TI 2013-060/III
Tinbergen Institute Discussion Paper

Censored Posterior and Predictive Likelihood in Left-Tail Prediction for Accurate Value at Risk Estimation

Lukasz Gatarek[1,4]
Lennart Hoogerheide[2,4]
Koen Hooning[3]
Herman K. van Dijk [1,2,4]

[1] *Erasmus School of Economics, Erasmus University Rotterdam;*
[2] *Faculty of Economics and Business Administration, VU University Amsterdam;*
[3] *TU Delft;*
[4] *Tinbergen Institute.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Censored posterior and predictive likelihood in left-tail prediction for accurate Value at Risk estimation ☆

Lukasz T. Gatarek[a], Lennart F. Hoogerheide[b], Koen Hooning[c], Herman K. van Dijk[d]

[a] *Econometric Institute and Tinbergen Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands*
[b] *Department of Econometrics and Tinbergen Institute, Vrije Universiteit Amsterdam, The Netherlands*
[c] *Delft University of Technology, The Netherlands*
[d] *Econometric Institute and Tinbergen Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Vrije Universiteit Amsterdam, The Netherlands*

Abstract

Accurate prediction of risk measures such as Value at Risk (VaR) and Expected Shortfall (ES) requires precise estimation of the tail of the predictive distribution. Two novel concepts are introduced that offer a specific focus on this part of the predictive density: the censored posterior, a density for the model parameters that results if we replace the likelihood by a so-called censored likelihood in the posterior; and the censored predictive likelihood, which is used for Model Averaging for forecast combination. We perform extensive experiments involving simulated and empirical data. Our results show the ability of these new approaches to outperform the standard posterior and traditional Bayesian Model Averaging techniques in applications of Value-at-Risk prediction in GARCH models.

JEL classification: C11, C15, C22, C51, C53, C58, G17

Keywords: censored likelihood, censored posterior, censored predictive likelihood, Model Averaging, Value at Risk, Metropolis-Hastings algorithm.

☆The concepts of this paper were presented at the European Seminar on Bayesian Econometrics (ESOBE) 2011 in Brussels, Belgium, and the Macromodels International Conference 2011 in Poznan, Poland. We are indebted to several participants at these meetings for constructive comments. Further, we are very grateful to Dennis Fok and Richard Paap for multiple very useful remarks on an earlier version of this paper. Lennart Hoogerheide was supported by Veni grant 451-09-028 of the Dutch National Science Foundation (NWO).

Email addresses: gatarek@tlen.pl (Lukasz T. Gatarek), l.f.hoogerheide@vu.nl (Lennart F. Hoogerheide), k.m.hooning@student.tudelft.nl (Koen Hooning), hkvandijk@ese.eur.nl (Herman K. van Dijk)

Preprint submitted to the Tinbergen Institute Discussion Paper series *March 6, 2014*

1. Introduction

In this paper we consider the issue of accurate estimation of the left tail of the predictive distribution, which is important for obtaining correct forecasts of risk measures such as Value at Risk (VaR) and Expected Shortfall (ES). Our benchmark is the Bayesian approach, which allows us to incorporate parameter uncertainty and to combine forecasts from multiple models using Bayesian Model Averaging (BMA). Typically, one has no specific focus on the left tail of the distribution of returns during the estimation of the posterior distributions of the model parameters or during the construction of model weights in cases of model combinations. The usual likelihood weights the observations in the tail of the distribution and those in the middle part equally. In this paper we present two novel measures that offer a specific focus on the left tail of the distribution during the estimation of model parameters and the combination of models: the censored posterior, which we define as the density of the model parameters that results if we replace the the likelihood by the censored likelihood in the posterior; and the censored predictive likelihood, which is a censored extension of the predictive likelihood. Note that the predictive likelihood is usually defined as the marginal likelihood when the first subset of the data is used for updating the prior. We do *not* define the censored likelihood as the likelihood for a censored data set, where all observations lying outside a particular area of interest are censored. We define the censored likelihood as the product of the conditional densities of the censored observations given the past observations (where only observations occurring in the area of interest such as the left tail remain uncensored), where all the past observations remain uncensored. We propose this specification for two reasons. First, the purpose is to improve the left-tail prediction based on the actually observed past observations. By censoring the past observations we would lose valuable information. Second, it would typically be much more difficult to compute the likelihood for censored data (where one would also condition on censored observations). This specification does imply that our censored posterior and censored predictive likelihood fall outside the framework of Bayesian statistics: the censored posterior and the censored predictive likelihood are *not* equal to the posterior and predictive likelihood for a censored data set, respectively. The censored likelihood has been used by Diks et al. (2011), but these authors only consider its use for testing the quality of (frequentist) left tail forecasts, without incorporating it in the estimation or combination of models. We perform extensive experiments, involving simulated and empirical data. Our results show the ability of the new measures to outperform standard posterior and traditional Bayesian Model Averaging techniques in applications of Value at Risk prediction in univariate GARCH models. Our approach is easily applied to different univariate time series models. Extension to multivariate time series models (for high-dimensional vectors of returns, for which the left tail of the distribution of portfolio returns may be considered) may require additional simulations, since the evaluation of the Cumulative Distribution Function (CDF) of the portfolio return is needed. The outline of this paper is as follows. In Section 2 we introduce the concept of the censored posterior. In Section 3 we consider Bayesian Model Averaging and introduce the concept of the censored predictive likelihood. In Section 4 we compare the performance of our proposed forecasts of percentiles in the left tail (i.e., Value at Risk forecasts) with traditional Bayesian forecasts for a large number of simulated data sets. In Section 5 we present a similar comparison for empirical data sets of well-known index returns. Section 6 concludes.

2. The censored posterior

Econometric models may be described by the joint probability distribution, known up to a parameter vector θ, of the random variables $y_{1:T} = \{y_1, \ldots, y_T\}$, where a set of T observations on these variables is available. Note that the typical element y_t may be a vector itself. Bayesian inference proceeds from the likelihood function $p(y_{1:T}|\theta)$, which is either the density of the data given the parameters in case of a continuous distribution or the probability function in case of a discrete distribution, and a prior density $p(\theta)$ reflecting prior beliefs on the parameters before the data set has been observed – see e.g., Hoogerheide et al. (2009). So, in the Bayesian approach the parameters θ are considered as random variables whose prior density $p(\theta)$ is updated with the information contained in the data, incorporated in the likelihood function $p(y_{1:T}|\theta)$, to obtain the posterior density of the parameters $p(\theta|y_{1:T})$. This process is formalized by Bayes' theorem, stating that the posterior density is given by:

$$p(\theta|y_{1:T}) = \frac{p(\theta)p(y_{1:T}|\theta)}{p(y_{1:T})}. \tag{1}$$

Note that this is merely a result of rewriting the identity $p(y_{1:T})p(\theta|y_{1:T}) = p(\theta)p(y_{1:T}|\theta)$, the two ways of decomposing the joint density $p(y_{1:T}, \theta)$ into a marginal and a conditional density. Equation (1) can be rewritten as

$$p(\theta|y_{1:T}) \propto p(\theta)p(y_{1:T}|\theta), \tag{2}$$

where the symbol $\propto$ means 'is proportional to', i.e., the left-hand side is equal to the right-hand side times a scaling constant $(1/p(y_{1:T}) = 1/\int p(\theta)p(y_{1:T}|\theta)d\theta)$ that does not depend on the parameters θ. That is, $p(\theta)p(y_{1:T}|\theta)$ is a kernel (=proportionality function) of the posterior density of θ, where this kernel merely has to be divided by a constant, the marginal likelihood $p(y_{1:T}) = \int p(\theta, y_{1:T})d\theta = \int p(y_{1:T}|\theta)p(\theta)d\theta$, in order to make it a proper (posterior) density. The marginal likelihood is the marginal density of the data $y_{1:T}$ after the parameters θ of the model have been integrated out with respect to their prior distribution. In Section 3 we consider how marginal likelihoods can be used for Bayesian Model Averaging (BMA), where the forecast distribution is a weighted average of the forecast distributions from different models.

In the likelihood

$$p(y_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t|y_1, \ldots, y_{t-1}, \theta) \tag{3}$$

and in the posterior density kernel in formulas (1) and (2) there is no specific focus on a particular region of interest $A_{1:T} = \{A_1, \ldots, A_T\}$ for the observations $y_{1:T} = \{y_1, \ldots, y_T\}$, where we can have $A_t = \{y_t | y_t \leq R_t\}$ for some (constant or time-varying) value R_t if we are interested in the left tail of the distribution of y_t. For this purpose we substitute the likelihood by a novel concept that we name the *censored likelihood*

$$p^{cs}(y_{1:T}|\theta) \equiv \prod_{t=1}^{T} p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta) \tag{4}$$

with $p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta)$ defined as the conditional density $p(\tilde{y}_t|y_1, \ldots, y_{t-1}, \theta)$ of the mixed continuous-discrete distribution for the censored variable $\tilde{y}_t$ defined as

$$\tilde{y}_t = \begin{cases} y_t & \text{if } y_t \in A_t, \\ C_t & \text{if } y_t \in A_t^C, \end{cases} \tag{5}$$

where A_t^C is the complement of A_t, where the value y_t is substituted by a value $C_t \in A_t^C$ if $y_t \in A_t^C$. So, the distribution of $\tilde{y}_t$ has a continuous density over A_t (i.e., the same continuous density as y_t) and a discrete probability $P(\tilde{y}_t = C_t|y_1, \ldots, y_{t-1}, \theta) = P(y_t \in A_t^C|y_1, \ldots, y_{t-1}, \theta)$. That is, $p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta)$ is given by:

$$\begin{aligned} p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta) \;\equiv\; & [p(y_t|y_1, \ldots, y_{t-1}, \theta)]^{I\{y_t \in A_t\}} \times \\ & \left[P(y_t \in A_t^C|y_1, \ldots, y_{t-1}, \theta)\right]^{I\{y_t \in A_t^C\}}, \qquad (6) \\ \\ = & \left[p(y_t|y_1, \ldots, y_{t-1}, \theta)\right]^{I\{y_t \in A_t\}} \times \\ & \left[\int_{y_t \in A_t^C} p(y_t|y_1, \ldots, y_{t-1}, \theta)\, dy_t\right]^{I\{y_t \in A_t^C\}}. \qquad (7) \end{aligned}$$

We assume that the original variable y_t has a continuous distribution in (7), otherwise the integral must obviously be replaced by a sum. Note that this censored likelihood is typically *not* equal to the likelihood in case of a data set where all values y_t in A_t^C are censored, since the conditional density $p(y_t|y_1, \ldots, y_{t-1}, \theta)$ depends on some of the observations $y_1, \ldots, y_{t-1}$, where observations y_s $(s = 1, \ldots, t-1)$ in A_s^C are *not* censored. Only if we would have $p(y_t|y_1, \ldots, y_{t-1}, \theta) = p(y_t|\theta)$, then the censored likelihood would be equal to the likelihood in case of a data set where all values y_t in A_t^C are censored. We propose this specification for two reasons. First, the purpose is to improve the left-tail prediction based on the actually observed past observations. By censoring the past observations we would lose valuable information. Second, it would typically be much more difficult to compute the likelihood for censored data (where one would also condition on censored observations).

Note that if y_t is one-dimensional, then it is straightforward to evaluate the integral $\int_{y_t \in A_t^C} p(y_t|y_1, \ldots, y_{t-1}, \theta)\, dy_t$ in (7), using either analytical or deterministic (quadrature) integration. If y_t is a high-dimensional vector, then simulation is required to evaluate this integral and the censored likelihood. In this paper we only consider examples where y_t is one-dimensional. In future research we will investigate the application to high-dimensional y_t – for example, a vector of returns in a large portfolio of stocks. We will consider the parallel implementation on graphics processing units (GPUs), for which the evaluation of many high-dimensional integrals is a natural application.

The density $p^{cs}(y_t|y_1,\dots,y_{t-1},\theta)$ in (7) is equal to the exponent of the censored likelihood score function of Diks et al. (2011), who consider Diebold-Mariano type tests for comparing the accuracy of two sequences of density forecasts $\hat{f}_t$ and $\hat{g}_t$. Diks et al. (2011) argue and show that the censored likelihood score function does not lead to biases toward densities with more probability mass in the region of interest A_t, unlike score functions that simply ignore the observations outside A_t. Moreover, Diks et al. (2011) find that the test based on the censored likelihood score function is typically more powerful than the test based on the conditional likelihood score function, where one considers the conditional density of y_t (conditionally upon the fact that $y_t \in A_t$). The latter finding is intuitively clear, since the censored likelihood contains more information than the conditional likelihood, as the censored likelihood also contains the information of how many observations occur outside A_t. For these reasons we consider the censored likelihood/posterior, rather than a conditional likelihood/posterior (or a likelihood/posterior where the observations outside A_t are simply ignored). Diks et al. (2011) also propose a smooth extension of the censored likelihood score function, where the indicator functions $I\{y_t \in A_t\}$ and $I\{y_t \in A_t^C\}$ in (7) are substituted by weight functions $w(y_t)$ and $1-w(y_t)$, taking values in the [0,1] interval. We leave these concepts of a *smoothly censored likelihood* and a *smoothly censored posterior* as topics for future research.

We define the novel concept of the censored posterior density as follows: a kernel of the censored posterior density $p^{cs}(\theta|y_{1:T})$ is obtained by multiplying the prior density with the censored likelihood:

$$p^{cs}(\theta|y_{1:T}) \propto p(\theta)p^{cs}(y_{1:T}|\theta). \tag{8}$$

That is, the censored posterior density is given by

$$p^{cs}(\theta|y_{1:T}) = \frac{p(\theta)p^{cs}(y_{1:T}|\theta)}{\int p(\theta)p^{cs}(y_{1:T}|\theta)d\theta}. \tag{9}$$

Typically, the censored posterior density $p^{cs}(\theta|y_{1:T})$ is a proper density in the same cases (i.e., under the same choices of the prior $p(\theta)$) where the posterior $p(\theta|y_{1:T})$ is a proper density (i.e., with finite integral $\int p(\theta)p^{cs}(y_{1:T}|\theta)d\theta < \infty$), as long as there are enough observations $y_t \in A_t$ that are not censored. In this paper, we consider proper (non-informative) prior distributions, which already ensures the properness of the censored posterior distributions in the models that we consider. The fact that the censored likelihood is not equal to the likelihood for a censored data set implies that our censored posterior falls outside the framework of Bayesian statistics: the censored posterior is *not* equal to the posterior for a censored data set.

In most models it is impossible to analytically evaluate the properties of interest of the censored posterior $p^{cs}(\theta|y_{1:T})$. Simulation is typically required. Several simulation methods can be used here. In this paper we use the independence chain Metropolis-Hastings method (Metropolis et al., 1953; Hastings, 1970), also known as the independent Metropolis-Hastings method, in order to evaluate the (censored) posterior density. For the candidate or proposal distribution we use Student's t-distribution around the mode with low degrees of freedom parameter to have fat tails, and with covariance matrix obtained by multiplying the 'standard choice' (minus the inverse Hessian) by a multiplication factor 1.5, which provides reasonable acceptance rates in our examples. Alternative methods include importance sampling (developed by Hammersley and Handscomb (1964) and introduced into econometrics and statistics by Kloek and Van Dijk (1978), and the random walk Metropolis(-Hastings) method of Metropolis et al. (1953). All these methods require only evaluations of the censored posterior density *kernel* in (8), so that the evaluation of the denominator in (9), $\int p(\theta)p^{cs}(y_{1:T}|\theta)d\theta$, is not required for investigating the censored posterior for a given model. In this paper, Student's t candidate distribution performs well in the sense of reasonably high acceptance rates in the independence chain Metropolis-Hastings method and reasonably low variance of the importance sampling weights (in the application of importance sampling for the evaluation of marginal and predictive likelihoods, which is discussed below). If Student's t-distribution would be a poor approximation of the posterior and lead to poor results – i.e., very low acceptance rates in the independence chain Metropolis-Hastings method, large or even infinite variance of the importance weights in importance sampling – then we recommend to change to the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) method of Hoogerheide et al. (2012a), which can be adopted to have a specific focus on the left tail by combining it with the algorithm of Hoogerheide and Van Dijk (2010).

In Bayesian analysis of a model with a regular, ‘uncensored’ posterior (based on the regular ‘uncensored’ likelihood), the predictive density of the variable y_{T+1}, given the data $y_{1:T} = \{y_1, \ldots, y_T\}$ up to time T, is given by:

$$p(y_{T+1}|y_{1:T}) = \int p(y_{T+1}|y_{1:T}, \theta) p(\theta|y_{1:T}) d\theta, \tag{10}$$

which is typically approximated by

$$p(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{j=1}^{N} p(y_{T+1}|y_{1:T}, \theta^{(j)}), \tag{11}$$

where $\theta^{(j)}$ $(j = 1, \ldots, N)$ are draws from the posterior $p(\theta|y_{1:T})$. In a similar fashion, in our case of a censored posterior we define the censored predictive density as follows:

$$p^{cs}(y_{T+1}|y_{1:T}) = \int p(y_{T+1}|y_{1:T}, \theta) p^{cs}(\theta|y_{1:T}) d\theta, \tag{12}$$

which is approximated by

$$p^{cs}(y_{T+1}|y_{1:T}) \approx \frac{1}{N} \sum_{j=1}^{N} p(y_{T+1}|y_{1:T}, \theta^{(j)}), \tag{13}$$

where $\theta^{(j)}$ $(j = 1, \ldots, N)$ are draws from the censored posterior $p^{cs}(\theta|y_{1:T})$.

3. Bayesian Model Averaging and the censored predictive likelihood

Since the seminal article of Bates and Granger (1969) several papers have shown that combinations of forecasts can outperform individual forecasts in terms of loss functions. For example, Stock and Watson (2004) find that for predicting output growth in seven countries forecast combinations generally perform better than forecasts based on single models. Marcellino (2004) has extended this analysis to a large European data set with broadly the same conclusion. In a Bayesian framework, Madigan and Raftery (1994) revitalize the concept of Bayesian Model Averaging (BMA). Geweke and Whiteman (2006) propose a BMA scheme based on the idea that a model is as good as its predictions, using predictive likelihoods instead of marginal likelihoods. Billio et al. (2013) make use of a Bayesian combination scheme with time-varying model weights.

In the case of Bayesian Model Averaging, where one considers m models M_i ($i = 1, \ldots, m$), the predictive density of the variable y_{T+1}, given the data $y_{1:T} = \{y_1, \ldots, y_T\}$ up to time T, is computed by averaging over the conditional predictive densities:

$$p(y_{T+1}|y_{1:T}) = \sum_{i=1}^{m} p(y_{T+1}|y_{1:T}, M_i)\ P(M_i|y_{1:T}), \tag{14}$$

where $p(y_{T+1}|y_{1:T}, M_i)$ is the conditional predictive density given data $y_{1:T}$ and model M_i, and $P(M_i|y_{1:T})$ is the posterior probability for model M_i. The conditional predictive density given data $y_{1:T}$ and model M_i is

$$p(y_{T+1}|y_{1:T}, M_i) = \int p(y_{T+1}|y_{1:T}, \theta_i, M_i)\ p(\theta_i|y_{1:T}, M_i)\ d\theta_i, \tag{15}$$

where θ_i is the parameter vector in model M_i. The posterior probability for model M_i is

$$P(M_i|y_{1:T}) = \frac{p(y_{1:T}|M_i)P(M_i)}{\sum_{k=1}^{m} p(y_{1:T}|M_k)P(M_k)}, \tag{16}$$

where $P(M_i)$ is the prior probability for model M_i and $p(y_{1:T}|M_i)$ is the marginal likelihood for model M_i given by

$$p(y_{1:T}|M_i) = \int p(y_{1:T}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i \tag{17}$$

with $p(\theta_i|M_i)$ the prior density for the parameters θ_i in model M_i. The integral in equation (17) can be evaluated analytically in the case of linear models, but typically not for more complex model specifications. Ardia et al. (2012) provide a comparative study of several Monte Carlo methods for marginal likelihood evaluation, and find that the importance sampling estimator is a computationally efficient and accurate estimator (on the condition that the importance density provides a reasonable approximation of the posterior). In this paper we use importance sampling, where the importance density is the same as the candidate density in the independence chain Metropolis-Hastings method. If there is no *a priori* preference for one of the models, then one typically specifies the prior probabilities as $P(M_i) = 1/m$ ($i = 1, \ldots, m$).

If one possesses highly informative, tight priors $p(\theta_i|M_i)$, then BMA will result in well-defined marginal likelihoods in (17) and usable posterior model probabilities in (16). If not, then BMA may result in unreliable posterior model probabilities. The reason of this phenomenon is known in the statistical literature as Bartlett's paradox, see Lindley

(1957) and Bartlett (1957). Bartlett's paradox may be interpreted as the fact that if we spread too much prior probability mass in the prior of model i, $p(\theta_i|M_i)$, over 'silly' values, i.e., we make the prior very wide as compared to the prior densities in the other models $p(\theta_k|M_k)$ $(k \neq i)$, we can typically make the marginal likelihood $p(y_{1:T}|M_i)$ in (17) and the posterior probability $P(M_i|y_{1:T})$ in (16) as small as we want, independent of the information in the data $y_{1:T}$. Therefore, another method is needed to compute posterior model probabilities if no highly informative priors are available for all the models under consideration. One alternative is to use the predictive likelihood instead of the marginal likelihood.

3.1. Bayesian Model Averaging using the predictive likelihood

If one desires to perform BMA in case of weakly informative or even improper prior densities, then one possible approach is to make use of the predictive likelihood, see Gelfand and Dey (1994) and Eklund and Karlsson (2007), who provide an overview of several definitions of predictive likelihoods including the specifications corresponding with the fractional Bayes factor of O'Hagan (1995) and the intrinsic Bayes factor of Berger and Pericchi (1996). We use the specification where the predictive likelihood for model M_i is given by

$$p(y_{r+1:T}|y_{1:r}, M_i) = \int p(y_{r+1:T}|\theta_i, y_{1:r}, M_i)\; p(\theta_i|y_{1:r}, M_i)\; d\theta_i, \tag{18}$$

with *training sample* $y_{1:r} = \{y_1, \ldots, y_r\}$ and *hold-out sample* $y_{r+1:T} = (y_{r+1}, \ldots, y_T)$. The predictive likelihood in (18) can be considered as a marginal likelihood where the posterior density $p(\theta_i|y_{1:r}, M_i)$ after the first r observations (forming the training sample) plays the role of the prior density, and where the observations in the hold-out sample $y_{r+1:T}$ play the role of 'the data set'. Bayes' rule implies that this posterior density $p(\theta_i|y_{1:r}, M_i)$ is given by

$$p(\theta_i|y_{1:r}, M_i) = \frac{p(y_{1:r}|\theta_i, M_i)p(\theta_i|M_i)}{\int p(y_{1:r}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}. \tag{19}$$

Substituting (19) into (18) yields

$$p(y_{r+1:T}|y_{1:r}, M_i) = \frac{\int p(y_{r+1:T}|\theta_i, y_{1:r}, M_i)p(y_{1:r}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}{\int p(y_{1:r}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}, \tag{20}$$

which is equal to

$$p(y_{r+1:T}|y_{1:r}, M_i) = \frac{\int p(y_{1:T}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}{\int p(y_{1:r}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}. \tag{21}$$

From (21) it is clear that the predictive likelihood is simply given by the ratio of the marginal likelihood of all observations over the marginal likelihood for the first r observations in the training sample. Roughly stated, the first r observations are used to delete the completely silly values from the original non-informative prior $p(\theta_i|M_i)$. The posterior density $p(\theta_i|y_{1:r}, M_i)$ after the training sample should not be crucially affected by the choice between different non-informative priors $p(\theta_i|M_i)$. Using the predictive likelihood the posterior model probabilities in the BMA can now be defined as

$$P(M_i|y_{1:T}) = \frac{p(y_{r+1:T}|y_{1:r}, M_i)P(M_i)}{\sum_{k=1}^{m} p(y_{r+1:T}|y_{1:r}, M_k)P(M_k)}. \tag{22}$$

For the evaluation of the two integrals, the two marginal likelihoods, in the numerator and denominator of (21) we make use of importance sampling, where for each case a fat-tailed Student's t-density around the mode is used as the importance density.

One remaining issue when using the predictive likelihood is how to divide the data in a training sample $y_{1:r}$ and a hold-out sample $y_{r+1:t}$. Gelfand and Dey (1994) and Eklund and Karlsson (2007) give an overview of different options. We simply use the equal sample split $r = T/2$, which should assure that both the training and the hold-out sample contain enough data to obtain reliable posterior model probabilities. A sensitivity analysis with respect to the choice of r is left as a topic for further research. Alternative choices include small values of r for which the hold-out sample $y_{r+1:T}$ is as large as possible and small values of $T - r$ for which the training sample contains almost the same information as the whole data set $y_{1:T}$.

3.2. Model Averaging using the censored predictive likelihood

In BMA using either the marginal likelihood or the predictive likelihood, the entire predictive density $p(y_{T+1}|y_{1:T})$ is considered as equally important. There is no particular focus on a particular part of the predictive density such as the left tail. For this purpose, we propose a novel concept that we name the *censored predictive likelihood*, which results by substituting the likelihood $p(y_{r+1:T}|\theta_i, y_{1:r}, M_i)$ in (18) with the censored likelihood

$$p^{cs}(y_{r+1:T}|\theta_i, y_{1:r}, M_i) = \prod_{t=r+1}^{T} p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta) \tag{23}$$

with $p^{cs}(y_t|y_1, \ldots, y_{t-1}, \theta)$ in (7). That is, the censored predictive likelihood is given by:

$$p^{cs}(y_{r+1:T}|y_{1:r}, M_i) = \int p^{cs}(y_{r+1:T}|\theta_i, y_{1:r}, M_i)\ p(\theta_i|y_{1:r}, M_i)\ d\theta_i. \tag{24}$$

Substituting (19) into (24) yields

$$p^{cs}(y_{r+1:T}|y_{1:r}, M_i) = \frac{\int p^{cs}(y_{r+1:T}|\theta_i, y_{1:r}, M_i)\ p(y_{1:r}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i}{\int p(y_{1:r}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i}. \tag{25}$$

We evaluate the two integrals in (25) – the numerator and the marginal likelihood in the denominator – by importance sampling using fat-tailed Student't t importance densities. The fact that the censored likelihood is not equal to the likelihood for a censored data set implies that our censored predictive likelihood falls outside the framework of Bayesian statistics: the censored predictive likelihood is *not* equal to the predictive likelihood for a censored data set.

Also note that we do *not* define the *censored predictive likelihood* as

$$\begin{aligned} p^{cs}(y_{r+1:T}|y_{1:r}, M_i) &= \int p^{cs}(y_{r+1:T}|\theta_i, y_{1:r}, M_i)\ p^{cs}(\theta_i|y_{1:r}, M_i)\ d\theta_i && (26) \\ &= \frac{\int p^{cs}(y_{r+1:T}|\theta_i, y_{1:r}, M_i)\ p^{cs}(y_{1:r}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i}{\int p^{cs}(y_{1:r}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i} && (27) \\ &= \frac{\int p^{cs}(y_{1:T}|\theta_i, y_{1:r}, M_i)\ p(\theta_i|M_i)\ d\theta_i}{\int p^{cs}(y_{1:r}|\theta_i, M_i)\ p(\theta_i|M_i)\ d\theta_i} && (28) \end{aligned}$$

where the posterior $p(\theta_i|y_{1:r}, M_i)$ in (24) is substituted by the censored posterior $p^{cs}(\theta_i|y_{1:r}, M_i)$, for several reasons. First, one of the purposes of the training data is to make sure that the model probabilities do not crucially depend on which non-informative priors are specified. For this purpose, there is no need to prefer the censored posterior over the posterior. In fact, the posterior is arguably more capable to 'delete' the effect of 'silly' parameter values included in the non-informative prior $p(\theta_i|M_i)$ than the censored posterior (for a given data window $y_{1:r}$). Second, the censored likelihood $p^{cs}(y_{1:r}|\theta_i, M_i)$ would be included in both the numerator and denominator of (27), so that it is anyway canceled in a certain sense; that is, the use of the censored likelihood $p^{cs}(y_{1:r}|\theta_i, M_i)$ (instead of the likelihood $p(y_{1:r}|\theta_i, M_i)$) would not increase the focus on a particular part of the predictive density such as the left tail.

Using the censored predictive likelihood we define the model probabilities in model averaging as:

$$P^{cs}(M_i|y_{1:T}) = \frac{p^{cs}(y_{r+1:T}|y_{1:r}, M_i)P(M_i)}{\sum_{k=1}^{m} p^{cs}(y_{r+1:T}|y_{1:r}, M_k)P(M_k)}. \tag{29}$$

If one has specified highly informative priors for each model that is included in the model averaging, and if one is particularly interested in a particular part of the predictive density such as the left tail, then one can obviously also make use of the *censored marginal likelihood*, defined as

$$p^{cs}(y_{1:T}|M_i) = \int p^{cs}(y_{1:T}|\theta_i, M_i)\; p(\theta_i|M_i)\; d\theta_i, \tag{30}$$

which is simply the censored predictive likelihood in (24) with $r = 0$. On the other hand, for $r = T$ we have the case of equal model weights.

The concepts of the censored posterior and the censored predictive likelihood imply that we have $2 \times 2 = 4$ alternatives, if we perform model averaging with the predictive likelihood or the censored predictive likelihood. First, one needs to choose between (1) the (uncensored) posterior and (2) the censored posterior. Second, one needs to choose between (a) the (uncensored) predictive likelihood and (b) the censored predictive likelihood. In the application to simulated data sets and in the first empirical application, we will focus on the approaches (1+a) and (2+b), in which censoring is not used or fully used. In the second empirical application we will compare (1+a) and (1+b), where we specifically focus on the effect of censoring in case of the predictive likelihood.

4. Application: simulated data sets from GARCH(2,2) model

In order to investigate the quality of our proposed methods in an application involving left tail prediction, we perform a very extensive simulation experiment, where we analyze

$S = 100$ data sets of $\tilde{T} = 1000$ observations that are simulated from the GARCH(2,2) model (see Bollerslev (1986))

$$y_t = \sigma_t \, \varepsilon_t \quad (t = 1, \ldots, \tilde{T}), \tag{31}$$

$$\sigma_t^2 = \beta_0 + \alpha_1 y_{t-1}^2 + \alpha_2 y_{t-2}^2 + \beta_1 \, \sigma_{t-1}^2 + \beta_2 \, \sigma_{t-2}^2, \tag{32}$$

where the i.i.d. ε_t have a Student's t-distribution with ν degrees of freedom. For the true parameters of the GARCH(2,2) model in (31)-(32) we specify $\beta_0 = \beta_1 = \beta_2 = \alpha_1 = \alpha_2 = 0.07$ and $\nu = 8$ degrees of freedom. We estimate this GARCH(2,2) model, as well as GARCH(1,1), GARCH(1,2), and GARCH(2,1) models that result by setting $\beta_2 = 0$ and/or $\alpha_2 = 0$ in (32). In the Bayesian estimation of each model, we specify non-informative, proper Gaussian prior densities $p(\theta_i|M_i)$ $(i = 1, 2, 3, 4)$ and equal prior model probabilities $P(M_i) = 1/4$. For each estimation, 10000 candidate draws are used in the Metropolis-Hastings algorithm and importance sampling; in the Metropolis-Hastings algorithm the first 1000 draws are discarded as a burn-in.

For each simulated data set, we consider one-period-ahead prediction of the 1%, 2%, ..., 10% percentiles, i.e., the one-period-ahead 99%, 98%, ..., 90% Value at Risk. We use a moving estimation window of $T = 500$ observations, i.e., the data set used for predicting the percentiles of y_{s+500} $(s = 1, 2, \ldots, 500)$ is given by $\{y_s, y_{s+1}, \ldots, y_{s+499}\}$. This implies that for each estimation window both the training sample $\{y_s, y_{s+1}, \ldots, y_{s+249}\}$ and hold-out sample $\{y_{s+250}, y_{s+251}, \ldots, y_{s+499}\}$ in the predictive likelihood approach contain 250 observations.

Note that the simulation experiment requires 200000 estimations (100 data sets × 500 estimation windows × 4 models) for both the posteriors and censored posteriors, and for the predictive likelihoods another 200000 estimations are required for the training sample. Therefore, even though computationally efficient C++ code has been used, an enormous amount of computing time was required. In future research, we will consider the parallel implementation on graphics processing units (GPUs), for which this experiment with many independent simulations and estimations is a natural application.

As mentioned before, we use a Student's t-distribution (with low degrees of freedom) in the independence chain Metropolis-Hastings method for the evaluation of the (censored) posterior and in the importance sampling method for the evaluation of the marginal likelihoods (and the denominator of the censored predictive likelihood in (25)), leading to reasonably high acceptance rates and reasonably low variances of the importance sampling weights.

For the censoring we consider the region of interest $A_t = \{y_t | y_t \leq R_t\}$ where the constant value $R_t = R$ is either the 20% or 30% percentile of the estimation window in case of the censored posterior, or the 20% or 30% percentile of the hold-out sample in case of the censored predictive likelihood. These choices ensure that a reasonable number of observation is uncensored: 100 or 150 observations in the estimation window for the censored posterior, 50 or 75 observations in the hold-out sample for the censored predictive likelihood. We consider two choices to investigate the sensitivity of our results with respect to the particular boundary value R_t. As an alternative one could choose R_t time-varying, such as a percentile of the conditional distribution of y_t given $\{y_1, \ldots, y_{t-1}\}$ in a certain model M. One disadvantage of that approach would be that the choice for a particular model M (upon which the conditional distribution's percentile is based) may affect the performance of the different models. We leave this as a topic for future research.

To investigate the quality of the predicted 1%, 2%, ..., 10% percentiles from the different models and the different predictive BMA approaches, we consider a simple measure, the root mean squared error (RMSE), where for each simulated data set the 'error' is the difference between the observed fraction of 'violations' (realizations y_t that are more negative than the predicted percentile) and the desired value of 1%, 2%, ..., 10%. The mean is taken over the $S = 100$ data sets (if a particular percentile is considered), or over both the $S = 100$ data sets and the 10 percentiles. Alternatively, we could have considered a Diebold-Mariano type test using the censored likelihood based scoring rule of Diks et al. (2011). This will be considered in the second empirical application. Table 1 shows the results. We draw the following conclusions. First, obviously, the GARCH(2,2) model has the lowest RMSE, since the data are simulated from a GARCH(2,2) model. For the GARCH(2,2) model, the (uncensored) posterior yields slightly lower RMSE (on average over the 10 percentiles) than the censored posterior. Making use of the independence across the $S = 100$ simulated data sets, we perform a one-sided t-test to assess whether in the GARCH(2,2) model the uncensored posterior performs significantly better than the censored posterior (on average over the 10 percentiles), where we bootstrap the distribution under the null hypothesis of equal performance (see Efron and Tibshirani (1993)). The p-value is 0.0405 (0.0614) in the test whether the (uncensored) posterior performs better than the censored posterior using the 20% (30%) percentile as the boundary value, so that the performance of the uncensored posterior is significantly better at a significance level of 5% (10%). In the true model, it is optimal to fully use all observations in order to estimate the parameters as accurately as possible. However, in practice it is not a priori known what the true model is for an empirical data set. Moreover, one often faces the situation where the true model is not included in the set of models under consideration. Second, disappointingly, the censored posterior yields similar results to the (uncensored) posterior for each false model (GARCH(1,1), GARCH(1,2), and GARCH(2,1)). The RMSEs are close; the differences are not significant in a t-test. Third, a very interesting result is found for the model averaging approaches. Here the censored posterior and censored predictive likelihood perform much better than the (uncensored) posterior and (uncensored) predictive likelihood. That is, the focus on the left tail during the estimation and combination of the models clearly pays off in terms of a higher quality of the left tail of the predictive density. The p-value is 0.0001 (0.0001) in the test whether the censored posterior and censored predictive likelihood using the 20% (30%) percentile as the boundary value performs better than the (uncensored) posterior and the (uncensored) predictive likelihood, so that the performance of the censored posterior and censored predictive likelihood is significantly better at a significance level of 0.01%.

Fourth, if we would use the GARCH(2,2) model with the true parameters, rather than the estimated parameters, then the number of violations (for the $100p\%$ percentile) would have a binomial distribution with 500 trials and probability of 'success' equal to p. Therefore the RMSE of the fraction of violations would be equal to the standard deviation $\sqrt{p(1-p)/500}$, which is an increasing function of p. A comparison of the results for the estimated GARCH(2,2) model and the true GARCH(2,2) model (in the bottom row of Table 1) shows that the harmful effect of the estimation errors on the quality of (the left tail of) the predictive density is huge.

Fifth, just like for the true model, for each estimated model and for each Model

Averaging approach, the RMSE is typically larger for larger percentiles. To perform a fair comparison of the *relative* performance between different percentiles, we consider the ratios of the RMSE and the RMSE under the true model. These ratios, reported by Table 2, show that for the false models (GARCH(1,1), GARCH(1,2), and GARCH(2,1)) the performance is worse for lower percentiles that are deeper in the left tail. This also holds true for the uncensored BMA approach. Arguably, the latter is caused by the problem that the predictive likelihood may still suffer from Bartlett's paradox, as the training sample may be too small; the data set $y_{1:250}$ may be too small to yield a good predictive density for $y_{251:500}$ in the GARCH(2,2) model. In future research we will perform similar experiments with larger estimation windows and larger training samples. On the other hand, the performance is approximately equally good for the different percentiles (including the deeper left tail) for the censored model averaging approach; this stresses that the focus on the left tail during the estimation and combination of the models is very beneficial. This conclusion can also be drawn from Table 3 which shows the ratio of the RMSE in Table 1 over the RMSE for the uncensored posterior in the corresponding model or the corresponding uncensored BMA approach.

Table 1: Simulation experiment using 100 simulated data sets from GARCH(2,2) models. The table shows $100\times$ *the root mean squared error (RMSE), where for each simulated data set the 'error' is the difference between the observed fraction of 'violations' (realizations* y_t *that are more negative than the predicted percentile) and the desired value of 1%, 2%, ..., 10%. The mean is taken over the* $S = 100$ *data sets (in the first 10 columns), or over both the* $S = 100$ *data sets and the 10 percentiles (in the last column). MA refers to Model Averaging based on the predictive likelihood, either using the (uncensored) posterior and the (uncensored) predictive likelihood or using the censored posterior and the censored predictive likelihood. The theoretical value of the true model (with true parameter values) refers to 100 times the theoretical RMSE, i.e., 100 times the standard deviation* $\sqrt{p(1-p)/500}$ *for the* $100p\%$ *percentile.*

		1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
GARCH(1,1)	uncensored	8.26	12.30	12.34	12.28	13.20	13.46	13.79	13.82	13.71	13.38	12.75
GARCH(1,1)	censored (20%)	8.76	12.54	12.53	12.61	13.41	13.56	13.86	13.81	14.17	13.86	13.00
GARCH(1,1)	censored (30%)	8.64	12.49	12.75	12.60	13.38	13.55	13.73	13.77	13.80	13.57	12.91
GARCH(1,2)	uncensored	2.35	4.09	3.87	3.45	4.55	5.39	6.14	6.18	6.94	6.53	5.15
GARCH(1,2)	censored (20%)	2.57	4.08	4.10	3.70	4.17	4.65	5.44	5.67	6.16	5.88	4.77
GARCH(1,2)	censored (30%)	2.34	3.98	3.75	3.43	4.26	4.95	5.84	5.86	6.45	6.07	4.87
GARCH(2,1)	uncensored	5.66	8.11	7.96	7.81	8.69	9.39	9.49	9.70	10.06	9.56	8.73
GARCH(2,1)	censored (20%)	5.60	7.98	7.79	7.53	8.62	9.30	9.40	9.82	10.12	9.75	8.69
GARCH(2,1)	censored (30%)	5.38	7.87	7.78	7.52	8.42	9.10	9.39	9.70	9.96	9.52	8.57
GARCH(2,2)	uncensored	1.57	2.21	2.41	2.59	3.09	3.32	3.58	4.00	4.54	4.64	3.34
GARCH(2,2)	censored (20%)	1.60	2.11	2.26	2.50	2.90	3.34	3.95	4.21	4.57	4.76	3.39
GARCH(2,2)	censored (30%)	1.64	2.29	2.44	2.59	3.04	3.41	3.84	4.28	4.64	4.83	3.45
MA	uncensored	8.55	12.81	12.85	12.89	13.79	14.20	14.54	14.55	14.39	14.13	13.38
MA	censored (20%)	2.25	3.59	3.46	3.46	4.15	4.68	5.19	5.65	6.10	6.30	4.66
MA	censored (30%)	2.30	3.22	3.24	3.38	4.09	4.78	5.07	5.42	5.93	6.17	4.53
true model		0.44	0.63	0.76	0.88	0.97	1.06	1.14	1.21	1.28	1.34	

Table 2: Simulation experiment using 100 simulated data sets from GARCH(2,2) models. The table shows the ratio of the RMSE in Table 1 over the RMSE for the true model.

		1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
GARCH(1,1)	uncensored	18.56	19.65	16.18	14.02	13.54	12.68	12.08	11.39	10.71	9.98
GARCH(1,1)	censored (20%)	19.69	20.02	16.43	14.39	13.76	12.77	12.15	11.39	11.07	10.33
GARCH(1,1)	censored (30%)	19.42	19.95	16.71	14.38	13.73	12.76	12.03	11.35	10.78	10.11
GARCH(1,2)	uncensored	5.28	6.53	5.07	3.94	4.67	5.07	5.38	5.09	5.42	4.87
GARCH(1,2)	censored (20%)	5.78	6.52	5.37	4.23	4.28	4.38	4.77	4.68	4.82	4.38
GARCH(1,2)	censored (30%)	5.26	6.35	4.92	3.92	4.37	4.66	5.12	4.83	5.04	4.52
GARCH(2,1)	uncensored	12.72	12.96	10.44	8.91	8.91	8.84	8.31	7.99	7.86	7.13
GARCH(2,1)	censored (20%)	12.59	12.74	10.21	8.60	8.84	8.75	8.24	8.10	7.91	7.27
GARCH(2,1)	censored (30%)	12.09	12.57	10.20	8.59	8.64	8.57	8.23	7.99	7.78	7.09
GARCH(2,2)	uncensored	3.53	3.52	3.16	2.95	3.17	3.12	3.14	3.29	3.55	3.46
GARCH(2,2)	censored (20%)	3.60	3.36	2.97	2.86	2.98	3.14	3.46	3.47	3.57	3.55
GARCH(2,2)	censored (30%)	3.70	3.65	3.20	2.95	3.12	3.21	3.36	3.53	3.63	3.60
MA	uncensored	19.21	20.45	16.85	14.71	14.15	13.37	12.74	11.99	11.25	10.54
MA	censored (20%)	5.06	5.74	4.54	3.95	4.25	4.41	4.55	4.65	4.76	4.70
MA	censored (30%)	5.18	5.14	4.25	3.86	4.20	4.50	4.45	4.47	4.64	4.60
true model		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Simulation experiment using 100 simulated data sets from GARCH(2,2) models. The table shows the ratio of the RMSE in Table 1 over the RMSE for the uncensored posterior in the corresponding model or the corresponding uncensored Model Averaging approach.

		1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
GARCH(1,1)	censored (20%)	1.06	1.02	1.02	1.03	1.02	1.01	1.01	1.00	1.03	1.04	1.02
GARCH(1,1)	censored (30%)	1.05	1.02	1.03	1.03	1.01	1.01	1.00	1.00	1.01	1.01	1.01
GARCH(1,2)	censored (20%)	1.10	1.00	1.06	1.07	0.92	0.86	0.89	0.92	0.89	0.90	0.92
GARCH(1,2)	censored (30%)	1.00	0.97	0.97	1.00	0.94	0.92	0.95	0.95	0.93	0.93	0.94
GARCH(2,1)	censored (20%)	0.99	0.98	0.98	0.96	0.99	0.99	0.99	1.01	1.01	1.02	1.00
GARCH(2,1)	censored (30%)	0.95	0.97	0.98	0.96	0.97	0.97	0.99	1.00	0.99	1.00	0.98
GARCH(2,2)	censored (20%)	1.02	0.95	0.94	0.97	0.94	1.01	1.10	1.05	1.01	1.03	1.02
GARCH(2,2)	censored (30%)	1.05	1.04	1.01	1.00	0.98	1.03	1.07	1.07	1.02	1.04	1.04
MA	censored (20%)	0.26	0.28	0.27	0.27	0.30	0.33	0.36	0.39	0.42	0.45	0.35
MA	censored (30%)	0.27	0.25	0.25	0.26	0.30	0.34	0.35	0.37	0.41	0.44	0.34

5. Empirical applications

5.1. Empirical application 1: GJR-GARCH models for stock index returns

We perform a similar experiment as in the previous section, using two empirical data sets of daily returns on the S&P 500 and Nikkei 225 stock indices. Again, we consider $\tilde{T} = 1000$ observations (the trading days from March 6 2007 to February 7 2011), where again a moving window of 500 observations is used for the estimation and combination of the models, where the purpose is the accurate one-day-ahead prediction of the left tail of the returns distribution. There are two main differences with the experiment using simulated data sets in the previous section. First, since a *leverage effect* – the phenomenon that a negative return has a larger effect on the variance of tomorrow's return than a positive return of the same size – is observed for many returns on stocks and stock indices, we consider the GJR-GARCH (GJR-GARCH(p, q, s)) model of Glosten et al. (1993)

$$\begin{aligned} y_t &= \sigma_t \, \varepsilon_t \quad (t = 1, \ldots, \tilde{T}), & (33) \\ \sigma_t^2 &= \beta_0 + \alpha_1 y_{t-1}^2 + \ldots + \alpha_q y_{t-q}^2 & \\ &\quad + \gamma_1 I\{y_{t-1} < 0\} y_{t-1}^2 + \ldots + \gamma_s I\{y_{t-s} < 0\} y_{t-s}^2 & \\ &\quad + \beta_1 \, \sigma_{t-1}^2 + \ldots + \beta_p \, \sigma_{t-p}^2, & (34) \end{aligned}$$

where the i.i.d. ε_t have a Student's t-distribution with ν degrees of freedom. We consider the 80 ($= 4 \times 4 \times 5$) GJR-GARCH(p, q, s) models with $p = 1, 2, 3, 4$, $q = 1, 2, 3, 4$, $s = 0, 1, 2, 3, 4$. Another reason for including more models than in the previous section, next to the leverage effect that is often observed for stock index returns, is that we investigate only two empirical data sets, instead of 100 simulated sets. For the simulation experiment in the previous section, the computing time was already enormous in the case of four models. The second difference is that we also consider the 'semi censored' predictive Model Averaging approaches, where only the posterior or the predictive likelihood is censored, whereas the other remains 'uncensored'.

Tables 4 - 6 show the results for the S&P 500. Tables 7 - 9 show the results for the Nikkei 225. We observe the following findings *for both data sets.* First, the worst results

Table 4: Empirical application to daily returns on S&P 500 using GJR-GARCH models. The table shows $100\times$ *the root mean squared error (RMSE), where for each simulated data set the 'error' is the difference between the observed fraction of 'violations' (realizations* y_t *that are more negative than the predicted percentile) and the desired value of 1%, 2%, . . ., 10%. For the individual percentiles the RMSE reduces to the absolute error (in the first 10 columns); for the RMSE in the last column the mean is taken over the 10 percentiles. We make use of Model Averaging based on the predictive likelihood, using the uncensored or censored posterior and the uncensored or censored predictive likelihood. The theoretical value of the true model (with true parameter values) refers to 100 times the theoretical RMSE, i.e., 100 times the standard deviation* $\sqrt{p(1-p)/500}$ *for the* $100p\%$ *percentile.*

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
uncensored	uncensored	0.80	1.60	2.40	3.00	3.20	4.00	5.00	5.20	5.60	6.00	4.04
uncensored	censored (20%)	1.00	1.60	2.00	2.60	3.60	3.80	4.20	4.20	5.00	5.80	3.68
uncensored	censored (30%)	1.00	1.60	2.00	2.60	3.60	3.80	4.00	4.20	5.20	5.80	3.69
censored (20%)	uncensored	1.00	1.80	2.40	3.00	3.60	4.60	4.40	4.60	5.20	6.20	3.99
censored (30%)	uncensored	1.00	1.80	2.20	2.80	3.60	4.20	4.40	4.40	5.20	6.00	3.86
censored (20%)	censored (20%)	1.00	1.40	2.20	2.80	3.20	3.80	4.20	4.60	4.80	5.40	3.62
censored (30%)	censored (30%)	1.00	1.60	2.40	2.40	3.00	3.20	4.00	4.60	5.40	6.20	3.73
true model		0.44	0.63	0.76	0.88	0.97	1.06	1.14	1.21	1.28	1.34	

– in the sense of the highest RMSE over the 10 percentiles – are obtained by the uncensored predictive BMA approach, where both the posterior and the predictive likelihood remain uncensored. Second, the best results are obtained by the approach where both the posterior and the predictive likelihood are censored, where the 'censoring boundary' R_t is taken equal to the 20% percentile of the observations. Third, the results are quite robust with respect to the choice of this 'censoring boundary' R_t: the results are similar for the approach where R_t is equal to the 30% percentile. Fourth, the censoring of the predictive likelihood is more beneficial than the censoring of the posterior: the approach where only the posterior is censored performs worse than the approach where only the predictive likelihood is censored. The latter's performance is close to the performance of the method where both the posterior and the predictive likelihood are censored. In the next subsection, we will analyze the difference in performance between the predictive likelihood and the censored predictive likelihood, where the posterior is left uncensored. Fifth, Tables 5 and 8 show that for our data window the relative performance, the ratio of the absolute error over the standard deviation of the error in a theoretical true model, typically becomes better for the lower percentiles in the deep tail.

Table 5: Empirical application to daily returns on S&P 500 using GJR-GARCH models. The table shows the ratio of the RMSE in Table 4 over the RMSE for the true model.

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
uncensored	uncensored	1.80	2.56	3.15	3.42	3.28	3.77	4.38	4.29	4.38	4.47
uncensored	censored (20%)	2.25	2.56	2.62	2.97	3.69	3.58	3.68	3.46	3.91	4.32
uncensored	censored (30%)	2.25	2.56	2.62	2.97	3.69	3.58	3.51	3.46	4.06	4.32
censored (20%)	uncensored	2.25	2.87	3.15	3.42	3.69	4.33	3.86	3.79	4.06	4.62
censored (30%)	uncensored	2.25	2.87	2.88	3.20	3.69	3.95	3.86	3.63	4.06	4.47
censored (20%)	censored (20%)	2.25	2.24	2.88	3.20	3.28	3.58	3.68	3.79	3.75	4.02
censored (30%)	censored (30%)	2.25	2.56	3.15	2.74	3.08	3.01	3.51	3.79	4.22	4.62
true model		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6: Empirical application to daily returns on S&P 500 using GJR-GARCH models. The table shows the ratio of the RMSE in Table 4 over the RMSE for the uncensored posterior and uncensored predictive likelihood.

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
uncensored	censored (20%)	1.25	1.00	0.83	0.87	1.13	0.95	0.84	0.81	0.89	0.97	0.91
uncensored	censored (30%)	1.25	1.00	0.83	0.87	1.13	0.95	0.80	0.81	0.93	0.97	0.91
censored (20%)	uncensored	1.25	1.13	1.00	1.00	1.13	1.15	0.88	0.88	0.93	1.03	0.99
censored (30%)	uncensored	1.25	1.13	0.92	0.93	1.13	1.05	0.88	0.85	0.93	1.00	0.96
censored (20%)	censored (20%)	1.25	0.88	0.92	0.93	1.00	0.95	0.84	0.88	0.86	0.90	0.90
censored (30%)	censored (30%)	1.25	1.00	1.00	0.80	0.94	0.80	0.80	0.88	0.96	1.03	0.92

Table 7: Empirical application to daily returns on Nikkei 225 using GJR-GARCH models. The table shows 100× *the root mean squared error (RMSE), where for each simulated data set the 'error' is the difference between the observed fraction of 'violations' (realizations y_t that are more negative than the predicted percentile) and the desired value of 1%, 2%, …, 10%. For the individual percentiles the RMSE reduces to the absolute error (in the first 10 columns); for the RMSE in the last column the mean is taken over the 10 percentiles. We make use of Model Averaging based on the predictive likelihood, using the uncensored or censored posterior and the uncensored or censored predictive likelihood. The theoretical value of the true model (with true parameter values) refers to 100 times the theoretical RMSE, i.e., 100 times the standard deviation* $\sqrt{p(1-p)/500}$ *for the* $100p\%$ *percentile.*

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
uncensored	uncensored	0.80	1.20	2.00	3.00	3.20	4.00	4.60	4.80	5.40	6.20	3.91
uncensored	censored (20%)	0.60	1.20	2.00	2.60	3.40	3.60	3.60	4.00	4.60	4.80	3.32
uncensored	censored (30%)	0.80	1.20	2.00	2.60	3.40	3.80	3.60	4.00	4.60	4.80	3.35
censored (20%)	uncensored	0.80	1.40	2.20	3.00	3.60	4.40	4.60	4.60	4.80	5.40	3.79
censored (30%)	uncensored	0.80	1.40	2.20	3.00	3.60	4.40	4.20	4.40	4.60	5.20	3.66
censored (20%)	censored (20%)	1.00	1.40	2.00	2.20	2.80	3.20	3.40	4.00	4.80	5.40	3.31
censored (30%)	censored (30%)	0.40	1.40	1.80	2.40	2.40	3.00	4.00	4.40	5.00	5.00	3.34
true model		0.44	0.63	0.76	0.88	0.97	1.06	1.14	1.21	1.28	1.34	

Table 8: Empirical application to daily returns on Nikkei 225 using GJR-GARCH models. The table shows the ratio of the RMSE in Table 7 over the RMSE for the true model.

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
uncensored	uncensored	1.80	1.92	2.62	3.42	3.28	3.77	4.03	3.96	4.22	4.62
uncensored	censored (20%)	1.35	1.92	2.62	2.97	3.49	3.39	3.15	3.30	3.59	3.58
uncensored	censored (30%)	1.80	1.92	2.62	2.97	3.49	3.58	3.15	3.30	3.59	3.58
censored (20%)	uncensored	1.80	2.24	2.88	3.42	3.69	4.14	4.03	3.79	3.75	4.02
censored (30%)	uncensored	1.80	2.24	2.88	3.42	3.69	4.14	3.68	3.63	3.59	3.88
censored (20%)	censored (20%)	2.25	2.24	2.62	2.51	2.87	3.01	2.98	3.30	3.75	4.02
censored (30%)	censored (30%)	0.90	2.24	2.36	2.74	2.46	2.82	3.51	3.63	3.91	3.73
true model		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 9: Empirical application to daily returns on Nikkei 225 using GJR-GARCH models. The table shows the ratio of the RMSE in Table 7 over the RMSE for the uncensored posterior and uncensored predictive likelihood.

posterior	predictive likelihood	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	1%-10%
uncensored	censored (20%)	0.75	1.00	1.00	0.87	1.06	0.90	0.78	0.83	0.85	0.77	0.85
uncensored	censored (30%)	1.00	1.00	1.00	0.87	1.06	0.95	0.78	0.83	0.85	0.77	0.86
censored (20%)	uncensored	1.00	1.17	1.10	1.00	1.13	1.10	1.00	0.96	0.89	0.87	0.97
censored (30%)	uncensored	1.00	1.17	1.10	1.00	1.13	1.10	0.91	0.92	0.85	0.84	0.94
censored (20%)	censored (20%)	1.25	1.17	1.00	0.73	0.88	0.80	0.74	0.83	0.89	0.87	0.85
censored (30%)	censored (30%)	0.50	1.17	0.90	0.80	0.75	0.75	0.87	0.92	0.93	0.81	0.85

5.2. Empirical application 2: GARCH, GJR-GARCH and EGARCH models for stock index returns

In this subsection we specifically analyze the difference in performance between the (uncensored) predictive likelihood and the censored predictive likelihood. As in the previous subsection, we consider the log-returns of the S&P 500 and Nikkei 225 stock indices. However, there are five differences. First, we consider larger data sets: we use a moving window of 2000 in-sample observations for estimation and model combination, in order to predict 1000 out-of-sample observations. Second, we also consider the EGARCH model of Nelson (1991):

$$y_t = \sigma_t \, \varepsilon_t \qquad (t = 1, \ldots, \tilde{T}), \tag{35}$$

$$\log(\sigma_t^2) = \beta_0 + \alpha_1 \varepsilon_{t-1} + \gamma_1 |\varepsilon_{t-1}| + \beta_1 \, \log(\sigma_{t-1}^2), \tag{36}$$

where the i.i.d. ε_t have a Student's t-distribution with ν degrees of freedom. The property that this specification of an EGARCH model with Student's t errors for the log-returns implies that the distribution of the stock index itself has no finite moments does not pose a problem for our analysis: the density for the log-return and the corresponding Value-at-Risk estimates are well-defined. Third, we only consider the GARCH(1,1), GJR-GARCH(1,1,1) and EGARCH(1,1,1) models, which are popular in practice. Fourth, we compare the performance using different criteria. We consider the Diebold-Mariano test based on the logarithmic scoring rule that was analyzed by Mitchell and Hall (2005), Amisano and Giacomini (2007), and Bao et al. (2004, 2007), and the Diebold-Mariano test based on the censored likelihood score function proposed by Diks et al. (2011), which can be interpreted as the censored version of the logarithmic scoring rule. The first test has no specific focus on the left tail, whereas the second test specifically focuses on the quality of the prediction of the left tail. Under the null hypothesis of equal performance, the test statistic in these Diebold-Mariano tests asymptotically has a standard normal distribution. Fifth, we also consider the validity of the 95% Value-at-Risk forecasts, using the tests for correct unconditional coverage, independence and correct conditional coverage of Christoffersen (1998).

First, we consider the results for the S&P 500. The results of the Diebold-Mariano test based on the logarithmic scoring rule are given by Table 10. From this it is clear that the GARCH model performs significantly worse than the GJR-GARCH and EGARCH models (due to significant leverage effects) and the model averaging methods. Further, there is no significant difference in performance between the model averaging methods based upon the predictive likelihood and the censored predictive likelihood. The censoring of the predictive likelihood does not seem to significantly harm the quality of the density forecasts. The results of the Diebold-Mariano test based on the censored likelihood score function are given by Table 11. For the prediction of the left tail, the GJR-GARCH model outperforms the GARCH and GJR-GARCH models, where the GARCH model somehow seems to outperform the EGARCH model. Moreover, the model averaging method based upon the censored predictive likelihood significantly outperforms (at a significance level of 10%) the model averaging method based upon the predictive likelihood. That is, the censoring of the predictive likelihood significantly improves the quality of the left-tail density forecasts. The results for the tests for the validity of the 95% Value-at-Risk forecasts — the tests for correct unconditional coverage, independence and correct conditional coverage of Christoffersen (1998) — are given by Table 12. The fraction of violations (with $y_t < \text{VaR}_t$) is substantially too large (i.e., larger than 0.05) for all models and model averaging methods, which causes that the validity is rejected in each case. Also the independence of the violations is rejected (at a significance level of 5%) for all models and model averaging methods. More advanced models such as regime-switching models or stochastic volatility models (or combinations thereof) may be able to yield valid Value-at-Risk forecasts; we leave this as a topic for further research.

Table 10: Results for the Diebold-Mariano test based on the logarithmic scoring rule (that has no specific focus on the left tail) for 1000 one-day-ahead density forecasts for the S&P 500 index corresponding to the period October 31, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood. Positive (Negative) values indicate better performance of the model/method in the row (column). The asterisks *, ** and *** indicate significance at 10%, 5% and 1% significance levels (in a two-sided test), respectively.

	GJR-GARCH	EGARCH	BMA (pl)	MA (cspl)
GARCH	-2.88***	-2.29**	-2.94***	-3.86***
GJR-GARCH		-0.01	-0.15	-0.24
EGARCH			-0.15	-0.08
BMA (pl)				0.01

Table 11: Results for the Diebold-Mariano test based on the censored likelihood scoring rule (that has a specific focus on the left tail) for 1000 one-day-ahead density forecasts for the S&P 500 index corresponding to the period October 31, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood. Positive (Negative) values indicate better performance of the model/method in the row (column). The asterisks *, ** and *** indicate significance at 10%, 5% and 1% significance levels (in a two-sided test), respectively.

	GJR-GARCH	EGARCH	BMA (pl)	MA (cspl)
GARCH	-1.88*	1.91*	0.37	-1.67*
GJR-GARCH		3.31***	2.43**	1.47
EGARCH			-3.25***	-2.93***
BMA (pl)				-1.87*

Table 12: Results for the tests for the validity of the 95% Value-at-Risk forecasts: the tests for correct unconditional coverage, independence and correct conditional coverage of Christoffersen (1998) applied to 1000 one-day-ahead forecasts for the S&P 500 index corresponding to the period October 31, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood.

	fraction $y_t <$ VaR$_t$	LR_{uc}	(p-value)	LR_{ind}	(p-value)	LR_{cc}	(p-value)
GARCH	0.071	8.26	(0.00)	5.36	(0.02)	13.62	(0.00)
GJR-GARCH	0.070	7.53	(0.01)	5.11	(0.02)	12.64	(0.00)
EGARCH	0.082	18.22	(0.00)	5.22	(0.02)	23.44	(0.00)
BMA (pl)	0.071	8.26	(0.00)	2.68	(0.10)	10.94	(0.00)
MA (cspl)	0.071	8.26	(0.00)	5.36	(0.02)	13.62	(0.00)

Table 13: Results for the Diebold-Mariano test based on the logarithmic scoring rule (that has no specific focus on the left tail) for 1000 one-day-ahead density forecasts for the Nikkei 225 index corresponding to the period September 16, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood. Positive (Negative) values indicate better performance of the model/method in the row (column). The asterisks *, ** and *** indicate significance at 10%, 5% and 1% significance levels (in a two-sided test), respectively.

	GJR-GARCH	EGARCH	BMA (pl)	MA (cspl)
GARCH	-2.88***	-2.74***	-2.75***	-3.21***
GJR-GARCH		-0.17	-0.05	-0.38
EGARCH			0.80	-0.08
BMA (pl)				-0.28

Second, we consider the results for the Nikkei 225. The results of the Diebold-Mariano test based on the logarithmic scoring rule are given by Table 13. As for the S&P500, the GARCH model performs significantly worse than the GJR-GARCH and EGARCH models (due to significant leverage effects) and the model averaging methods. Further, there is no significant difference in performance between the model averaging methods based upon the predictive likelihood and the censored predictive likelihood. The censoring of the predictive likelihood does not seem to significantly harm the quality of the density forecasts. The results of the Diebold-Mariano test based on the censored likelihood score function are given by Table 14. For the prediction of the left tail, the GJR-GARCH model outperforms the GARCH model. Moreover, the model averaging method based upon the censored predictive likelihood significantly outperforms (at a significance level of 10%) the model averaging method based upon the predictive likelihood. That is, the censoring of the predictive likelihood significantly improves the quality of the left-tail density forecasts. The results for the tests for the validity of the 95% Value-at-Risk forecasts — the tests for correct unconditional coverage, independence and correct conditional coverage of Christoffersen (1998) — are given by Table 15. The fraction of violations (with $y_t < \mathrm{VaR}_t$) larger than 0.05 for all models and model averaging methods. However, at a significance level of 1% or 2.5%, the validity of the VaR forecasts is not

Table 14: Results for the Diebold-Mariano test based on the censored likelihood scoring rule (that has a specific focus on the left tail) for 1000 one-day-ahead density forecasts for the Nikkei 225 index corresponding to the period September 16, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood. Positive (Negative) values indicate better performance of the model/method in the row (column). The asterisks *, ** and *** indicate significance at 10%, 5% and 1% significance levels (in a two-sided test), respectively.

	GJR-GARCH	EGARCH	BMA (pl)	MA (cspl)
GARCH	-2.23**	-0.78	-0.63	-1.83*
GJR-GARCH		1.61	1.99**	1.21
EGARCH			1.74*	-1.32
BMA (pl)				-1.83*

Table 15: Results for the tests for the validity of the 95% Value-at-Risk forecasts: the tests for correct unconditional coverage, independence and correct conditional coverage of Christoffersen (1998) applied to 1000 one-day-ahead forecasts for the Nikkei 225 index corresponding to the period September 16, 2006 to October 20, 2010. BMA (pl) refers to Bayesian Model Averaging with model weights based on the (uncensored) predictive likelihood. MA (cspl) refers to Model Averaging with model weights based on the censored predictive likelihood.

	fraction $y_t <$ VaR$_t$	LR_{uc}	(p-value)	LR_{ind}	(p-value)	LR_{cc}	(p-value)
GARCH	0.073	9.81	(0.00)	0.09	(0.76)	9.91	(0.01)
GJR-GARCH	0.069	6.83	(0.01)	0.85	(0.36)	7.68	(0.02)
EGARCH	0.068	6.16	(0.01)	0.74	(0.39)	6.90	(0.03)
BMA (pl)	0.066	4.92	(0.03)	0.54	(0.46)	5.46	(0.07)
MA (cspl)	0.065	4.35	(0.04)	0.45	(0.50)	4.79	(0.09)

rejected for the model averaging methods (in any of the three tests), whereas correct unconditional coverage is rejected for the individual models. Moreover, the censored predictive likelihood yields a slightly better performance than the predictive likelihood.

6. Conclusion

We have introduced two novel concepts, the censored posterior and the censored predictive likelihood, that offer a specific focus on a particular part such as the left tail of the predictive density for forecasting of the Value at Risk. Extensive experiments are reported, involving simulated and empirical data. The obtained results show the ability of these innovative approaches to outperform the standard posterior and the traditional Bayesian Model Averaging techniques in applications of Value at Risk prediction in GARCH models. Especially, we find that the censored predictive likelihood can provide significantly and substantially better results than the (uncensored) predictive likelihood.

Multiple suggestions for further research have already been mentioned throughout the paper. In any case, the use of parallel computations on Graphics Processing Units (GPUs) should be considered to reduce the enormous computing time of the experiments, especially when applying our computationally intensive method to large numbers of (simulated) data sets. Moreover, we intend to investigate multivariate models (e.g., the Dynamic Conditional Correlation (DCC) model of Engle (2002), and different univariate models (e.g., GARCH models with different variance equations and with different distributions for the standardized error terms, or stochastic volatility models). Further, we intend to analyze other data sets (e.g., exchange rates), larger estimation windows, different values for the 'censoring boundary percentile' (e.g., the 10% percentile of a larger estimation window), and different tests such as the tests proposed by Hoogerheide et al. (2012b) who comment on the forecast rationality tests of Patton and Timmermann (2012). As an alternative to the model combination framework involving the predictive likelihood, the concept of censoring can be introduced within the forecast combination framework of Hoogerheide et al. (2010), which involves a certain type of time-varying model weights.

References

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business and Economic Statistics 25, 177–190.

Ardia, D., Baştürk, N., Hoogerheide, L. F., Van Dijk, H. K., 2012. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. Computational Statistics & Data Analysis 56 (11), 3398–3414.

Bao, Y., Lee, T.-H., Saltoğlu, B., 2004. A test for density forecast comparison with applications to risk management. Working Paper 04-08, UC Riverside.

Bao, Y., Lee, T.-H., Saltoğlu, B., 2007. Comparing density forecast models. Journal of Forecasting 26, 203–225.

Bartlett, M. S., 1957. A comment on D.V. Lindley's statistical paradox. Biometrika 44 (3/4), 533–534.

Bates, J. M., Granger, C. W. J., 1969. Combination of forecasts. Operational Research Quarterly 20, 451–468.

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91 (433), 109–122.

Billio, M., Casarin, R., Ravazzolo, F., Van Dijk, H. K., 2013. Time-varying combinations of predictive densities using nonlinear filtering. Journal of Econometrics, forthcoming.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307–327.

Christoffersen, P. F., 1998. Evaluating interval forecasts. International Economic Review 39 (4), 841–862.

Diks, C., Panchenko, V., Van Dijk, D., 2011. Likelihood-based scoring rules for comparing density forecasts in tails. Journal of Econometrics 163, 215–230.

Efron, B., Tibshirani, R., 1993. An Introduction to Bootstrap. Chapman and Hall, New York, USA.

Eklund, J., Karlsson, S., 2007. Forecast combination and model averaging using predictive measures. Econometric Reviews 26, 329–363.

Engle, R., 2002. Dynamic Conditional Correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20, 339–350.

Gelfand, A., Dey, D., 1994. Bayesian model choice: Asymptotics and exact calculations. Journal of the Royal Statistical Society Series B 56 (3), 501–514.

Geweke, J., Whiteman, C., 2006. Bayesian forecasting. In: Elliot, G., Granger, C. W. J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. North-Holland: Amsterdam, pp. 3–80.

Glosten, L. R., Jaganathan, R., Runkle, D. E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48 (5), 1779–1801.

Hammersley, J., Handscomb, D., 1964. Monte Carlo Methods, 1st Edition. Methuen, London.

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hoogerheide, L. F., Kleijn, R., Ravazzolo, F., Van Dijk, H. K., Verbeek, M., 2010. Forecast accuracy and economic gains from Bayesian Model Averaging using time varying weights. Journal of Forecasting 29, 251–269.

Hoogerheide, L. F., Opschoor, A., van Dijk, H. K., 2012a. A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation. Journal of Econometrics 171 (2), 101–120.

Hoogerheide, L. F., Ravazzolo, F., Van Dijk, H. K., 2012b. Comment on forecast rationality tests based on multi-horizon bounds. Journal of Business & Economic Statistics 30 (1), 30–33.

Hoogerheide, L. F., Van Dijk, H. K., 2010. Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling. International Journal of Forecasting 26 (2), 231–247.

Hoogerheide, L. F., Van Dijk, H. K., Van Oest, R. D., 2009. Simulation based Bayesian econometric inference: Principles and some recent computational advances. In: Belsley, D. A., Kontoghiorghes, E. (Eds.), Handbook of Computational Econometrics. Wiley, pp. 215–280.

Kloek, T., Van Dijk, H. K., 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. Econometrica 46, 1–20.

Lindley, D. V., 1957. A statistical paradox. Biometrika 44 (1/2).

Madigan, D., Raftery, A., 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. Journal of the American Statistical Association 89, 1335–1346.

Marcellino, M., 2004. Forecasting pooling for short time series of macroeconomic variables. Oxford Bulletin of Economic and Statistics 66, 91–112.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1092.

Mitchell, J., Hall, S. G., 2005. Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr 'fan' charts of inflation. Oxford Bulletin of Economics and Statistics 67, 995–1033.

Nelson, D. B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59 (2), 347–370.

O'Hagan, A., 1995. Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society Series B 57 (1), 99–138.

Patton, A. J., Timmermann, A., 2012. Forecast rationality tests based on multi-horizon bounds. Journal of Business & Economic Statistics 30 (1), 1–17.

Stock, J. H., Watson, M., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405–430.