ can be obtained by taking the limiting result from the cases of $1 \leq \theta < 2$ or $\theta > 2$ when $\theta \to 2$.

$$\frac{\partial \tau(\theta,\delta)}{\partial \delta} = \begin{cases} -2/[(\theta-2)\delta^2] + 4B(2+\delta, 2/\theta - 1)[\psi(2+\delta) - \psi(2/\theta+\delta+1) - 1/\delta]/(\theta^2\delta), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 1 \leq \theta < 2; \\ [\psi(2+\delta) - \delta\psi_1(2+\delta) - \psi(1) - 1]/\delta^2, \qquad\qquad\qquad\qquad \theta = 2; \\ -\dfrac{2}{(\theta-2)\delta^2} - \dfrac{4\pi[\psi(3+\delta) - \psi(2/\theta+\delta+1) - 2(1+\delta)/(2\delta+\delta^2)]}{(2+\delta)\delta\theta^2\sin(2\pi/\theta)B(1+\delta+2/\theta, 2-2/\theta)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \theta > 2. \end{cases}$$

For the Joe-Clayton copula, u and v are exchangeable, and we only present the derivative with respect to u:

$$\begin{aligned} \frac{\partial \log c(u,v,\theta,\delta)}{\partial u} =& (1+\delta)\theta\Delta_4(0) + (1-\theta)[(1-u)^{\theta-2}/T_2(u) + (1-v)^{\theta-2}/T_2(v)] \\ & - 2(1+\delta)\Delta_4(-\delta)/L_1 - (1/\theta - 2)L_1^{-1/\delta-1}\Delta_4(-\delta)/(1-L_1^{-1/\delta}) \\ & - (1+\delta)\theta L_1^{1/\delta-1}\Delta_4(-\delta)/\left[(1+\delta)\theta L_1^{1/\delta} - \theta\delta - 1\right] \end{aligned}$$

where $\Delta_4(d) = -T_1(u)^{d-1}(1-u)^{\theta-1}\theta - T_1(v)^{d-1}(1-v)^{\theta-1}\theta$.

A.3. **Gradients for parameters in marginal distributions.** The direct derivative of CDF function with respect to its parameters is straightforward for most densities. We only document the split-t case. Let $I = \kappa$ if $y > \mu$ and $I = 1$ elsewhere, $J = 1$, if $y > \mu$ and $J = -1$, and $A = I^2\nu\phi^2/[(y-\mu)^2 + I^2\nu\phi^2]$, the gradient for the split-t CDF function with respect to it parameters μ, ϕ, κ, ν are as follows,

$$\begin{aligned} \frac{\partial u_{split-t}(y,\mu,\phi,\kappa,\nu)}{\partial \mu} =& -\frac{2I\sqrt{\frac{1}{(y-\mu)^2+I^2\nu\phi^2}}A^{\nu/2}}{(1+\kappa)\text{Beta}\left[\frac{\nu}{2},\frac{1}{2}\right]}, \\ \frac{\partial u_{split-t}(y,\mu,\phi,\kappa,\nu)}{\partial \phi} =& -\frac{2I(y-\mu)\sqrt{\frac{1}{(y-\mu)^2+I^2\nu\phi^2}}A^{\nu/2}}{(1+\kappa)\phi\text{Beta}\left[\frac{\nu}{2},\frac{1}{2}\right]}, \\ \frac{\partial u_{split-t}(y,\mu,\phi,\kappa,\nu)}{\partial \phi} =& \begin{cases} -\frac{\text{Beta}_R\left[A,\frac{\nu}{2},\frac{1}{2}\right]}{(1+\kappa)^2} & y < \mu \\ -\frac{2(1+\kappa)(y-\mu)\sqrt{\frac{1}{(y-\mu)^2+\kappa^2\nu\phi^2}}A^{\nu/2} + \text{Beta}\left[A,\frac{\nu}{2},\frac{1}{2}\right]}{(1+\kappa)^2\text{Beta}\left[\frac{\nu}{2},\frac{1}{2}\right]} & \text{elsewhere,} \end{cases} \\ \frac{\partial u_{split-t}(y,\mu,\phi,\kappa,\nu)}{\partial \phi} =& \frac{I}{2(1+\kappa)\nu^2\text{Beta}\left[\frac{\nu}{2},\frac{1}{2}\right]}\left\{4JA^{\nu/2}{}_pF_q\left[\left\{\frac{1}{2},\frac{\nu}{2},\frac{\nu}{2}\right\},\left\{1+\frac{\nu}{2},1+\frac{\nu}{2}\right\},A\right]\right. \\ & \left. +\nu\left(-2(y-\mu)\sqrt{\frac{1}{(y-\mu)^2+\kappa^2\nu\phi^2}}A^{\nu/2}\right.\right. \end{aligned}$$

$$- \nu J \text{ Beta} \left[A, \frac{\nu}{2}, \frac{1}{2} \right] \left(\log(A) - \psi\left(\frac{\nu}{2}\right) + \psi\left(\frac{1+\nu}{2}\right) \right) \Bigg\}$$

where $\text{Beta}_R$is the regularized beta function, ${}_pF_q$ is the generalized hypergeometric function.

However there are exceptions when this derivative is numerically unstable in practice. In this situation, we propose an alternative approach. Note that

$$\frac{\partial u}{\partial \varphi} = \frac{\partial F(y,\varphi)}{\partial \varphi} = \int_{-\infty}^{y} \frac{\partial f(x,\varphi)}{\partial \varphi} \mathrm{d}x \tag{4}$$

where $u = F(y,\varphi)$ is the CDF function of density $f(y,\varphi)$, and calculating $\partial f(y,\varphi)/\partial\varphi$ is usually easier than $\partial F(y,\varphi)/\partial\varphi$. When the integral cannot be easily obtained analytically, numerical methods can be applied in the last stage. Furthermore, Equation (4) has the advantage of connecting existing derivatives of the PDF with respect to its parameters, e.g. Li et al. (2010) (asymmetric student-*t* density where asymmetric normal and symmetric student-*t* densities are its special cases), Li et al. (2011) (gamma and log-normal densities) and Villani et al. (2012) (negative binomial, beta and generalized Poisson densities) and Li & Villani (2013) (spline model with free knots as unknown parameters).

REFERENCES

AAS, K., CZADO, C., FRIGESSI, A. & BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44**, 182–198.

BOUYÉ, E. & SALMON, M. (2009). Dynamic copula quantile regressions and tail area dynamic dependence in Forex markets. *The European Journal of Finance* **15**, 721–750.

CLAYTON, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

CZADO, C. (2010). Pair-copula constructions of multivariate copulas. In *Copula theory and its applications: proceedings of the workshop held in Warsaw, 25-26 September 2009*, P. Jaworski, F. Durante, W. K. Härdle & T. Rychlik, eds. Springer, pp. 93–109.

CZADO, C., SCHEPSMEIER, U. & MIN, A. (2012). Maximum likelihood estimation of mixed c-vines with application to exchange rates. *Statistical Modelling* **12**, 229–255.

DOBRIĆ, J. & SCHMID, F. (2005). Nonparametric estimation of the lower tail dependence λ_L in bivariate copulas. *Journal of Applied Statistics* **32**, 387–407.

DOROTA, K. (2010). *Dependence Modeling: Vine Copula Handbook*. World Scientific.

DRAISMA, G., DREES, H., FERREIRA, A. & DE HAAN, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli* **10**, 251–280.

DURRLEMAN, V., NIKEGHBALI, A. & RONCALLI, T. (2000). A simple tranformation of copulas. *Working paper* Groupe de Recherce Operationelle, Credit Lyonnais, France.

GARCÍA, J. E., GONZÁLEZ-LÓPEZ, V. & NELSEN, R. (2013). A new index to measure positive dependence in trivariate distributions. *Journal of Multivariate Analysis* **115**, 481–495.

JAWORSKI, P., DURANTE, F., HÄRDLE, W. K. & RYCHLIK, T. (2010). *Copula theory and its applications: proceedings of the workshop held in Warsaw, 25-26 September 2009*, vol. 198. Springer.

JOE, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis* **46**, 262–282.

JOE, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London.

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* **94**, 401–419.

LEDFORD, A. W. & TAWN, J. A. (1997). Modelling Dependence within Joint Tail Regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 475–499.

LI, F. & VILLANI, M. (2013). Efficient Bayesian multivariate surface regression. *Scandinavian Journal of Statistics* **forthcoming**.

LI, F., VILLANI, M. & KOHN, R. (2010). Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference* **140**, 3638–3654.

LI, F., VILLANI, M. & KOHN, R. (2011). Modeling conditional densities using finite smooth mixtures. In *Mixtures: estimation and applications*, K. Mengersen, C. Robert & M. Titterington, eds. John Wiley & Sons Inc, Chichester, pp. 123–144.

MARDIA, K. & KENT, J. (1979). *Multivariate analysis*. Academic Press, London.

NELSEN, R. (2006). *An introduction to copulas*. Springer Verlag.

NOTT, D. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.

PATTON, A. (2006). Modelling asymmetric exchange rate dependence. *International economic review* **47**, 527–556.

PATTON, A. (2012a). Copula methods for forecasting multivariate time series. *Handbook of Economic Forecasting* , 1–76.

PATTON, A. (2012b). A review of copula models for economic time series. *Journal of Multivariate Analysis* **110**, 4–18.

PITT, M., CHAN, D. & KOHN, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554.

SCHMIDT, R. & STADTMULLER, U. (2006). Non-parametric Estimation of Tail Dependence. *Scandinavian Journal of Statistics* **33**, 307–335.

SKLAR, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229–231.

SMITH, M. & KHALED, M. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association* **107**, 290–303.

VACH, K., SAUERBREI, W. & SCHUMACHER, M. (2001). Variable selection and shrinkage: comparison of some approaches. *Statistica Neerlandica* **55**.

VILLANI, M., KOHN, R. & GIORDANI, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* **153**, 155–173.

VILLANI, M., KOHN, R. & NOTT, D. J. (2012). Generalized smooth finite mixtures. *Journal of Econometrics* **171**, 121–133.

XU, J. (1996). *Statistical modelling and inference for multivariate and longitudinal discrete response data*. Ph.D. thesis, Department of Statistics, University of British Columbia.

COMPUTATIONAL IMPLEMENTATION FOR FLEXIBLE BAYESIAN CONDITIONAL DENSITY ESTIMATION

FENG LI

We will here give a brief overview of the computational implementation of modeling conditional densities with R packages that we have developed in this thesis. For details regarding the MCMC schemes, see the corresponding paper and the help in the software. The packages can be downloaded from the author's homepage.

In general, implementing the core code requires three modules.

(1) *The model module*
In this part, the likelihood function and priors for the model should be specified. In particular, an efficient MCMC usually requires the analytical gradient for the posterior with respect to the model parameters to be coded. The model can also be extended and generalized by following the existing code structure.

(2) *The MCMC module*
The MCMC part contains the MCMC samplers, i.e. the Metropolis-Hastings within Gibbs or other types of algorithms. Variable selection is also coded in this part.

(3) *Model prediction and evaluation module*
This part of code implements the predictive distribution, methods of evaluating conditional densities and log predictive density score or other types of criteria for model evaluation. This part is not essential if one only needs to fit a flexible density model.

We briefly describe the core functions in the `movingknots` package that is mainly designed for the analysis in Li & Villani (2013), `cdcopula` that is mainly designed for Li (2013), and a utility package `flutils`. Other functions such as code to reproduce the R plots in the papers can also be found in the source code. The packages are self-contained and do not depend on external packages. The packages are extendable to other flexible models.

1. THE `MOVINGKNOTS` PACKAGE

The `movingknots` package is written in R and is mainly used for modeling regression surface with splines (Li & Villani, 2013) where the locations of basis functions are treated as unknown estimated parameters.

1.1. **The model module.**

The model used inLi & Villani (2013) is a Gaussian model but it is possible to extend it to the generalized linear model framework. We just briefly describe the core functions for the Gaussian model.

Feng Li (`feng.li@stat.su.se`): Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.

- `make.knots()`
  Generating various types basis functions for both additive splines and interactive splines.
- `d.matrix()`
  Efficient implementation for calculating the design matrix of a spline model with complicated basis functions.
- `linear_logpost()`
  The conditional and joint log posterior function.
- `linear_gradhess()`
  The conditional gradient for the posterior with respect to the parameters (knots locations, shrinkages, covariates coefficients and covariance matrix). Great efforts have been done to implement the calculations efficiently. For details on how to handle the sparsity problem with big design matrix, see the `etc` directory in the source code.
  - `delta.xi()`
    Efficient implementation of the gradient for the design matrix with respect to knot locations for various basis functions.
- `linear_IWishart()`
  Calculating degrees of freedom and location matrix in the inverse Wishart distribution when a conjugate prior is used.
- `linear_post4coef()`
  Direct sampling of the coefficients from multivariate normal distribution when the coefficients can be integrated out analytically.

1.2. **The MCMC module.**

- `MovingKnots_MCMC()`
  A Gibbs sampler where the Metropolis–Hastings algorithm is used in each Gibbs step. The sampler allows the model parameters to be updated sequentially, jointly, or by blocks.
- `MHPropMain()`
  The main function for Metropolis–Hastings algorithm where the random walk Metropolis and the Metropolis–Hastings algorithm with an integrated finite-step Newton algorithm.
  - `MHPropWithIWishart()`
    A sub-function for the Metropolis-Hastings algorithm that the proposal density is from inverse Wishart distribution.
  - `MHPropWithKStepNewtonMove()`
    A sub-function for the Metropolis-Hastings algorithm that the proposal density is from multivariate student's t distribution where the proposal mode and covariance matrix are from finite-step Newton iteration.
    - `KStepNewtonMove()`
      Finite-step Newton iteration for the log posterior of a model.
  - `RandomWalkMetropolis()`
    Random walk Metropolis algorithm.
- `MCMC.trajectory()`

The summary of posterior inference including posterior means, standard deviation, acceptance probability during the MCMC evaluation, and inefficiency factor that monitors the MCMC efficiency periodically.

1.3. **The DGP module.**

- DGP.hwang(), DGP.surface()
  Data generating processes for different surfaces.

1.4. **The model evaluation module.**

- `FitDiagnosis()`, `FitDiagnosis.hwang()`
  Model diagnosis procedures based on LOSS functions, Kullback–Leibler divergence and L_2 distance criteria.
- `PostPredIF()`
  The inefficiency factor for the out-of-sample predictive surface.
- `LPDS()`
  The log predictive density score with numerical standard errors for the cross-validation of the model.

2. THE CDCOPULA PACKAGE

The `cdcopula` package is mainly designed for modeling covariate-dependent copula models in Li (2013). A bash script `inst/run/CplRun` in the source file can run the MCMC with cross-validation and prediction in parallel non-interactive mode. The user only needs to supply a model configuration file, see `inst/config/config-main-sp100-sp600.R` file for the configuration example.

2.1. **The model module.**

- `kendalltau()`, `kendalltauGrad()`, `kendalltauInv()`, `kendalltauTabular()`
  Computes the analytical Kendall's τ, the gradient of the Kendall's τ and the inverse of Kendall's τ in a copula. The inverse of Kendall's can be evaluated numerically or with a dictionary-lookup algorithm.
- `logCplLik()`, `logCplGrad()`
  Computes the copula likelihood and its gradient with respect to copula parameters.
  - `HessApprox()`
    Approximating the Hessian matrix based on the outer-product method. This is used when the Hessian is difficulty to obtain, or unstable.
- `MargiModel()`, `MargiModelGrad()`
  Specify marginal models and their gradients with respect to the marginal parameters. The marginal models can be retrieved from any other univariate models.
- `logPriors()`, `logPriorsGradHess()`
  The prior specification of the copula model and its gradient and Hessian matrix for the prior distribution.

  - `any2any()`
    A function that automatically specifies priors in the intercept for covariate dependence structure, see Li (2013) for details. The strategy is that firstly puts prior information on the model parameters (e.g. τ and λ_L in the Joe-Clayton copula), and then derive the implied prior on the intercept β_0 under the assumption that the covariates are at their means..
- `logPost()`, `logPostOptim()`
  The function `logPost()` computes the log posterior of the copula model. Bayesian inference and the classical two-stage approach can be applied directly to this function. The function `logPostOptim()` is used to optimize the initial values for the MCMC.
- `logPredDens()`, `logPredDensScore()`
  Evaluating the log predictive likelihood and log predictive density score based on MCMC posterior inference.

2.2. **The MCMC module.** The MCMC part of this package allows evaluating the log predictive density score with cross-validation in parallel. See the configuration files for detailed information. The Metropolis-Hastings with tailored proposal based on finite-step Newton approximation is the analogue of Li et al. (2010) and Li et al. (2011).

- `CplMain()`
  The Main file for MCMC of the copula model that performs the MCMC.
- `GNewtonMove()`
  The generalized Newton method that allows the dimension of the proposal draws to change in the Metropolis-Hastings algorithm.
- `MHWithGNewtonMove()`
  Metropolis–Hastings algorithm with generalized Newton method and variable selection integrated.
- `CplMCMC.summary()`
  The posterior summary for the copula model including the detailed summary report for the marginal models and the copula model.

2.3. **The simulation module.**

- `DGPCpl()`
  Copula model data generating process.
- `ruCpl()`
  Random variable generators for common copulas, Joe-Clayton copula, Gaussian copula, multivariate t copula, fgm copula, gumbel copula, etc. The empirical Kendall's τ and theoretical Kendall's τ are also reported.
- `uCpl(), cCpl()`
  Compute the copula function and copula density with given margins.
- `hatCpl()`
  Empirical copula function estimation.
- `emltdc()`
  Empirical lower tail-dependence coefficient.

- `u2qtl()`
  Transform percentile to quantile according to marginal densities.

2.4. **The prediction and model evaluation module.** This part is same as the one in `movingknots` package in section 1.4.

3. THE FLUTILS PACKAGE

The `flutils` package is the utility package that includes R functions which are not available in R base packages or is not implemented efficiently elsewhere. We briefly describe some functions here that are related to the thesis. Other functions with general purposes are available and are documented in the package.

3.1. Distribution functions.

- `psplitt()`, `dsplitt()`, `rsplitt()`
  Density and distribution functions and random number generators for the split-t distributionused in Li et al. (2010) and Li (2013)
- `pbeta2()`,`dbeta2()`, `rbeta2()`, `plnorm2()` `dlnorm2()`, `rlnorm2()`
  Density and distribution functions and random number generators for beta and log-normal distributions are reparametrized in terms of their mean and variance used in Li (2013).
- `splitt.mean()`, `splitt.var()`, `splitt.skewness()`, `splitt.kurtosis()`
  The mean, variance, skewness and kurtosis for the split-t distribution.
- `piwishart()`, `diwishart()`, `riwishart()`
  Density and distribution functions and random number generators for the inverse Wishart distribution used inLi & Villani (2013).

3.2. MCMC functions.

- `ineff()`
  Inefficiency factor of a given MCMC chain.
- `parLinkFun()`, `parMeanFun()`, `parMeanFunGrad()`
  Functions that calculate the linear predictors, mean functions and their gradients for common link functions. And the generalized logit link, generalized log link. The conditional link functions are also considered in the function's implementation.

3.3. Special mathematical functions.

- `K.X()`, `K()`
  Efficient implementation of commutation matrix multiplications.
- `ghypergeo()`
  The generalized hypergeometric function ${}_pF_q$ for real numbers.
- `pochhammer()`
  The series expansion of Pochhammer symbol.
- `mvgamma()`
  The multivariate gamma function.

- `rdist()`
  Fast calculation of Euclidean distance matrix between two matrices.
- `ibeta()`
  Incomplete beta function and regularized incomplete beta function.
- `harmonic()`
  The harmonic numbers.
- `zeta()`
  The Riemann zeta function.
- `rps()`
  Generate random variable from a positive stable distribution.

3.4. **Data manipulate functions.**

- `stock2covariates()`
  Constructing sensible covariates from stock market data. The data source is from Yahoo Finance and the output format are suitable for the data used in Li et al. (2010) and Li (2013).
- `StdData()`
  Data standardization that standardizes a vector or colums of a matrix to e.g. mean zero and unit standard deviation for the MCMC and the inverse operation to transform back in the prediction and interpretation phase.
- `data.partition()`
  The data partitioning procedure used in cross-validation with methods like, *systematic*, *random*, *ordered* and for the special case of time series.

4. Notes on improving computing performance

The performance of the code is relatively fast in our real data application. Here we consider a few situations that can speed up R in general.

R base package `compiler` can compile R functions into machine code which speeds up the calculations. The function `sourceDir()` in package `flutils` can compile and load R functions in a very flexible way. Compiling R from source with fast compiler and BLAS (Basic Linear Algebra Subprograms) library, e.g. Intel compiler and Intel Math Kernel Library (MKL) can significantly speed up the matrix calculations.

Cross-validation is always costly and one may consider using parallel computing with R. A simple approach is the explicit parallelism where within each fold the MCMC are run sequentially. This can be achieved via the `parallel` package in R. A complete example using R's default parallelism scheme is documented in the file `inst/bin/CplRun` in the package `cdcopula`. For information of other types on parallelism with R, see the document *CRAN Task View: High-Performance and Parallel Computing with R*.

5. Implementing matrix derivatives

The efficient MCMC scheme for flexible Bayesian density estimation requires analytical expressions for the gradient and Hessian of the log posterior with respect to the parameters in the model. The appendices in the papers include the results for each models. The models included

in the thesis can be extended by providing the gradient and Hessian for the new model. Common rules for matrix derivatives are available in Lütkepohl (1996). However great care should be taken when dealing with big sparse matrices where one should avoid direct calculations which are memory-intensive and costly in CPU time. Sparse methods like those in the `Matrix` package in R should be used for that purpose. Checking the results with numerical gradients and Hessians can help to debug errors in coding.

REFERENCES

LI, F. (2013). Modeling covariate-contingent correlation and tail-dependence with copulas. *Manuscript* .

LI, F. & VILLANI, M. (2013). Efficient Bayesian multivariate surface regression. *Scandinavian Journal of Statistics* **in press**.

LI, F., VILLANI, M. & KOHN, R. (2010). Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference* **140**, 3638–3654.

LI, F., VILLANI, M. & KOHN, R. (2011). Modeling conditional densities using finite smooth mixtures. In *Mixtures: estimation and applications*, K. Mengersen, C. Robert & M. Titterington, eds. John Wiley & Sons Inc, Chichester, pp. 123–144.

LÜTKEPOHL, H. (1996). *Handbook of matrices*. John Wiley & Sons, Chichester.