# 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 $\theta$ combines data information $y$ with prior beliefs about $\theta$ 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 $\theta$ 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 $\theta$ 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 $\theta$ 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 $\omega_k$ 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 $\mu_k$ 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

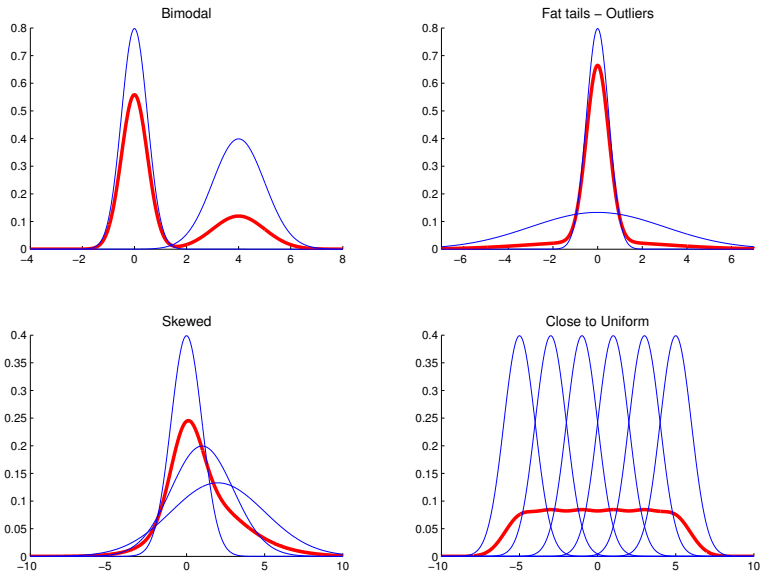


**Figure 1.1:** Using mixture of normal densities (thin lines) to mimic a flexible density (bold line)

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 $\sigma^2$.

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 $\xi_i$ is the location of $i$:th basis function, often referred to as a knot. Each basis function is defined by a knot $\xi_i$ 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 $\tau$

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

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