Partially Censored Posterior for Robust and Efficient Risk Evaluation.*

Agnieszka Borowska[(a,b)], Lennart Hoogerheide[(a,b)], Siem Jan Koopman[(a,b,c)] and Herman K. van Dijk[(b,d,e)]

[(a)] Vrije Universiteit Amsterdam

[(b)] Tinbergen Institute

[(c)] CREATES, Aarhus University

[(d)] Erasmus University Rotterdam

[(e)] Norges Bank

July 2019

Abstract

A novel approach to inference for a specific region of the predictive distribution is introduced. An important domain of application is accurate prediction of financial risk measures, where the area of interest is the left tail of the predictive density of logreturns. Our proposed approach originates from the Bayesian approach to parameter estimation and time series forecasting, however it is robust in the sense that it provides a more accurate estimation of the predictive density in the region of interest in case of misspecification. The first main contribution of the paper is the novel concept of the Partially Censored Posterior (PCP), where the set of model parameters is partitioned into two subsets: for the first subset of parameters we consider the standard marginal posterior, for the second subset of parameters (that are particularly related to the region of interest) we consider the conditional censored posterior. The censoring means that observations outside the region of interest are censored: for those observations only the probability of being outside the region of interest matters. This quasi-Bayesian approach yields more precise parameter estimation than a fully censored posterior for all parameters, and has more focus on the region of interest than a standard Bayesian approach. The second main contribution is that we introduce two novel methods for computationally efficient simulation: Conditional MitISEM, a Markov chain Monte Carlo method to simulate model parameters from the Partially Censored Posterior, and PCP-QERMit, an Importance Sampling method that is introduced to further decrease the numerical standard errors of the Value-at-Risk and Expected Shortfall estimators. The third main contribution is that we consider the effect of using a time-varying boundary of the region of interest, which may provide more information about the left tail of the distribution of the standardized innovations. Extensive simulation and empirical studies show the ability of the introduced method to outperform standard approaches.

Keywords: Bayesian inference; censored likelihood; censored posterior; partially censored posterior; misspecification; density forecasting; Markov chain Monte Carlo; importance sampling; mixture of Student's t; Value-at-Risk; Expected Shortfall.

*This working paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. We would like to thank two anonymous referees and Jeroen Rombouts for detailed and constructive comments on earlier drafts of the article. We are also grateful to Marc Baardman, Paolo Gorgi, Anne Opschoor, the participants of the 1st EcoSta 2017 conference (Hong Kong, 15–17 June 2017), the 8th ESOBE 2017 (Maastricht, 26–27 October 2017) and the 12th NESG meeting (Amsterdam, 25 May 2018) for providing useful comments and suggestions.

1 Introduction

The issue of accurate estimation of the left tail of the predictive distribution of returns is crucial from the risk management perspective and is thus commonly investigated by both academics and practitioners. One of the main reasons for its importance is that it is used to obtain measures of downside risk for investments such as Value-at-Risk (VaR) and Expected Shortfall (ES), cf. McNeil and Frey (2000) and McNeil et al. (2015). The task of tail prediction is a special case of density forecasting where the focus is on a specific subset of the domain of the predictive distribution. Density forecasting in general has been rapidly growing in econometrics, finance and macroeconomics due to increased understanding of the limited informativeness of point forecasts, cf. Diks et al. (2011). In contrast to these, density forecasts provide a full insight into the forecast uncertainty. For a survey of the evolution of density forecasting in economics, see Aastveit et al. (2019).

A natural framework, therefore, for analysing density forecasts is the Bayesian framework, as it treats all unobserved quantities as parameters to be estimated; see e.g. Geweke and Amisano (2010) for a comparison and evaluation of Bayesian predictive distributions. This includes the predictions for the observation process. Importantly, the Bayesian approach incorporates the parameter uncertainty into analysis and facilitates dealing with model uncertainty, usually via Bayesian Model Averaging. However, the issue of Bayesian model misspecification still seems to be an open question.[1] A formal approach to this problem is provided by Kleijn and van der Vaart (2006), who show (under stringent conditions) that given an incorrectly specified model, the posterior concentrates *'close'* to the points in the support of the prior that minimise the Kullback-Leibler divergence with respect to the true data generating process (DGP). This result can be seen as the Bayesian counterpart of the MLE being consistent for the pseudo-true values in frequentist statistics. Nevertheless, differently than the asymptotic distribution of the MLE, the estimated posterior variance is incorrect in case of misspecification (Kleijn and van der Vaart, 2006). Müller (2013) shows that one can rescale the posterior so that credible sets have the correct coverage. As a practical solution to the problem, Geweke and Amisano (2012) apply the so-called model pooling, which relaxes the key assumption behind model averaging that the true model is in the set of models under consideration.

In the context of tail forecasting, the crucial question is: *what if "close" is not close enough*? From the perspective of accurate tail prediction obtaining estimates being just *"close"* to their real values is likely to lead to incorrect risk measures and hence to poor managerial decisions in cases where the misspecification is severe. To improve inference on a particular region of the predictive density, Gatarek et al. (2013) introduce the Censored Posterior (CP) for estimation and the censored predictive likelihood for model combination using Model Averaging. A concept underlying their approach is the censored likelihood scoring function of Diks et al. (2011), an adaptation (with specific focus on the left tail) of the popular logarithmic scoring rule, cf. Hall and Mitchell (2007) and Amisano and Giacomini (2007). Diks et al. (2011) use the censored likelihood scoring function only for comparing density forecasts in tails, not for estimation. The censoring means that observations outside the region of interest are censored: for those observations only the probability of being outside the region of interest matters. However, as we discuss in the later part of this paper, for densely parametrised models applied in practice the Censored Posterior approach is likely to lose too much information.

To overcome these shortcomings the first main contribution of this paper is the novel concept of the Partially Censored Posterior (PCP), where the set of model parameters is partitioned into two subsets: the first, for which we consider the standard marginal posterior, and the second, for which we consider a conditional censored posterior. In the second subset we choose parameters that are expected to especially

[1] At the time of writing there is an active, ongoing debate in the Bayesian community about the issue of Bayesian model misspecification. Interestingly, it seems that there is no common ground on it (yet)! Cf. Robert (2017) and Cross Validated (2017).

benefit from censoring (due to their particular relationship with the tail of the predictive distribution). This quasi-Bayesian approach leads to more precise parameter estimation than a fully censored posterior for all parameters, and has more focus on the region of interest than the standard Bayesian approach (that is, with no censoring).

The second main contribution is that we introduce two novel simulation methods. The first method is a Markov chain Monte Carlo (MCMC) method to simulate model parameters from the Partially Censored Posterior. Here we extend the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) algorithm of Hoogerheide et al. (2012) to propose the *Conditional MitISEM* approach, where we approximate the joint censored posterior with a mixture of Student's t distributions and use the resulting conditional mixture of Student's t distributions as a candidate distribution for the conditional censored posterior. The high quality of the (conditional) candidate distributions leads to a computationally efficient MCMC method. The second method is an Importance Sampling method that is introduced to further decrease the numerical standard errors of the VaR and ES estimators. Here we adapt the *Quick Evaluation of Risk using Mixture of t approximations* (QERMit) algorithm of Hoogerheide and van Dijk (2010) to propose the *PCP-QERMit* method, where an adaptation is required since we do not have a closed-form formula for the partially censored posterior density kernel.

The third main contribution is that we consider the effect of using a time-varying boundary of the region of interest. To the best of our knowledge, the literature on the censored likelihood scoring rule, the censored likelihood and the censored posterior has been limited to a time-constant threshold defining the left tail. However, a constant threshold might be suboptimal when we focus on the left tail of the conditional distribution (given past observations). Even if the interest is in the unconditional left tail, then the time-varying threshold may be still more advantageous than the time-constant one. This is simply because the time-varying threshold allows us to obtain more information about the left tail of the distribution of the standardized innovations compared to the time-constant one.

The outline of this paper is as follows. In Section 2 we consider the risk measure concepts and discuss the censored posterior. Moreover, we introduce our novel concept of the Partially Censored Posterior and the novel simulation methods of Conditional MitISEM and PCP-QERMit. As an other extension of the existing literature on censored likelihood based methods, in Section 3 we introduce a time-varying threshold for censoring. In Section 4 we provide an empirical application using an AGARCH model with skewed-t innovations, a GAS-skewed-t model and GAS-t model for daily IBM logreturns. Section 5 concludes.

2 Censored posterior and partially censored posterior

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series of daily logreturns on a financial asset price, with $y_{1:T} = \{y_1, \ldots, y_T\}$ denoting the (in-sample) observed data. We denote $y_{s:r} = \{y_s, y_{s+1}, \ldots, y_{r-1}, y_r\}$ for $s \leq r$. We assume that $\{y_t\}_{t\in\mathbb{Z}}$ is subject to a dynamic stationary process parametrised by θ, on which we put a prior $p(\theta)$. We are interested in the conditional predictive density of $y_{T+1:T+H}$, given the observed series $y_{1:T}$. In particular, we are interested in the standard risk measure given by the $100(1-\alpha)\%$ VaR (in the sense of McNeil and Frey, 2000), the $100\alpha\%$ quantile of the predictive distribution of $\sum_{t=T+1}^{T+H} y_t$ given $y_{1:T}$. We also consider the ES as an alternative risk measure, due to its advantageous properties compared to the VaR, mainly sub-additivity (which makes ES a coherent risk measure in the sense of Artzner et al., 1999):

$$100(1-\alpha)\%\,\mathrm{ES} = \mathbb{E}\left[\sum_{t=T+1}^{T+H} y_t \,\middle|\, \sum_{t=T+1}^{T+H} y_t < 100(1-\alpha)\%\,\mathrm{VaR}\right].$$

The regular (uncensored) likelihood is given by the standard formula

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

and the posterior predictive density is

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

Given the data $y_{1:T}$ and a set of parameter draws $\{\theta^{(i)}\}_{i=1}^{M}$ from the posterior, the posterior predictive density can be estimated as:

$$p(y_{T+1:T+H}|y_{1:T}) \approx \frac{1}{M}\sum_{i=1}^{M} p(y_{T+1:T+H}|y_{1:T}, \theta^{(i)}). \tag{2.1}$$

2.1 Censored likelihood and censored posterior

As mentioned above, we are interested in a particular region of the predictive distribution, i.e. the left tail. For generality let us denote the region of interest by $A = \{A_1, \ldots, A_T\}$, where $A_t = \{y_t|y_t < C_t\}$ with threshold C_t potentially time-varying. For assessing the performance of forecast methods, i.e. comparing accuracy of density forecasts for such a region, Diks et al. (2011) introduce the censored likelihood (CSL) scoring function, which Gatarek et al. (2013) employ to define the censored likelihood (CL), where the CL is obtained by taking the exponential transformation of the CSL. The CL is given by

$$p^{cl}(y_{1:T}|\theta) = \prod_{t=1}^{T} p^{cl}(y_t|\theta, y_{1:t-1}), \tag{2.2}$$

where $p^{cl}(y_t|\theta, y_{1:t-1})$ is the conditional density of the mixed continuous-discrete distribution for the censored variable $\tilde{y}_t$

$$\tilde{y}_t = \begin{cases} y_t, & \text{if } y_t \in A_t, \\ R_t, & \text{if } y_t \in A_t^C. \end{cases} \tag{2.3}$$

Definition (2.3) means that the censored variable $\tilde{y}_t$ is equal to the original one in the region of interest, while everywhere outside it it is equal to the value $R_t \in A_t^C$. In consequence, the distribution of $\tilde{y}_t$ is mixed: continuous (in A_t) and discrete (in R_t). We have:

$$\begin{aligned} p^{cl}(y_t|y_{1:t-1}\theta) &= [p(y_t|y_{1:t-1}, \theta)]^{I\{y_t \in A_t\}} \times \left[\mathbb{P}(y_t \in A_t^C|y_{1:t-1}, \theta)\right]^{I\{y_t \in A_t^C\}} \\ &= [p(y_t|y_{1:t-1}, \theta)]^{I\{y_t \in A_t\}} \times \left[\int_{A_t^C} p(x|y_{1:t-1}, \theta)dx\right]^{I\{y_t \in A_t^C\}}. \end{aligned} \tag{2.4}$$

Differently than with a likelihood of a *censored dataset* where all $y_t \in A_t^C$ are censored and their exact values are completely ignored, with the censored likelihood the exact value of $y_t \in A_t^C$ still plays a role in conditioning in subsequent periods, in the sense that we condition on the *uncensored* past observations $y_{t-1}, y_{t-2}, \ldots$. Only in the case of i.i.d. observations when $p(y_t|y_{1:t-1}, \theta) = p(y_t|\theta)$ both approaches would be equivalent. We do this 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 $y_{t-1}, y_{t-2}, \ldots$ 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 past observations). Therefore, the (Partially) Censored Posterior is a *quasi*-Bayesian concept.

Gatarek et al. (2013) use the CL to define the censored posterior (CP) density as

$$p^{cp}(\theta|y_{1:T}) \propto p(\theta)p^{cl}(y_{1:T}|\theta), \tag{2.5}$$

where $p(\theta)$ is the prior density kernel on the model parameters. That is, the CP does not result from Bayes' rule, that the posterior density is proportional to the product of prior density and likelihood; the CP is proportional to the product of prior density and censored likelihood. Typically, the censored posterior density $p^{cp}(\theta|y_{1:T})$ is a proper density in the same cases (i.e., under the same choices of the prior $p(\theta)$) where the regular posterior $p(\theta|y_{1:T})$ is a proper density (i.e., with finite integral $\int p(\theta)p^{cl}(y_{1:T}|\theta)d\theta < \infty$), as long as there are enough observations $y_t \in A_t$ that are not censored. Note that Berkowitz (2001) uses a censored approach to Value-at-Risk (VaR) testing and has a similar focus on large losses. In Appendix A we illustrate the advantages and disadvantages of estimation based on the censored posterior in a simple simulation study in which we consider three data generating processes (DGPs), where we assume a split normal distribution, a skewed-t distribution or a mixture of two normal distributions for i.i.d. y_t.

2.2 Partially Censored Posterior

Not all of the parameters are typically expected to particularly relate to the region of interest of the predictive distribution. For this reason we propose the *Partially Censored Posterior*, where only a selected subset of parameters is estimated with the conditional CP, while for the remaining parameters we consider the regular posterior.

2.2.1 Definition and MCMC algorithm *Conditional MitISEM*

Below we formally define the Partially Censored Posterior (PCP) and devise an MCMC algorithm to simulate from it. The PCP is a novel concept based on combining the standard posterior for the "common" parameters and the Censored Posterior of Gatarek et al. (2013) for the parameters that particularly affect the properties of the region of interest. Consider a vector of model parameters θ and suppose that some subset of parameters, call it θ_2, is particularly related to the (left) tail of the predictive distribution so that it may benefit from censoring, while the other parameters, in the subset θ_1, would not benefit from censoring, or could even be adversely affected by censoring. In other words, we consider a partitioning $\theta = (\theta_1', \theta_2')'$. How this partitioning is done depends on the model under consideration. We propose that a sensible way is to in collect θ_2 the parameters determining the shape of the conditional distribution of y_t (e.g., the degrees of freedom parameter of a Student's t distribution, the shape parameter of a Generalized Error Distribution), but also parameters for the (unconditional) mean and variance. Next, we propose to collect in θ_1 the other parameters, such as the coefficients determining the dynamic behaviour of the conditional mean/variance in ARMA/GARCH models.

Definition and algorithm We define the PCP as

$$p^{pcp}(\theta_1, \theta_2|y) = p(\theta_1|y)p^{cp}(\theta_2|\theta_1, y),$$

where $p(\theta_1|y)$ is the standard marginal posterior of θ_1 and $p^{cp}(\theta_2|\theta_1, y)$ is the *conditional* censored posterior of θ_2 given θ_1. For a given value of θ_1, a kernel of the *conditional* censored posterior density of θ_2

given θ_1 is given by:

$$p^{cp}(\theta_2|\theta_1, y) = \frac{p^{cp}(\theta_1, \theta_2|y)}{p^{cp}(\theta_1|y)} \propto p^{cp}(\theta_1, \theta_2|y) \propto p(\theta_1, \theta_2)p^{cl}(y|\theta_1, \theta_2),$$

with prior density kernel $p(\theta_1, \theta_2)$ and censored likelihood $p^{cl}(y|\theta_1, \theta_2)$ in (2.2). We propose the following MCMC procedure to simulate from the PCP, the *Conditional MitISEM* method:

1. Simulate $(\theta_1^{(i)}, \theta_2^{(i)})$, $i = 1, \ldots, M$, from posterior $p(\theta_1, \theta_2|y)$ using the independence chain Metropolis-Hastings (IC-MH) algorithm, using as a candidate density a mixture of Student's t densities obtained by applying the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) algorithm of Hoogerheide et al. (2012) to the posterior density kernel $p(\theta_1, \theta_2|y)$.
2. Keep $\theta_1^{(i)}$ and ignore $\theta_2^{(i)}$, $i = 1, \ldots, M$.
3. For each $\theta_1^{(i)}$ simulate $\theta_2^{(i,j)}$, $j = 1, \ldots, N$, from the conditional censored posterior $p^{cp}(\theta_2|\theta_1^{(i)}, y)$:
   3.1. Construct joint candidate density $q_{mit}(\theta_1, \theta_2)$, a mixture of Student's t densities obtained by applying the MitISEM algorithm to the censored posterior density kernel $p^{cp}(\theta_1, \theta_2|y)$;
   3.2. Use conditional candidate density $q_{cmit}(\theta_2|\theta_1 = \theta_1^{(i)})$, the mixture of Student's t densities implied by the joint candidate density $q_{mit}(\theta_1, \theta_2)$, as a candidate density to simulate $\theta_2^{(i,j)}$ from $p^{cp}(\theta_2|\theta_1^{(i)}, y)$ in a run of the independence chain MH algorithm.

The use of MitISEM in step 3.1. implies that this step is efficiently performed with a relatively high acceptance rate in the IC-MH algorithm. To perform the conditional sampling in step 3.2. we use the fact that the conditional distribution of a joint mixture of Student's t distributions is itself a mixture of Student's t distributions and we provide its details in Appendix B.

This implies that if we have obtained $q_{mit}(\theta_1, \theta_2)$, a mixture of Student's t densities that approximates the joint censored posterior $p^{cp}(\theta_1, \theta_2|y)$, then we can use the M implied conditional mixtures of Student's t densities $q_{cmit}(\theta_2|\theta_1 = \theta_1^{(i)})$, $(i = 1, \ldots, M)$, as candidate densities for $p^{cp}(\theta_2|\theta_1^{(i)}, y)$ $(i = 1, \ldots, M)$. Hence, we only need one MitISEM approximation to obtain all the conditional candidate densities. In step 3.2. we do need a separate run of the IC-MH algorithm to simulate $\theta_2^{(i,j)}$ for each given $\theta_1^{(i)}$ $(i = 1, \ldots, M)$. However, given the typically high quality of the conditional MitISEM candidate density, a small burn-in will typically suffice, after which we can choose to use $N = 1$ draw $\theta_2^{(i,j)}$. Note that step 3.2. can be performed in a parallel fashion. As an alternative, to further speed up the simulation method with only a small loss of precision, we can also choose to use $N \geq 2$ draws $\theta_2^{(i,j)}$ $(j = 1, \ldots, N)$ from each run, for example $N = 10$, combined with a thinning approach for $\theta_1^{(i)}$, where only every Nth draw of $\theta_1^{(i)}$ is used.

2.2.2 Variance reduction with *PCP-QERMit*

Putting much effort in obtaining more accurate estimates of risk measures such as VaR and ES, using the specific left-tail focus of the PCP, might be wasteful if counteracted by large simulation noise affecting these estimates (i.e. high numerical standard errors). Hence, we aim to increase numerical efficiency of the proposed PCP method. For this purpose, we adapt the *Quick Evaluation of Risk using Mixture of t approximations* (QERMit) algorithm of Hoogerheide and van Dijk (2010) for efficient VaR and ES estimation.

QERMit is an importance sampling (IS) based method in which an increase in efficiency is obtained by oversampling "high-loss" scenarios and assigning them lower importance weights. The theoretical result of Geweke (1989) prescribes that the optimal importance density (in the sense of minimising the numerical standard error for a given number of draws) for Bayesian estimation of a probability of a given

set (here, the left tail of the predictive distribution) is composed of two equally weighted components, one for the high-loss scenarios (corresponding to the tail) and one for remaining realisations of returns. I.e. there is a 50%-50% division between "high-loss" draws and other draws. Such an approach allows for a substantial increase in efficiency compared to the so-called *direct approach* for VaR evaluation, in which predictions are obtained by simply sampling posterior draws of model parameters and combining these with the future innovations from the model to generate future paths of returns. One then simply computes the VaR estimate as the required percentile of the sorted (in ascending order) simulated returns. The QERMit method of Hoogerheide and van Dijk (2010) works for the regular (uncensored) Bayesian approach, i.e. based on the regular posterior and the regular predictive distribution. This method does require a closed-form formula for the target density, which is used as the numerator of the IS weights in the final step where the draws from the importance distribution are used to estimate the VaR. In case of the PCP we do not have a closed-form formula for the target density $p^{pcp}(\theta_1, \theta_2|y) = p(\theta_1|y)p^{cp}(\theta_2|\theta_1, y)$, since we do not have closed-form formulas for the density kernels $p(\theta_1|y)$ and $p^{cp}(\theta_2|\theta_1, y)$.

New IS algorithm To overcome this problem, we propose a new IS-based method to reduce the variance of the H-step-ahead VaR estimator obtained with the PCP. Given the draws of $(\theta_1^{(i)}, \theta_2^{(i)})$, $i = 1, \ldots, M$, from the PCP, we aim to sample the future innovations in the model $\varepsilon_{T+1:T+H}$ *conditionally* on $(\theta_1^{(i)}, \theta_2^{(i)})$ such that the resulting joint draws $(\theta_1^{(i)}, \theta_2^{(i)}, \varepsilon_{T+1:T+H})$ will lead to "high losses". This relates to the idea of oversampling the negative scenarios underlying the QERMit approach of Hoogerheide and van Dijk (2010), however we do not require to evaluate the target density kernel of the PCP. The proposed *PCP-QERMit* algorithm proceeds as follows.

1. **Preliminary steps**

   1.1. Obtain a set of draws from the PCP, $(\theta_1^{(i)}, \theta_2^{(i)})$, $i = 1, \ldots, M$, using the *Conditional MitISEM* algorithm of the previous subsection.

   1.2. Simulate future innovations $\varepsilon_{T+1:T+H}^{(i)}$ from their model distribution.

   1.3. Calculate the corresponding future returns $y_{T+1:T+H}^{(i)}$.

   1.4. Consider those joint draws $(\theta_1^{(i)}, \theta_2^{(i)}, \varepsilon_{T+1:T+H}^{(i)})$ that have led to e.g. the 10% lowest returns $\sum_{t=T+1}^{T+H} y_t^{(i)}$ (the "high loss draws").

2. **High loss draws**

   2.1. Use the "high loss draws" from step 1.4. to approximate the joint PCP "high-loss" density of θ and $\varepsilon_{T+1:T+H}$ with a mixture of Student's t densities $q_{mit}(\theta_1, \theta_2, \varepsilon_{T+1:T+H})$ by applying the MitISEM algorithm to the draws $(\theta_1^{(i)}, \theta_2^{(i)}, \varepsilon_{T+1:T+H}^{(i)})$.

   2.2. Sample $\tilde{\varepsilon}_{T+1:T+H}^{(i)}|\theta_1^{(i)}, \theta_2^{(i)}$, $i = 1, \ldots, M$, from its conditional importance density (aimed at high losses) $q_{cmit}(\varepsilon_{T+1:T+H}|\theta_1^{(i)}, \theta_2^{(i)})$, the conditional mixture of Student's t distributions implied by $q_{mit}(\theta_1, \theta_2, \varepsilon_{T+1:T+H})$ (cf. Appendix B).

3. **IS estimation of the VaR (or ES)**

   3.1. Compute the importance weights of the draws $(\theta_1^{(i)}, \theta_2^{(i)}, \tilde{\varepsilon}_{T+1:T+H}^{(i)})$, $i = 1, \ldots, M$, as

$$w^{(i)} = \frac{p(\tilde{\varepsilon}_{T+1:T+H}^{(i)}|\theta_1^{(i)}, \theta_2^{(i)})}{q(\tilde{\varepsilon}_{T+1:T+H}^{(i)}|\theta_1^{(i)}, \theta_2^{(i)})},$$

   where the numerator $p(\varepsilon_{T+1:T+H}^{(i)}|\theta_1^{(i)}, \theta_2^{(i)})$ is simply the density of the innovations in the model (and where the kernel of the partially censored posterior density $p^{pcp}(\theta_1, \theta_2|y) = p(\theta_1|y)p^{cp}(\theta_2|\theta_1, y)$ drops out of the importance weight, as it appears in both numerator and denominator).

3.2. Compute the future returns $y_{T+1:T+H}^{(i)}$ corresponding to the joint draws $(\theta_1^{(i)}, \theta_2^{(i)}, \tilde{\varepsilon}_{T+1:T+H}^{(i)})$, $i = 1, \ldots, M$, and the resulting total return over H periods $\sum_{t=T+1}^{T+H} y_t$.

3.3. Estimate the $100(1-\alpha)\%$VaR as the value C such that

$$\widehat{\mathbb{P}}\left(\sum_{t=T+1}^{T+H} y_t < C\right) = \alpha,$$

with

$$\widehat{\mathbb{P}}\left(\sum_{t=T+1}^{T+H} y_t < C\right) = \frac{1}{M}\sum_{i=1}^{M} w^{(i)} \mathbb{I}\left(\sum_{t=T+1}^{T+H} y_t^{(i)} < C\right), \tag{2.6}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

For the ES the method continues in a similar fashion. Step 2.2. is crucial in the above algorithm, as it allows us to "guide" the future disturbances to the "high-loss" region without the necessity of evaluating the kernel of the partially censored posterior density $p^{pcp}(\theta_1, \theta_2|y) = p(\theta_1|y)p^{cp}(\theta_2|\theta_1, y)$. Note that we do not need to use the 50%-50% division between "high-loss" draws and other draws, which was the case in the regular QERMit method for Bayesian VaR/ES prediction, but we can fully focus on the high losses. Such a concentration of all the mass of the importance density in the "high-loss region" is valid since we do not use the *self-normalised* IS weights $w^{(i)}/\sum_{j=1}^{M} w^{(j)}$. Normalising of the IS weights is necessary in Bayesian IS estimation whenever only the posterior kernel is available. Since we have the exact target and candidate densities of the innovations $\varepsilon_{T+1:T+H}$, we use the *unscaled* IS weights $w^{(i)}$ that only occur in (2.6) in the product $w^{(i)}\mathbb{I}\left(\sum_{t=T+1}^{T+H} y_t^{(i)} < C\right)$, so that the weights $w^{(i)}$ only matter for "high-loss" draws for which the indicator $\mathbb{I}\left(\sum_{t=T+1}^{T+H} y_t^{(i)} < C\right)$ is equal to 1.

Illustration To illustrate the benefits of the PCP-QERMit method we consider a simple example involving the AR(1) model. We consider the true DGP of the form

$$y_t = \mu(1-\rho) + \rho y_{t-1} + \varepsilon_t,$$

with split normally distributed innovations $\varepsilon_t \sim \mathcal{SN}(\delta, \tau_1^2, \tau_2^2)$ with $\delta = \frac{\tau_2 - \tau_1}{\sqrt{2\pi}}$ so that $E(\varepsilon_t) = 0$, see Appendix A for a brief discussion of this split normal distribution. We simulate $T = 1000$ observations from the model with $\mu = 0$, $\tau_1 = 1$, $\tau_2 = 2$ and $\rho = 0.8$.

We estimate the AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. We specify the usual non-informative prior $p(\mu, \sigma, \rho) \propto \frac{1}{\sigma}$ (for $\sigma > 0$, $-1 < \rho < 1$).

We estimate the 1-step-ahead 99.5%, 99% and 95% VaR and ES (and compute the numerical standard error from 50 MC replications) using the PCP where $\theta_1 = \{\rho\}$ stems from the regular marginal posterior, whereas $\theta_2 = \{\mu, \sigma\}$ stems from the conditional censored posterior. Both the PCP direct approach (Conditional MitISEM) and the PCP-QERMit method make use of 10000 draws. (The PCP has a time-constant threshold C_t given by the 10% quantile of the in-sample data.) Table 1 shows the results, where the smaller numerical standard errors stress the usefulness of the PCP-QERMit method for obtaining more accurate estimates of both VaR and ES.

2.3 Simulation study: AR(1) model

Below, we compare the quality of the left-tail density forecasts from the PCP with the regular posterior and the full CP. We consider the same estimated model and the same DGP as in the previous subsection: an estimated AR(1) model with normally distributed innovations for data from an AR(1) model with split normally distributed innovations.

Risk measure	PCP direct approach	PCP-QERMit
99.5% VaR	-4.3557	-4.3379
	[0.1050]	[0.0500]
99.5% ES	-4.9877	-4.9786
	[0.1328]	[0.0830]
99% VaR	-3.8461	-3.8308
	[0.0813]	[0.0340]
99% ES	-4.5311	-4.5183
	[0.1003]	[0.0587]
95% VaR	-2.4682	-2.4675
	[0.0429]	[0.0100]
95% ES	-3.3130	-3.3055
	[0.0524]	[0.0228]

Table 1: Estimated AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ for $T = 1000$ observations from DGP of AR(1) model with split normally distributed innovations $\varepsilon_t \sim \mathcal{SN}(\delta = \frac{1}{\sqrt{2\pi}}, \tau_1 = 1, \tau_2 = 2)$. Results of estimated 1-step-ahead 99.5%, 99% and 95% VaR and ES (and numerical standard error from 50 MC replications within brackets). The PCP direct approach (Conditional MitISEM) and PCP-QERMit method make use of 10000 draws. (The PCP has a time-constant threshold C_t given by the 10% quantile of the in-sample data.)

We keep $\mu = 0$, $\rho = 0.8$ and $\tau_1 = 1$ in the DGP. We do vary the level of misspecification by considering the correctly specified case of $\tau_2 = 1$ and the misspecified cases of $\tau_2 = 1.5$ and $\tau_2 = 2$. Further, we analyse the effect of the sample size T by considering estimation windows of size $T = 100$, 200, 500 and 1000.

For each DGP we consider 1000 out-of-sample observations for 20 simulated datasets, where for each observation we compute the (one-step-ahead) censored likelihood (CSL) score function of Diks et al. (2011) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns), given by

$$\begin{aligned} S^{csl}(p(y_{T+1}|y_{1:T})) &= \mathbb{I}(y_{T+1} < C_{T+1}) \log p(y_{T+1}|y_{1:T}) \\ &\quad + \mathbb{I}(y_{T+1} \geq C_{T+1}) \log \left(\int_{C_{T+1}}^{\infty} p(s|y_{1:T}) ds \right). \end{aligned} \tag{2.7}$$

For each simulated dataset we compute the Diebold-Mariano test statistic (with Newey-West standard error; see Diebold and Mariano, 1995), where the loss differential is the difference in the censored likelihood score function. We use the average of the 20 Diebold-Mariano test statistics to test the null hypothesis of equal left-tail density prediction, where the critical values in a two-sided test at 5% significance are simply given by $\pm\frac{1.96}{\sqrt{20}} \approx \pm 0.44$ (as the 20 simulated datasets are independent, and the test statistics have approximately the standard normal distribution under the null). The standard Bayesian concept of the Bayes factor is not suitable in our situation. First, if we would use the Bayes factor for all uncensored data, then the (partially) censored posterior would be expected to perform substantially worse than the standard posterior, since the (partially) censored posterior only aims to provide a good prediction of the predictive distribution in the region of interest, i.e. the left tail. Outside the region of interest the standard posterior is expected to provide much better density forecasts. Second, the Bayes factor for the censored data (conditioning on past censored observations) would be hard to evaluate and would also not reflect the purpose of the (partially) censored posterior to improve the left-tail prediction based on all information provided by the actually observed past observations.

Table 2 shows the results. We observe the following findings. First, as expected, in the case without misspecification ($\tau_2 = 1$), the regular posterior performs better than the PCP or CP. In this case it is obviously optimal to use all observations in an uncensored way. Moreover, in this case the PCP performs better than the CP, as "the less censoring, the better". Second, in the cases of misspecification and a large estimation window ($T = 500$ or $T = 1000$) the PCP and CP outperform the regular posterior. The more severe the misspecification, the smaller the sample size T for which censoring becomes beneficial.

Third, in the case of misspecification and a small estimation window ($T = 100$ or $T = 200$) the regular posterior outperforms the CP and the PCP, caused by the loss of information due to censoring. Fourth, the PCP is never significantly outperformed by the CP. In the case of misspecification and a large estimation window, we do not reject the equality of their performance. In the cases of no misspecification and/or a small estimation window the PCP significantly outperforms the CP.

In order to analyse the robustness of our conclusions with respect to the choice of the quality measure and the distribution of the errors in the AR(1) model, we perform a similar study based on the 99.5%, 99% and 95% VaR (instead of the censored likelihood score function), where we simulate $T = 100$, 1000 or 10000 draws y_t from the AR(1) model where the errors have the skewed-t distribution $\mathcal{SKT}(0, 1, \nu = 5, \lambda)$ of Hansen (1994), see Appendix A for a brief discussion of this skewed-t distribution. Table 3 shows the results. The conclusions are similar to those for the censored likelihood score function in the AR(1) model with the split-normal errors. The PCP outperforms the regular posterior if the misspecification is large enough (i.e., if the asymmetry parameter λ in the DGP is far enough from 0), if the VaR of interest lies deep enough in the left tail and if the number of observations T is large enough.

T	$\tau_2 = 1$	$\tau_2 = 1.5$	$\tau_2 = 2$
100	7.379***	5.868***	2.137***
200	4.315***	1.097***	**-0.872*** **
500	5.261***	-0.367	**-1.221*** **
1000	2.026***	**-0.959*** **	**-1.648*** **

(a) Posterior vs PCP.

T	$\tau_2 = 1$	$\tau_2 = 1.5$	$\tau_2 = 2$
100	4.471***	3.957***	1.894***
200	2.987***	1.458***	-0.739***
500	1.923***	0.065	-1.370***
1000	1.084***	-0.778***	-1.810***

(b) Posterior vs CP.

T	$\tau_2 = 1$	$\tau_2 = 1.5$	$\tau_2 = 2$
100	**-1.561*** **	**-2.157*** **	**-2.312*** **
200	**-2.041*** **	**-0.924*** **	-0.419*
500	**-1.410*** **	-0.135	0.320
1000	**-0.857*** **	0.031	-0.157

(c) CP vs PCP.

Table 2: Estimated AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ for $T = 100, 200, 500, 1000$ observations from DGP of AR(1) model with split normally distributed innovations $\varepsilon_t \sim \mathcal{SN}(\delta = \frac{\tau_2 - \tau_1}{\sqrt{2\pi}}, \tau_1 = 1, \tau_2)$. We consider the correctly specified case of $\tau_2 = 1$ and the misspecified cases of $\tau_2 = 1.5$ and $\tau_2 = 2$. The tables show the average of 20 Diebold-Mariano test statistics (with Newey-West standard errors) for 20 simulated data sets. The loss differential (computed for $H = 1000$ out-of-sample observations for each simulated dataset) is the difference in the censored likelihood score function (2.7) with time-constant threshold $C_t = C$ given by the 5% quantile of the returns. Positive values indicate superior left-tail forecast performance of the first approach; negative values indicate superior left-tail forecast performance of the second approach. The significance (in a two-sided test) is indicated by * for $p \leq 0.1$, ** for $p \leq 0.05$ and *** for $p \leq 0.01$. Bold numbers indicate a significantly better performance of our proposed PCP approach (at 5% significance level).

	λ	-0.50	-0.40	-0.30	-0.20	-0.10	0.10	0.20	0.30	0.40	0.50
$T = 100$											
	CP 10% - posterior	-6.13	-3.09	-2.39	-1.01	-0.24	3.82	5.69	3.91	0.25	-1.30
VaR 99.5%	PCP 10% - posterior	-6.43	-3.48	-2.96	-1.34	-0.41	2.86	3.93	0.60	-3.10	-6.16
	PCP 10% - CP10%	-0.98	-2.55	-2.16	-1.80	-1.10	-2.16	-5.16	-5.17	-5.41	-6.85
	CP 10% - posterior	-2.95	-0.38	-0.14	1.13	1.65	5.83	6.71	3.12	0.03	-1.01
VaR 99%	PCP 10% - posterior	-3.69	-1.25	-0.88	0.35	1.12	4.10	3.67	-0.81	-4.30	-7.12
	PCP 10% - CP10%	-1.93	-3.88	-3.16	-3.00	-2.56	-3.42	-6.79	-5.11	-5.68	-6.92
	CP 10% - posterior	3.95	7.17	5.87	5.53	6.24	5.57	6.69	2.36	1.66	1.25
VaR 95%	PCP 10% - posterior	2.45	4.97	3.90	3.96	3.91	2.39	0.68	-3.83	-5.24	-8.01
	PCP 10% - CP10%	-3.88	-5.37	-5.52	-4.28	-4.05	-5.63	-7.83	-4.89	-5.58	-6.55
$T = 1000$											
	CP 10% - posterior	-45.45	-45.30	-37.78	-30.44	-11.16	-0.18	8.44	-3.17	-16.35	-38.39
VaR 99.5%	PCP 10% - posterior	-46.53	-42.38	-33.51	-32.68	-10.28	-1.83	6.62	-4.71	-16.70	-44.72
	PCP 10% - CP10%	-3.95	-4.40	-5.49	-4.16	-2.70	-5.46	-3.28	-4.56	-4.41	-4.02
	CP 10% - posterior	-39.40	-24.82	-21.84	-17.71	-6.85	5.73	3.58	-7.86	-20.30	-35.84
VaR 99%	PCP 10% - posterior	-44.91	-23.78	-23.23	-21.51	-6.59	2.63	1.03	-10.65	-21.22	-43.33
	PCP 10% - CP10%	-5.36	-4.60	-5.97	-5.58	-2.93	-5.42	-3.38	-4.35	-3.84	-3.65
	CP 10% - posterior	3.61	4.97	6.54	6.90	10.63	0.80	-3.83	-8.20	-19.69	-23.39
VaR 95%	PCP 10% - posterior	3.10	4.22	5.69	5.89	9.00	-0.61	-7.92	-14.73	-21.63	-40.91
	PCP 10% - CP10%	-0.13	0.92	0.10	-0.89	0.02	-2.12	-2.88	-3.90	-2.94	-3.87
$T = 10000$											
	CP 10% - posterior	-118.70	-52.62	-77.94	-48.23	-39.92	-3.58	13.93	-7.85	-37.13	-55.64
VaR 99.5%	PCP 10% - posterior	-137.96	-62.99	-90.51	-53.61	-36.44	-6.62	11.04	-7.80	-36.74	-77.33
	PCP 10% - CP10%	-3.91	-4.94	-5.86	-6.24	-4.21	-3.42	-4.14	-1.93	-2.27	-3.19
	CP 10% - posterior	-105.01	-45.95	-70.72	-38.43	-24.25	1.30	-1.13	-36.57	-44.67	-46.73
VaR 99%	PCP 10% - posterior	-166.18	-86.78	-139.48	-65.77	-35.01	-3.07	-4.99	-40.53	-51.26	-65.40
	PCP 10% - CP10%	-10.37	-7.53	-10.61	-8.51	-5.73	-4.13	-5.80	-2.35	-3.07	-1.37
	CP 10% - posterior	5.62	3.90	9.97	9.73	20.83	-2.69	-17.52	-20.66	-30.08	-33.05
VaR 95%	PCP 10% - posterior	8.65	6.67	13.13	16.23	18.47	-0.28	-17.91	-22.48	-46.90	-57.81
	PCP 10% - CP10%	4.24	4.02	6.77	2.39	2.34	3.34	2.41	-0.15	-1.95	-1.20

Table 3: Estimated AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ for $T = 100, 1000, 10000$ observations from DGP of AR(1) model where the errors have the skewed-t distribution $\mathcal{SKT}(0, 1, \nu = 5, \lambda)$ of Hansen (1994). Results of t-statistic in the Diebold-Mariano test with 50 or 20 loss differentials given by the differences in RMSE of the estimated VaR for 50 simulated datasets (for $T = 100, 1000$) or 20 simulated datasets (for $T = 10000$). A value ≥ 1.96 (≤ -1.96) means that the mean of the RMSEs is (asymptotically) significantly larger (smaller) for the first approach than for the second approach at 5% significance level.

3 Time-varying threshold

Notice that the region of interest A_t used to define the censored variable in (2.3) is potentially time-varying. However, to the best of our knowledge, the literature on the censored likelihood scoring function, the censored likelihood and the censored posterior has been limited to a time-constant threshold. Gatarek et al. (2013) set the "censoring boundary" to the 20% or 30% percentile of the estimation window, leaving the topic of a time-varying threshold for further research. Opschoor et al. (2016) focus on the 15% percentile of a two-piece Normal distribution or a certain percentile (15% or 25%) of the empirical distribution of the data. Diks et al. (2011) investigate the impact of a time-varying threshold, which, however, is understood slightly differently. These authors evaluate the forecasting methods using a rolling window scheme and set the time-varying constant equal to the empirical quantile of the observations in the relevant estimation window. Obviously, a time-constant threshold implied by a certain empirical percentile differs between different data windows.

However, a constant threshold might be suboptimal when we focus on the left tail of the conditional distribution (given past observations). Even if the interest is in the unconditional left tail, so only in the most negative returns, then the time-varying threshold might be still more advantageous than the time-constant one. This is simply because the time-varying threshold provides more information about the left tail of the distribution of the standardized innovations compared to the time-constant one.

Therefore, we consider the time-varying threshold C_t given by a certain percentile of the estimated conditional distribution of y_t (given the past) that is implied by the Maximum Likelihood Estimate (MLE) $\hat{\theta}_{ML}$. Note that the threshold C_t must be equal for all draws $\theta^{(i)}$ $(i = 1, \ldots, M)$ from the (partially) censored posterior, as the threshold C_t affects the (partially) censored posterior. Making C_t depend on draws $\theta^{(i)}$ $(i = 1, \ldots, M)$ from the (partially) censored posterior would lead to a circular reasoning. Hence, the MLE $\hat{\theta}_{ML}$ provides a usable solution. As an alternative, one could use the regular posterior mean of θ.

The above discussion relates to *estimation* based on a (partially) censored posterior. However, note that the choice of a threshold C_{T+1} can also be important *for the assessment of the quality of the left-tail prediction*. Indeed, (2.7) can be computed with time-varying C_{T+1}. In our empirical study in Section 4 we consider, next to time-constant thresholds for the CSL rule (the 0.5%, 1% and 5% percentiles of the in-sample data), time-varying thresholds given by the 0.5%, 1% and 5% percentiles of the MLE-implied conditional distribution.

4 Empirical application

In this section we compare the left-tail forecasting performance for the regular posterior, the censored posterior and the partially censored posterior using empirical data. We consider daily logreturns of the IBM stock, from the 4th January 2007 to the 28th December 2018 (3019 observations, see Figure 4.1).

We consider three models. The first model is the AGARCH(1,1) model, the Asymmetric GARCH model of Engle and Ng (1993), with innovations following the skewed-t distribution of Hansen (1994). This model accounts for the skewness and the leverage effect often observed for stock returns. We adopt the following specification

$$
\begin{aligned}
y_t &= \mu_1 + \sqrt{h_t}\varepsilon_t, \\
\varepsilon_t &\sim \mathcal{SKT}(0, 1, \nu, \lambda), \\
h_t &= \omega(1 - \alpha - \beta) + \alpha(y_{t-1} - \mu_2)^2 + \beta h_{t-1},
\end{aligned}
$$

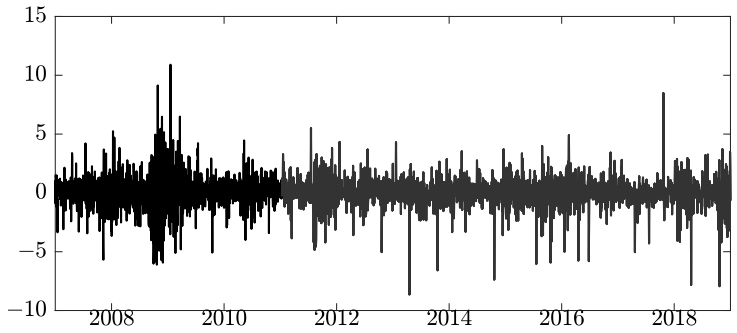


Figure 4.1: The daily logreturns of the IBM stock from the 4th January 2007 to the 28th December 2018.

where $\mathcal{SKT}(0,1,\nu,\lambda)$, denotes the skewed-$t$ distribution of Hansen (1994) with zero mean, unit variance, ν degrees of freedom and skewness parameter λ. We put flat priors on variance dynamics parameters to impose its positivity and stationarity: $\omega > 0$, $\alpha \in (0,1)$, $\beta \in (0,1)$ with $\alpha + \beta < 1$. For $\nu - 2$ we use an uninformative yet proper exponential prior (with prior mean 100) and for $\lambda \sim \mathcal{U}(-1,1)$.

Creal et al. (2013) propose the Generalised Autoregressive Score (GAS) model in which a time-varying parameter is updated with the scaled score of a new observation's contribution to the loglikelihood function. Our second and third model are GAS models with skewed-t and Student's t innovations. Neither of these GAS models accounts for a leverage effect, and only the GAS-skewed-t model accounts for skewness. The GAS(1,1)-skewed-t model (with time-varying parameter given by the logarithm of the variance $\log(h_t)$) is given by the following specification

$$
\begin{aligned}
y_t &= \mu + \sqrt{h_t}\varepsilon_t, \\
\varepsilon_t &\sim \mathcal{SKT}(0,1,\nu,\lambda), \\
\log(h_t) &= \omega + A\left(\frac{(\nu+1)bz_{t-1}(bz_{t-1}+a)}{2(\nu-2)(1+I_{t-1}\lambda)^2 + 2(bz_{t-1}+a)^2} - \frac{1}{2}\right) + B\log(h_{t-1}), \\
I_{t-1} &= \begin{cases} -1, & z_{t-1} < -a/b, \\ 1, & z_{t-1} \geq -a/b, \end{cases} \\
z_{t-1} &= \frac{y_{t-1}-\mu}{\sqrt{h_{t-1}}},
\end{aligned}
$$

where the constants a, b and c are given by $a = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right)$, $b^2 = 1 + 3\lambda^2 - a^2$ and $c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}$.

The GAS(1,1)-t model (with time-varying parameter given by the variance h_t) is given by the following specification

$$
\begin{aligned}
y_t &= \mu + \sqrt{\frac{\nu-2}{\nu}h_t}\varepsilon_t, \\
\varepsilon_t &\sim t(\nu), \\
h_t &= \omega + A\frac{\nu+3}{\nu}\left(\frac{(\nu+1)(y_{t-1}-\mu)^2}{\nu - 2 + \frac{(y_{t-1}-\mu)^2}{h_{t-1}}} - h_{t-1}\right) + Bh_{t-1}.
\end{aligned}
$$

That is, in the GAS(1,1)-skewed-t and GAS(1,1)-t models we have parameter vectors $\theta = (\mu, \omega, A, B, \nu, \lambda)'$ and $\theta = (\mu, \omega, A, B, \nu)'$. We put flat priors on μ, ω, A and B, with $\omega > 0$ in the GAS(1,1)-t model and $B \in (0,1)$. For $\nu - 2$ we use an uninformative yet proper exponential prior (with prior mean 100). For the GAS(1,1)-skewed-t model we specify $\lambda \sim \mathcal{U}(-1,1)$.

As a benchmark and the starting point for the PCP approach, we first carry out the standard posterior

analysis; second, we perform the estimation based on the CP. Each time we run $M = 10000$ iterations (after a burn-in of 1000) of the IC-MH using as a candidate the mixture of Student's t distributions obtained with the MitISEM algorithm of Hoogerheide et al. (2012) For the PCP, given the posterior draws of $\theta_1 = \{\mu_2, \alpha, \beta\}$ or $\theta_1 = \{A, B\}$ of the parameters describing the dynamics (including the parameter μ_2 in the AGARCH model, which describes the leverage effect), we conditionally sample $\theta_2 = \{\mu_1, \omega, \nu, \lambda\}$, $\theta_2 = \{\mu, \omega, \nu, \lambda\}$ or $\theta_2 = \{\mu, \omega, \nu\}$ from the conditional censored posterior. θ_2 contains the parameters that determine the unconditional mean and variance and the shape of the distribution of y_t, these parameters are particularly related to the left tail of the predictive distribution of y_t. For the threshold C_t we consider multiple quantiles, both the constant value given by the quantile of the in-sample data and the time-varying quantile of the MLE-implied conditional distribution.

In our forecasting study we consider $H = 2007$ out-of-sample density forecasts, where we have an in-sample period of $T = 1012$ observations. As our primary interest is accurate left-tail density prediction, we compare the density forecasts based on the censored likelihood (CSL) scoring rule (2.7) of Diks et al. (2011). A novelty of this paper is that we also allow the threshold *for the assessment of the quality of the left-tail prediction* to be time-varying, which we set to the 0.5%, 1% and 5% percentile of the MLE-implied conditional distribution. We also consider a time-constant threshold for evaluation, as in the previous literature, which we set at the 0.5%, 1% and 5% percentile of the in-sample data.

Tables 4, 5 and 6 present the results of the Diebold-Mariano test based on the censored likelihood scoring rule with time-constant and time-varying threshold, respectively, for the estimated AGARCH(1,1) model with skewed-t errors, the GAS(1,1)-skewed-t model and the GAS(1,1)-t model, respectively. A positive number indicates that the first approach provides better left-tail density forecasts (in terms of the CSL) than the second approach.

Table 7 gives a summary of the results. For example, in the AGARCH(1,1)-skewed-t model the PCP beats the regular posterior in 40 out of 48 cases (with 4 constant and time-varying quantiles for the estimation of the PCP, and 3 constant and time-varying quantiles for the assessment of the quality), where the outperformance is significant (at 5% level) in 29 cases. On the other hand, in the AGARCH(1,1)-skewed-t model the PCP is only beaten by the regular posterior in 8 out of 48 cases, never significantly. We observe that in each of the three models the role of *partial* censoring is crucial. With multiple parameters to be estimated based on a dataset where many observations have been censored, it is harder for the fully censored posterior to provide accurate left-tail density forecasts. With an appropriately chosen subset of parameters to apply censoring, we can often achieve better left-tail density forecasts than with the regular posterior or the fully censored posterior. However, we note that we expect the results to be contingent on the data used. After all, if a model is not misspecified (or if the misspecification is negligible), then we do not expect the (partially) censored posterior to outperform the standard posterior.

In the AGARCH(1,1)-skewed-t model the time-varying threshold (during estimation) leads to better PCP results than its counterpart with a constant threshold (during estimation) in 22 out of 24 cases (with 4 quantiles for the estimation of the PCP, and 3 constant and time-varying quantiles for the assessment of the quality), where the outperformance is significant (at 5% level) in 21 cases. This stresses the potential usefulness of making the threshold for estimation of the partially censored posterior time-varying. However, in the GAS(1,1)-skewed-t model the better performance of the PCP with a time-varying threshold is observed less often, and in the GAS(1,1)-t model the PCP with a constant threshold appears to perform better. This suggests that the preference for a time-varying threshold or a constant threshold (for the estimation of the partially censored posterior) may crucially depend on the model specification.

		time-constant threshold for evaluation of CSL: quantile			time-varying threshold for evaluation of CSL: quantile		
		0.5%	1%	5%	0.5%	1%	5%
20%	PCP (const. threshold) - posterior	-1.21	-0.87	1.16	-0.41	0.12	1.95
	PCP (const. threshold) - CP (const. threshold)	2.47	2.32	0.11	2.07	1.50	0.20
	CP (const. threshold) - posterior	-1.76	-1.46	0.82	-1.08	-0.47	1.32
	PCP (time-var. threshold) - posterior	-0.71	-0.33	2.15	2.09	3.67	5.35
	PCP (time-var. threshold) - CP (time-var. threshold)	2.56	2.43	0.52	2.56	2.43	0.52
	CP (time-var. threshold) - posterior	-1.78	-1.49	1.09	1.83	3.46	5.26
	PCP (time-var. threshold) - PCP (const. threshold)	2.51	2.38	3.22	2.28	3.84	5.55
30%	PCP (const. threshold) - posterior	-0.29	0.27	3.04	0.62	1.23	4.31
	PCP (const. threshold) - CP (const. threshold)	1.54	1.21	-2.12	0.49	-0.50	-1.07
	CP (const. threshold) - posterior	-0.69	-0.18	3.11	0.31	1.08	3.15
	PCP (time-var. threshold) - posterior	-0.28	0.22	2.74	2.16	3.72	5.30
	PCP (time-var. threshold) - CP (time-var. threshold)	2.36	2.22	1.17	2.36	2.22	1.17
	CP (time-var. threshold) - posterior	-1.69	-1.34	0.69	1.97	3.58	5.15
	PCP (time-var. threshold) - PCP (const. threshold)	0.35	-0.43	-3.65	2.19	3.73	5.13
40%	PCP (const. threshold) - posterior	2.37	2.62	3.82	2.74	2.94	5.34
	PCP (const. threshold) - CP (const. threshold)	2.30	2.20	2.06	1.86	1.78	3.07
	CP (const. threshold) - posterior	-1.18	-0.83	0.12	-0.27	-0.08	-0.28
	PCP (time-var. threshold) - posterior	2.51	2.70	3.78	2.79	2.98	5.17
	PCP (time-var. threshold) - CP (time-var. threshold)	2.33	2.20	2.17	1.92	1.86	3.30
	CP (time-var. threshold) - posterior	-0.50	-0.05	0.76	0.55	0.65	0.29
	PCP (time-var. threshold) - PCP (const. threshold)	2.63	2.67	3.44	2.74	2.86	4.41
50%	PCP (const. threshold) - posterior	0.63	-0.33	3.49	1.48	1.93	5.02
	PCP (const. threshold) - CP (const. threshold)	-0.21	-0.31	-3.61	-1.56	-2.48	-2.39
	CP (const. threshold) - posterior	0.74	-1.49	5.22	2.22	3.22	5.35
	PCP (time-var. threshold) - posterior	1.58	2.00	3.75	2.19	2.57	4.79
	PCP (time-var. threshold) - CP (time-var. threshold)	1.97	1.81	-0.06	1.43	0.87	0.83
	CP (time-var. threshold) - posterior	-0.10	0.52	4.34	1.52	2.46	4.41
	PCP (time-var. threshold) - PCP (const. threshold)	2.50	2.38	2.65	2.62	2.69	2.52

Table 4: Empirical application to daily IBM logreturns: estimated AGARCH(1,1) model of Engle and Ng (1993) with skewed-t innovations of Hansen (1994). Results of t-statistic in the Diebold-Mariano test with loss differentials (for the $H = 2007$ days in the out-of-sample period) given by the differences in censored likelihood (CSL) score function in (2.7), where the time-constant threshold for evaluation is the 0.5%, 1% or 5% percentile of the in-sample data, and where the time-varying threshold is the 0.5%, 1% or 5% percentile of the MLE-implied conditional distribution. For estimation of the (partially) censored posterior we use constant and time-varying thresholds given by the 20%, 30%, 40% and 50% quantiles, which are given by the percentiles of the in-sample data and the percentiles of the MLE-implied conditional distribution, respectively. A value ≥ 1.96 (≤ -1.96) means that the mean of the CSL is significantly larger (smaller) for the first approach than for the second approach (at 5% significance level).

		time-constant threshold for evaluation of CSL: quantile			time-varying threshold for evaluation of CSL: quantile		
		0.5%	1%	5%	0.5%	1%	5%
10%	PCP (const. threshold) - posterior	1.25	0.55	0.97	0.93	0.57	0.44
	PCP (const. threshold) - CP (const. threshold)	1.42	1.20	2.38	2.68	3.41	3.13
	CP (const. threshold) - posterior	0.51	-0.17	0.25	-0.03	-0.55	-0.57
	PCP (time-var. threshold) - posterior	1.53	0.92	1.16	1.17	0.80	0.64
	PCP (time-var. threshold) - CP (time-var. threshold)	-0.01	-0.71	0.46	0.88	1.41	1.15
	CP (time-var. threshold) - posterior	1.66	1.26	1.04	0.86	0.37	0.26
	PCP (time-var. threshold) - PCP (const. threshold)	2.41	2.37	2.27	2.78	2.66	2.50
20%	PCP (const. threshold) - posterior	1.99	1.58	1.79	1.92	1.57	1.35
	PCP (const. threshold) - CP (const. threshold)	1.43	0.94	1.01	1.55	1.77	1.69
	CP (const. threshold) - posterior	1.89	1.58	1.61	1.47	1.04	0.78
	PCP (time-var. threshold) - posterior	2.02	1.65	1.84	2.01	1.65	1.44
	PCP (time-var. threshold) - CP (time-var. threshold)	2.17	1.92	1.32	1.82	1.98	2.31
	CP (time-var. threshold) - posterior	1.48	1.09	1.39	1.31	0.83	0.37
	PCP (time-var. threshold) - PCP (const. threshold)	-0.75	-0.09	-1.02	-0.57	-0.51	-0.31
30%	PCP (const. threshold) - posterior	2.12	1.80	2.06	2.13	1.85	1.74
	PCP (const. threshold) - CP (const. threshold)	1.71	1.22	0.66	1.37	1.20	1.03
	CP (const. threshold) - posterior	2.07	1.83	2.04	1.88	1.58	1.39
	PCP (time-var. threshold) - posterior	2.32	2.11	2.43	2.52	2.35	2.36
	PCP (time-var. threshold) - CP (time-var. threshold)	1.06	1.04	-1.15	-0.13	-0.71	-0.50
	CP (time-var. threshold) - posterior	2.28	2.06	2.70	2.65	2.57	2.44
	PCP (time-var. threshold) - PCP (const. threshold)	1.74	2.51	-0.07	0.58	0.84	0.50
40%	PCP (const. threshold) - posterior	2.34	2.13	2.45	2.52	2.36	2.39
	PCP (const. threshold) - CP (const. threshold)	0.29	0.39	-1.65	-0.85	-1.37	-0.85
	CP (const. threshold) - posterior	2.26	2.03	2.76	2.68	2.65	2.48
	PCP (time-var. threshold) - posterior	2.36	2.20	2.66	2.65	2.61	2.76
	PCP (time-var. threshold) - CP (time-var. threshold)	-0.54	-0.24	-1.71	-1.34	-1.40	-0.62
	CP (time-var. threshold) - posterior	2.33	2.13	2.81	2.70	2.67	2.55
	PCP (time-var. threshold) - PCP (const. threshold)	-0.86	-0.29	0.16	-0.13	0.63	0.80

Table 5: Empirical application to daily IBM logreturns: estimated GAS(1,1) model with skewed-t innovations. Results of t-statistic in the Diebold-Mariano test with loss differentials (for the $H = 2007$ days in the out-of-sample period) given by the differences in censored likelihood (CSL) score function in (2.7), where the time-constant threshold for evaluation is the 0.5%, 1% or 5% percentile of the in-sample data, and where the time-varying threshold is the 0.5%, 1% or 5% percentile of the MLE-implied conditional distribution. For estimation of the (partially) censored posterior we use constant and time-varying thresholds given by the 10%, 20%, 30% and 40% quantiles, which are given by the percentiles of the in-sample data and the percentiles of the MLE-implied conditional distribution, respectively. A value ≥ 1.96 (≤ -1.96) means that the mean of the CSL is significantly larger (smaller) for the first approach than for the second approach (at 5% significance level).

		time-constant threshold for evaluation of CSL: quantile			time-varying threshold for evaluation of CSL: quantile		
		0.5%	1%	5%	0.5%	1%	5%
40%	PCP (const. threshold) - posterior	2.48	2.46	2.76	3.06	2.57	2.84
	PCP (const. threshold) - CP (const. threshold)	1.71	1.69	-0.86	1.06	0.04	-0.56
	CP (const. threshold) - posterior	2.18	2.29	3.05	3.14	2.71	2.78
	PCP (time-var. threshold) - posterior	1.84	1.59	-0.83	0.57	0.13	-3.94
	PCP (time-var. threshold) - CP (time-var. threshold)	2.66	2.24	3.35	3.25	3.00	5.93
	CP (time-var. threshold) - posterior	1.64	1.42	-2.06	-0.14	-0.70	-5.09
	PCP (time-var. threshold) - PCP (const. threshold)	1.40	1.11	-2.25	-0.85	-1.10	-4.64
50%	PCP (const. threshold) - posterior	2.65	2.47	3.05	3.12	2.46	3.42
	PCP (const. threshold) - CP (const. threshold)	1.65	1.60	-0.48	1.16	0.28	0.29
	CP (const. threshold) - posterior	2.39	2.21	3.04	3.08	2.39	2.92
	PCP (time-var. threshold) - posterior	2.55	2.42	2.30	2.74	2.71	1.19
	PCP (time-var. threshold) - CP (time-var. threshold)	0.49	0.54	-1.60	-0.27	-0.94	-0.71
	CP (time-var. threshold) - posterior	2.40	2.26	2.67	2.67	2.84	1.38
	PCP (time-var. threshold) - PCP (const. threshold)	1.88	1.86	-1.02	-0.42	0.23	-2.09

Table 6: Empirical application to daily IBM logreturns: estimated GAS(1,1) model with Student's t innovations of Creal et al. (2013). Results of t-statistic in the Diebold-Mariano test with loss differentials (for the $H = 2007$ days in the out-of-sample period) given by the differences in censored likelihood (CSL) score function in (2.7), where the time-constant threshold for evaluation is the 0.5%, 1% or 5% percentile of the in-sample data, and where the time-varying threshold is the 0.5%, 1% or 5% percentile of the MLE-implied conditional distribution. For estimation of the (partially) censored posterior we use constant and time-varying thresholds given by the 40% and 50% quantiles, which are given by the percentiles of the in-sample data and the percentiles of the MLE-implied conditional distribution, respectively. A value ≥ 1.96 (≤ -1.96) means that the mean of the CSL is significantly larger (smaller) for the first approach than for the second approach (at 5% significance level).

			AGARCH(1,1)-skewed-t						GAS(1,1)-skewed-t						GAS(1,1)-t					
PCP	-	posterior	40	-	8	[29	-	0]	48	-	0	[24	-	0]	22	-	2	[17	-	1]
PCP	-	CP	38	-	10	[20	-	4]	32	-	16	[7	-	0]	40	-	8	[6	-	0]
CP	-	posterior	29	-	19	[14	-	0]	44	-	4	[20	-	0]	20	-	4	[17	-	2]
PCP (time-var.)	-	PCP (const.)	22	-	2	[21	-	1]	14	-	10	[7	-	0]	5	-	7	[0	-	3]

Table 7: Number of cases in which the PCP, CP or posterior outperforms the other approach [or significantly outperforms at 5% significance level] out of 48 cases (with 4 constant and time-varying quantiles for the estimation of the PCP and CP, and 3 constant and time-varying quantiles for the assessment of the quality) for the AGARCH(1,1)-skewed-t and GAS(1,1)-skewed-t models, or out of 24 cases (with 2 constant and time-varying quantiles for the estimation of the PCP and CP, and 3 constant and time-varying quantiles for the assessment of the quality) for the GAS(1,1)-t model. And number of cases in which the PCP with time-varying threshold outperforms the PCP with constant threshold or vice versa [or significantly outperforms at 5% significance level].

5 Conclusions

We have proposed a novel approach to inference for a specific region of interest of the predictive distribution. Our Partially Censored Posterior method falls outside the framework of regular Bayesian statistics as we do not work with the regular likelihood but with the censored likelihood based on the censored likelihood scoring rule of Diks et al. (2011). This allows us to keep the merits of the regular Bayesian analysis, e.g. taking into account parameter uncertainty, and at the same time to allow for robust inference focused on the left tail in cases of potential model misspecification. The latter is vital for risk management, where the shape of the left tail of the conditional distribution is of crucial importance.

Partitioning of the parameter set into two subsets, one of which is likely to benefit from censoring, increases the precision of the parameter estimates compared to the fully censored posterior of Gatarek et al. (2013) and allows us to obtain better left-tail density forecasts. Further, we have introduced two novel simulation methods, the MCMC method of Conditional MitISEM and the importance sampling method of PCP-QERMit. Finally, we have considered novel ways of time-varying censoring, which allow us for an even better focus on the left tail of the distribution of the standardized innovations. We have demonstrated the usefulness of our methods in extensive simulation and empirical studies.

To further exploit the power of our quasi-Bayesian framework, in future research we intend to employ the PCP in the context of forecast combination via Model Averaging using partially censored predictive likelihoods, or in a (quasi-)Bayesian framework with time-varying weights for pairs of models and estimation methods (and possibly investment strategies), extending Bastürk et al. (2019). Also extensions of the classical approach of Opschoor et al. (2016) based on so-called pooling are relevant in this regard. Another interesting extension will be to investigate the impact of using the smoothly-censored likelihood of Diks et al. (2011) in our PCP setting, to make the PCP approach even more robust w.r.t. the choice of the threshold C_t. An important domain of application of the proposed PCP methodology would be portfolio optimization and portfolio risk management, where the evaluation of the probability of y_t lying outside the region of interest ($\mathbb{P}(y_t \in A_t^C | y_{1:t-1}, \theta)$) may require an efficient simulation method. An interesting extension would be the analysis of credit risk and defaults.

There are multiple possible applications beyond the field of financial econometrics. Risk estimation is of interest in many areas, not only in finance. For example, statistical models for weather forecasting and climatology. But also in financial econometrics quite different applications can be considered, such as in electricity markets where one may be particularly interested in the right tail of the distribution of energy prices.

Finally, models with latent variables (such as regime switching models and stochastic volatility models) and models with a realized variance measure would be interesting extensions. However, optimal simulation methods for such models would require an adaptation of the simulation methods presented in this article. We will consider such simulation methods in future research.

Bibliography

Aastveit, Knut Are, James Mitchell, Francesco Ravazzolo, and Herman K. Van Dijk (2019), "The Evolution of Forecast Density Combinations in Economics." To appear in *Oxford Research Encyclopedia of Economics and Finance*.

Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests." *Journal of Business & Economic Statistics*, 25, 177–190.

Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1999), "Coherent Measures of Risk." *Mathematical Finance*, 9, 203–228.

Ausín, M.C., and P. Galeano (2007). "Bayesian Estimation of the Gaussian Mixture GARCH Model", *Computational Statistics & Data Analysis*, 51(5), 2636–2652.

Bastürk, N., Borowska, A., Grassi, S., Hoogerheide, L.F., Van Dijk, H.K. (2019). "Forecast density combinations of dynamic models and data driven portfolio strategies", *Journal of Econometrics*, 210 (1), 170–186.

Berkowitz, J. (2001). "Testing Density Forecast, With Applications to Risk Management", *Journal of Business & Economic Statistics*, 19 (4), 465–474.

Creal, D., S. J. Koopman, and A. Lucas. (2013), "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, 28, 777–795.

Cross Validated (2017), "Why should I be Bayesian when my model is wrong?" `https://stats.stackexchange.com/questions/274815/why-should-i-be-bayesian-when-my-model-is-wrong`. Accessed: 2017-07-18.

De Roon, F. and P. Karehnke (2016), "A Simple Skewed Distribution with Asset Pricing Applications." *Review of Finance*, 1–29.

Diebold, F. X. and R. S. Mariano (1995), "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics*, 13, 253–263.

Diks, C., V. Panchenko, and D. van Dijk (2011), "Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails." *Journal of Econometrics*, 163, 215–230.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation." *Econometrica*, 50, 987–1007.

Engle, R. F. and V. K. Ng (1993), "Measuring and Testing the Impact of News on Volatility." *Journal of Finance*, 48, 1749–1778.

Gatarek, L. T., L. F. Hoogerheide, K. Hooning, and H. K. van Dijk (2013), "Censored Posterior and Predictive Likelihood in Bayesian Left-tail Prediction for Accurate Value at Risk Estimation." Technical Report TI 2013-060/III, Tinbergen Institute Discussion Paper.

Geweke, J. (1989), "Bayesian Inference in Econometric Models using Monte Carlo Integration." *Econometrica*, 57, 1317–1739.

Geweke, J. and G. Amisano (2010), "Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns." *International Journal of Forecasting*, 26, 216–230.

Geweke, J. and G. Amisano (2012), "Prediction with Misspecified Models." *The American Economic Review*, 102, 482–486.

Hall, S. G. and J. Mitchell (2007), "Combining Density Forecasts." *International Journal of Forecasting*, 23, 1–13.

Hansen, B. E. (1994), "Autoregressive Conditional Density Estimation." *International Economic Review*, 705–730.

Hoogerheide, L. F., A. Opschoor, and H. K. van Dijk (2012), "A Class of Adaptive Importance Sampling Weighted EM Algorithms for Efficient and Robust Posterior and Predictive Simulation." *Journal of Econometrics*, 171, 101–120.

Hoogerheide, L. F. and H. K. van Dijk (2010), "Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling." *International Journal of Forecasting*, 26, 231–247.

Kleijn, B. J. K. and A. W. van der Vaart (2006), "Misspecification in Infinite-Dimensional Bayesian Statistics." *The Annals of Statistics*, 34, 837–877.

McNeil, A. J. and R. Frey (2000), "Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach." *Journal of Empirical Finance*, 7, 271–300.

McNeil, A. J., R. Frey, and P. Embrechts (2015), *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

Müller, U. K. (2013), "Misspecification in Infinite-Dimensional Bayesian Statistics." *Econometrica*, 81, 1805–1849.

Opschoor, A., D. van Dijk, and M. van der Wel (2016), "Combining Density Forecasts using Focused Scoring Rules." *Tinbergen Institute Discussion Paper*, 14-090/III.

Robert, C. (2017), "Why should I be Bayesian when my model is wrong?" `https://xianblog.wordpress.com/2017/05/09/why-should-i-be-bayesian-when-my-model-is-wrong/`. Accessed: 18 July 2017.

Roth, M. (2013), "On the Multivariate t Distribution." Technical Report LiTH-ISY-R-3059, Automatic Control Group at Linköpings Universitet.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*. Wiley.

A Advantages and disadvantages of the censored posterior: a simulation study with i.i.d. data

To illustrate the advantages and disadvantages of estimation based on the censored posterior, we perform a simple simulation study in which we consider three data generating processes (DGPs), where we assume a split normal distribution, a skewed-t distribution or a mixture of two normal distributions for i.i.d. y_t.

The density of the split normal distribution $\mathcal{SN}(\delta, \tau_1^2, \tau_2^2)$, analysed by e.g. Geweke (1989) and De Roon and Karehnke (2016), is given by

$$p(y_t) = \begin{cases} \phi(y_t; \delta, \tau_1^2), & y_t > \delta, \\ \phi(y_t; \delta, \tau_2^2), & y_t \leq \delta, \end{cases}$$

where $\phi(x; m, s)$ denotes the Gaussian density with mean m and variance s evaluated at x. The mean of a random variable distributed according to $\mathcal{SN}(0, \tau_1^2, \tau_2^2)$, i.e. with a split at zero, is equal to $-\frac{\tau_2 - \tau_1}{\sqrt{2\pi}}$, which is non-zero for any asymmetric case. The variance is equal to $\kappa = \frac{1}{2}\left(\left(\tau_1^2 + \tau_2^2\right) - \frac{(\tau_2 - \tau_1)^2}{\pi}\right)$. Hence, shifting of the split point accordingly to the chosen parameters τ_1^2 and τ_2^2 allows us to consider a zero-mean random variable: $y_t \sim \mathcal{SN}(\delta, \tau_1^2, \tau_2^2)$ with $\delta := \frac{\tau_2 - \tau_1}{\sqrt{2\pi}}$ results in $\mathbb{E}[y_t] = 0$. The reason behind the use of the split normal distribution is to be able to obtain one correctly specified tail, when we estimate a model with a normal distribution $\mathcal{N}(\mu, \sigma^2)$.

We consider two cases of the true parameters of the DGP: a symmetric case with $\tau_1 = 1$ and $\tau_2 = 1$; and an asymmetric case with $\tau_1 = 1$ and $\tau_2 = 2$. In that latter case we set $\delta = \frac{1}{\sqrt{2\pi}}$ to impose $\mathbb{E}[y_t] = 0$. For both cases we generate $T = 100$, $T = 1000$ and $T = 10000$ observations from the true model. We are interested in evaluating the 95% and 99% VaR, i.e. in the estimation of the 5% and 1% quantiles of the distribution of y_t. For the symmetric case the true values for these quantities are -1.6449 and -2.3263, while for the asymmetric case -2.8908 and -4.2538.

For each case we estimate an i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model with unknown mean μ and variance σ^2. We specify the usual non-informative prior $p(\mu, \sigma) \propto \frac{1}{\sigma}$ (for $\sigma > 0$). We perform an estimation based on the uncensored posterior and two specifications for the censored posterior. In each the threshold value C is constant over time, $A_t = \{y_t : y_t \leq C\}$, where we consider two different values for the threshold C: one equal to the 10% quantile of the generated sample (CP10%) and another one equal to zero (CP0). In both cases all the uncensored observations stem from the left half of the distribution. In other words:

- for the regular posterior, all uncensored data are used;
- for CP0 all generated negative values are used, and all (generated) positive values contribute to the censored posterior via the probability that they are positive;
- for CP10% all values below the 10% quantile of the simulated dataset are used and all other values contribute to the censored posterior via the probability that they are larger than the quantile.

All the simulations are carried out with $M = 10000$ posterior draws after a burn-in of 1000 using an independence chain Metropolis-Hastings (IC-MH) algorithm with target density kernel (2.5) where the candidate density is a single Student's t distribution.

Tables 8a and 8b report simulation results for Monte Carlo (MC) experiments of 100 simulated datasets for the symmetric and asymmetric case, respectively. Figures A.1 and A.2 present kernel density estimates of the 99.5%, 99% and 95% VaR for a single simulation for $T = 100, 1000, 10000$ for the symmetric and asymmetric case, respectively. For example, for the 95% VaR the (censored) posterior density of $\mu - 1.645\sigma$

Value	True	Posterior	CP10%	CP0
		$T = 100$		
99% VaR	-2.3263	-2.1245	-2.2322	-2.1519
		[**0.5763**]	[0.6812]	[0.6041]
95% VaR	-1.6449	-1.4899	-1.4668	-1.4951
		[**0.2922**]	[0.3123]	[0.2986]
		$T = 1000$		
99% VaR	-2.3263	-2.0998	-2.1020	-2.1074
		[**0.5464**]	[0.5500]	[0.5476]
95% VaR	-1.6449	-1.4816	-1.4823	-1.4858
		[**0.2725**]	[0.2734]	[0.2735]
		$T = 10000$		
99% VaR	-2.3263	-2.0965	-2.0876	-2.0972
		[**0.5427**]	[0.5432]	[0.5428]
95%VaR	-1.6449	-1.4802	-1.4767	-1.4815
		[**0.2712**]	[0.2713]	[0.2713]

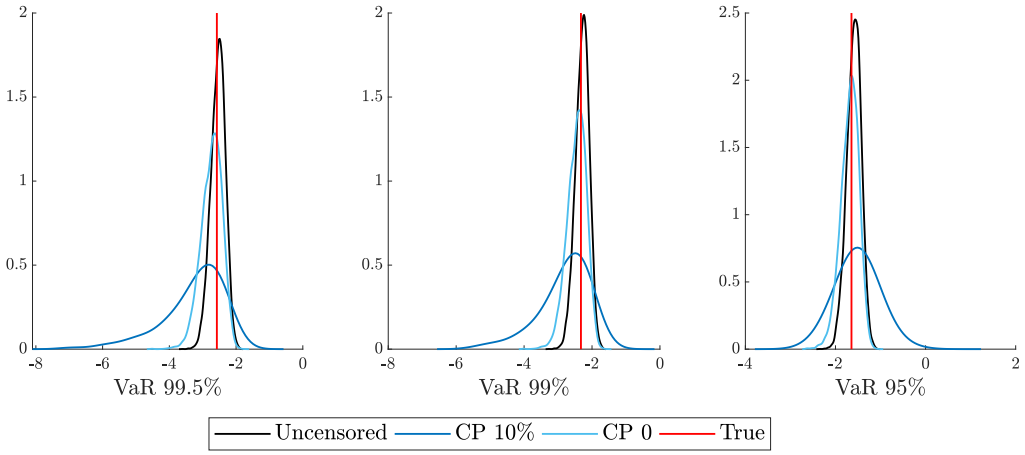
(a) Symmetric (correctly specified) case: $\tau_2 = 1$.

Value	True	Posterior	CP10%	CP0
		$T = 100$		
99% VaR	-4.2538	-3.6551	-4.5968	-4.4697
		[0.5082]	[0.6438]	[**0.3506**]
95% VaR	-2.8908	-2.5675	-2.8886	-2.9773
		[0.1984]	[0.2612]	[**0.1402**]
		$T = 1000$		
99% VaR	-4.2538	-3.5549	-4.2739	-4.2701
		[0.5063]	[0.0527]	[**0.0293**]
95% VaR	-2.8908	-2.5101	-2.8895	-2.8882
		[0.1540]	[0.0158]	[**0.0145**]
		$T = 10000$		
99% VaR	-4.2538	-3.5654	-4.2610	-4.2583
		[0.4787]	[0.0098]	[**0.0091**]
95% VaR	-2.8908	-2.5226	-2.8919	-2.8917
		[0.1369]	[0.0031]	[**0.0029**]

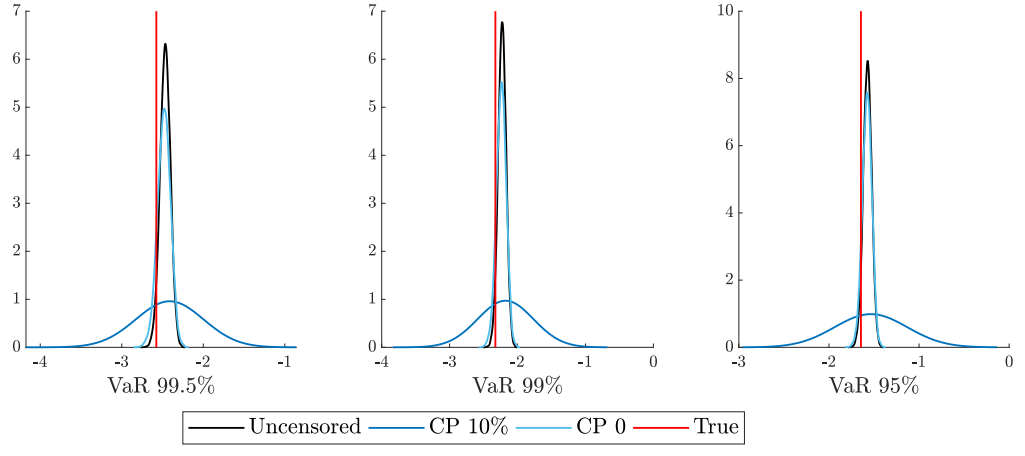
(b) Asymmetric (misspecified) case: $\tau_2 = 2$.

Table 8: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for data from DGP of (a) i.i.d. $y_t \sim \mathcal{SN}(\delta = 0, \tau_1 = 1, \tau_2 = 1)$ (which is equivalent with the standard normal distribution $\mathcal{N}(\mu = 0, \sigma = 1)$) and (b) i.i.d. split normal $y_t \sim \mathcal{SN}(\delta = \frac{1}{\sqrt{2\pi}}, \tau_1 = 1, \tau_2 = 2)$. Simulation results for the regular posterior and for the censored posterior with threshold at 0 (CP0) and threshold at the 10% data percentile (CP10%). MSEs across 100 simulated datasets in brackets, with the best MSE in boldface.

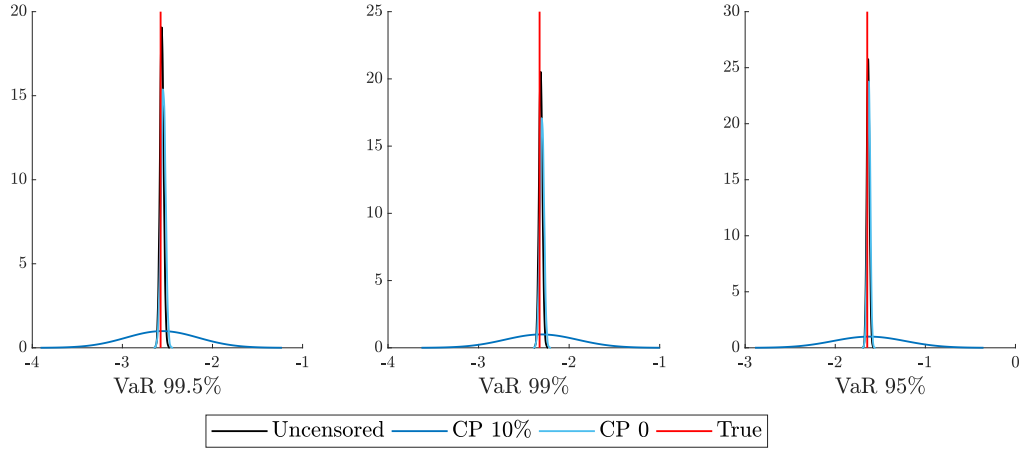
is shown. Note that the true values of the 99.5%, 99% and 95% VaR are the 0.5%, 1% and 5% percentiles of the $\mathcal{N}(\delta, \tau_2^2)$ distribution. In the misspecified case the regular posterior provides incorrect estimates from the left tail perspective, because the estimated model aims to approximate the distribution over the whole domain. The CP provides parameter estimates with a much better location (regarding the left tail of the predictive distribution) by focusing on the relevant region. The cost of a better location is, however, a larger variance of the estimates due to the loss of information caused by censoring. Obviously, the precision of the estimates from the CP depends on the degree of censoring: the more censoring, the less information, the lower the precision. In the symmetric case we can see that, as expected, the only cost of censoring is a higher variance, but the locations of the regular posterior and the CP are similar. In this specific case of the split normal distribution, where the left tail is perfectly described by the left tail of a normal distribution, the optimal threshold seems to be the one where we leave all observations from the left half uncensored, whereas we censor all observations from the right half. That is, a threshold equal to the value δ where the density 'jumps' between the left and right halves. The threshold 0 leads to better results than the threshold of the 10% quantile, since the 10% quantile lies further from the optimal threshold. We observe that for the larger datasets ($T = 1000$ and $T = 10000$) the VaR from the regular posterior is only slightly better (in the sense of a slightly smaller MSE) in the case of no misspecification (with a normal DGP), whereas in the case of misspecification (with a split normal DGP) the censored posterior leads to much more accurate VaR estimates. However, in case of a small dataset ($T = 100$) the VaR is substantially better for the regular posterior than for the censored posterior where the loss in precision due to censoring has a severe effect. We introduce the Partially Censored Posterior (PCP) exactly for the reason of limiting this harmful effect of loss of information due to censoring.



(a) $T = 100$

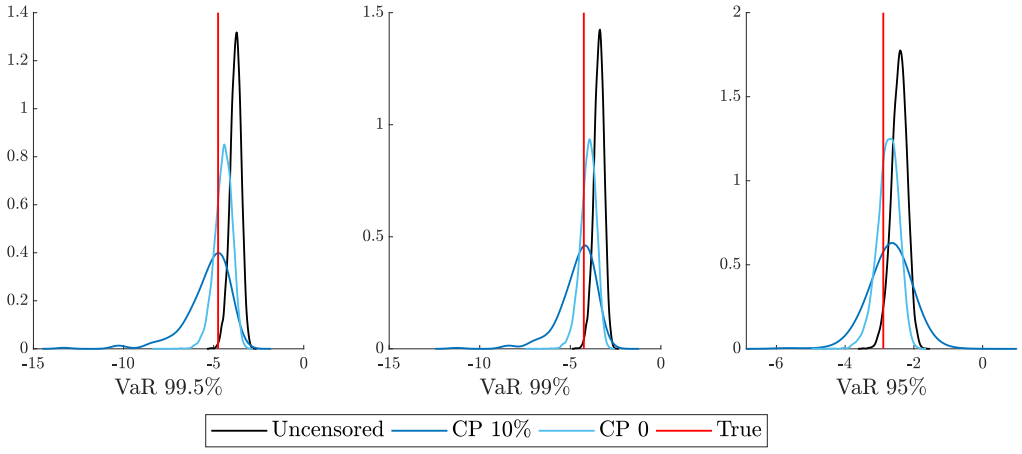


(b) $T = 1000$

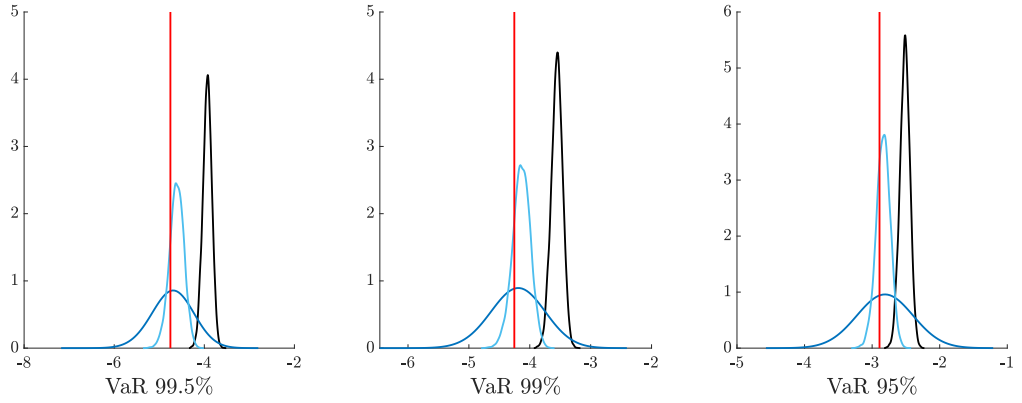


(c) $T = 10000$

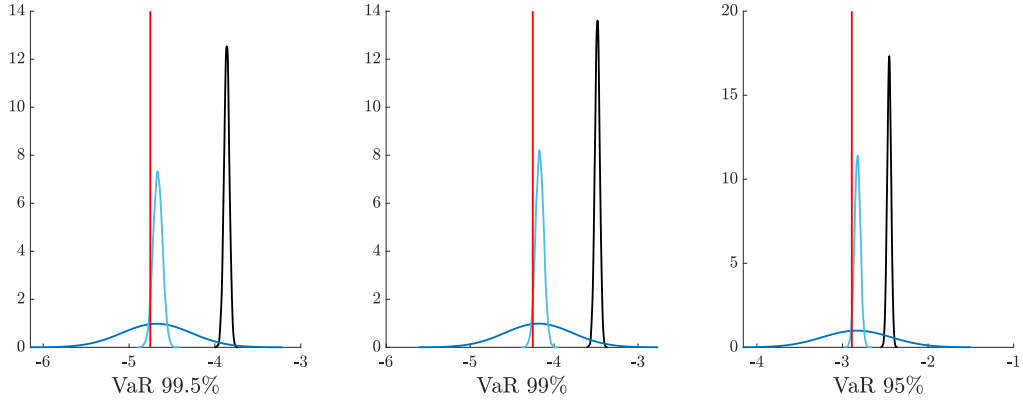
Figure A.1: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. split normal $\mathcal{SN}(\delta = 0,\ \tau_1 = 1,\ \tau_2 = 1)$ (which is equivalent with the standard normal distribution $\mathcal{N}(\mu = 0, \sigma = 1)$). Kernel density estimates of 99.5%, 99% and 95% VaRs obtained using regular posterior and censored posterior (CP) with threshold at 0 (CP0) and with threshold at the 10% data percentile (CP10%) together with the true VaR values. For example, for the 95% VaR the (censored) posterior density of $\mu - 1.645\sigma$ is shown.



(a) $T = 100$



(b) $T = 1000$



(c) $T = 10000$

Figure A.2: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. split normal $\mathcal{SN}(\delta = \frac{1}{\sqrt{2\pi}}, \tau_1 = 1, \tau_2 = 2)$. Kernel density estimates of 99.5%, 99% and 95% VaRs obtained using regular posterior and censored posterior (CP) with threshold at 0 (CP0) and with threshold at the 10% data percentile (CP10%) together with the true VaR values. For example, for the 95% VaR the (censored) posterior density of $\mu - 1.645\sigma$ is shown.

The discontinuous nature of the split normal density makes it very artificial for finance applications. A continuous density would better fit with standard modelling practices for return data. For this reason and to check the robustness of our results in case of different distributions where the estimated normal distributions will never be able to provide a perfect description of the left tail, we also consider simulated datasets from a mixture of normals and from a skewed-t distribution.

We consider the mixture of normals that is used by Ausín and Galeano (2007) for the standardized innovations in their Gaussian Mixture GARCH (1,1) model:

$$y_t \sim \begin{cases} \mathcal{N}(0, \sigma^2) & \text{with probability } \rho, \\ \mathcal{N}(0, \sigma^2/\lambda) & \text{with probability } 1-\rho, \end{cases} \tag{A.1}$$

where $\sigma^2 = \frac{1}{\rho+(1-\rho)/\lambda}$ so that $\text{var}(y_t) = 1$, and where $0 < \lambda < 1$. The inverse $\frac{1}{\lambda}$ indicates how much the variance in the 'wild' regime is amplified, where we consider multiple values of λ. The closer λ is to 0, the larger the kurtosis, and the larger the misspecification in the estimated model with the normal distribution. ρ is the probability of the 'calm' regime, which we set at 0.75.

For each value of λ and for each number of observations T (T = 100, 1000, or 10000) we simulate 100 datasets and use the different methods to estimate the 99.5%, 99% and 95% VaR. We perform the Diebold-Mariano test with loss differentials given by the 100 differences in absolute errors of the estimated VaR. Table 9 gives the results. We conclude the following. First, we obtain better results for censoring (as compared to the regular posterior) if we move from $T = 100$ to $T = 1000$ and $T = 10000$. If we have few observations, then the loss of information due to the censoring does more harm than when we have many observations. Second, censoring becomes more beneficial if the distribution of the DGP becomes "further" from the estimated normal distribution (with λ further from 1 and closer to 0). Third, for the 99.5% and 99% VaR CP10% performs better than CP0 (for T large enough and λ small enough), whereas for the 95% VaR CP0 performs better than CP10%. Using the 10% quantile as the threshold means that we have a more precise focus on the left tail, whereas with threshold 0 we have a broader focus on approximately the left half of the distribution. Fourth, for the 99.5% and 99% VaR censoring is more beneficial than for the 95% VaR. The 5% quantile is not far in the tail. However, for small values of λ (where the deviations from normality are substantial) the CP0 with large enough sample size T can still provide a more accurate 95% VaR than the regular posterior. The reason for this is that we estimate both μ and σ^2 of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. If we use CP0, then we only need to aim at fitting the shape of approximately the left half of the distribution, which is easier than aiming at the shape of the whole distribution. Note: if $\mu = 0$ would be fixed in the estimated distribution, then it would not matter that we only need to aim at the left half of the distribution, since both tails have the same shape. The DGP is a symmetric distribution. But with μ free we can use $\mu \neq 0$ to approximate the shape of the left half of the distribution better. A normal distribution with $\mu > 0$ and σ larger than the actual standard deviation can provide an approximation to a fat left tail. Censoring can also be useful when estimating a symmetric distribution if the DGP is a *different* symmetric distribution.

In the general case of misspecification the left tail is not perfectly described by the estimated model. Then there is typically a clear trade-off between the variance and the bias: the less censoring (that is, the less negative the threshold), the smaller the variance, but the larger the bias.

We also consider the skewed-t distribution of Hansen (1994): $\mathcal{SKT}(0, 1, \nu, \lambda)$ with zero mean, unit variance, ν degrees of freedom and skewness parameter λ, where $\lambda > 0$, $\lambda = 0$ and $\lambda < 0$ imply right-skewed, symmetric (Student's t) and left-skewed distributions, respectively. We take $\nu = 5$ to allow for the fat tails that are typical for data on financial returns and we consider multiple values of λ. Table 10 gives the results. From these we can draw similar conclusions as for the mixture of normal distributions. Note that

	λ	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.60	0.70	0.80	0.90
T=100															
	Posterior - CP0	-0.67	0.77	1.77	1.72	0.65	1.90	-0.81	-1.21	-2.37	-1.98	-3.53	-4.01	-3.83	-4.09
99.5% VaR	Posterior - CP10%	3.32	0.42	-0.98	-0.08	-2.45	-3.88	-4.55	-5.55	-7.52	-7.06	-8.98	-7.80	-9.55	-6.82
	CP0 - CP10%	3.16	0.15	-1.34	-0.63	-2.53	-4.48	-4.63	-5.32	-6.61	-6.89	-7.76	-6.63	-7.59	-5.66
	Posterior - CP0	-1.44	-0.42	-0.60	-0.07	-1.01	0.09	-2.20	-2.10	-3.86	-2.68	-3.28	-4.11	-4.01	-4.09
99% VaR	Posterior - CP10%	3.22	-0.19	-2.04	-1.59	-3.81	-5.20	-5.55	-5.75	-7.63	-6.99	-7.86	-6.81	-7.66	-5.80
	CP0 - CP10%	3.43	-0.05	-1.82	-1.49	-3.40	-5.43	-5.25	-5.56	-6.34	-6.26	-6.75	-5.30	-5.13	-4.40
	Posterior - CP0	-4.17	-2.00	-2.80	-0.82	-3.87	-3.20	-2.11	-2.54	-4.97	-2.52	-1.78	-3.10	-2.47	-1.99
95% VaR	Posterior - CP10%	-6.48	-4.60	-4.80	-2.91	-5.90	-3.36	-2.74	-2.88	-4.33	-3.81	-2.13	-3.57	-2.67	-1.97
	CP0 - CP10%	-3.93	-5.08	-3.74	-4.02	-3.63	-0.68	-1.44	-0.98	0.48	-2.05	-0.91	-1.53	-0.38	-0.39
T=1000															
	Posterior - CP0	-1.66	-0.63	0.26	-0.46	1.66	-0.87	1.86	0.28	-1.94	-0.86	-1.47	-1.34	-2.19	-0.55
99.5% VaR	Posterior - CP10%	50.55	42.34	35.03	29.61	23.05	14.24	12.47	5.75	2.29	1.79	-3.54	-3.59	-4.50	-3.52
	CP0 - CP10%	47.44	44.82	34.28	36.37	24.21	14.03	9.94	5.17	3.66	2.52	-2.58	-2.66	-2.91	-3.05
	Posterior - CP0	-3.07	-3.40	0.37	-1.60	0.25	-1.36	0.77	-0.17	-2.48	-1.31	-1.91	-1.62	-1.92	-0.64
99% VaR	Posterior - CP10%	48.91	39.04	28.88	22.00	16.36	5.65	5.12	1.03	-0.57	-0.55	-4.43	-3.51	-4.03	-3.10
	CP0 - CP10%	49.56	42.42	27.63	26.09	15.28	5.75	4.37	1.11	1.01	0.58	-2.90	-2.55	-3.02	-3.00
	Posterior - CP0	3.09	3.11	1.04	2.25	-0.69	-0.24	-0.43	-2.02	-2.97	-1.05	-2.26	-1.77	-1.11	-0.20
95% VaR	Posterior - CP10%	-1.91	-2.33	-5.10	-3.75	-5.39	-5.07	-3.12	-3.53	-3.53	-2.63	-2.67	-2.67	-0.52	-0.15
	CP0 - CP10%	-4.82	-6.10	-7.20	-6.36	-5.26	-5.57	-3.33	-2.09	-1.29	-2.24	-0.50	-1.13	0.61	0.02
T=10000															
	Posterior - CP0	-2.67	-2.00	-0.46	-2.52	0.14	-0.66	0.77	0.39	-0.03	0.36	-0.78	-0.59	-1.10	-0.35
99.5% VaR	Posterior - CP10%	100.21	74.23	67.66	49.69	39.04	32.48	24.08	19.55	15.82	11.02	0.11	-1.50	-1.79	-0.88
	CP0 - CP10%	101.49	83.81	63.07	52.51	38.95	37.47	24.51	21.74	15.76	10.05	0.88	-1.11	-0.63	-0.47
	Posterior - CP0	-5.84	-4.16	-1.44	-2.90	-0.09	-0.67	0.62	-0.16	-0.81	0.58	-0.31	-1.27	-2.21	-1.48
99% VaR	Posterior - CP10%	101.29	78.06	70.67	58.43	43.10	32.61	25.06	14.43	10.27	7.28	-0.05	-2.36	-2.79	-1.47
	CP0 - CP10%	119.18	88.04	72.54	60.92	39.47	32.06	23.65	15.19	10.58	6.14	0.27	-1.10	-0.80	0.13
	Posterior - CP0	10.78	11.37	8.08	6.17	5.26	2.44	0.50	-0.57	0.03	0.11	-0.79	-0.11	-0.19	-0.40
95% VaR	Posterior - CP10%	2.32	-1.93	-2.39	-4.21	-2.86	-4.39	-3.34	-5.03	-1.95	-2.98	-1.41	-1.17	-1.85	0.53
	CP0 - CP10%	-6.29	-10.37	-8.25	-9.87	-7.97	-5.69	-3.89	-4.76	-2.22	-3.32	-0.74	-0.91	-1.61	0.92

Table 9: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. y_t from the mixture of two normal distributions of Ausín and Galeano (2007) in (A.1), where λ closer to 0 implies a higher kurtosis (and larger deviation from the estimated normal distribution). Results of t-statistic in Diebold-Mariano test with loss differentials given by the 100 differences in absolute errors of the estimated VaR for 100 simulated datasets. A value ≥ 1.96 (≤ -1.96) means that the mean of the absolute errors is significantly larger (smaller) for the first approach than for the second approach (at 5% significance level).

for $\lambda = 0$ the DGP is a Student's t distribution, which is obviously different from the normal distribution that is estimated. Therefore also for $\lambda = 0$ censoring can help.

In order to analyse the effect of the threshold on the performance of the censored posterior we consider results for the skewed-t distribution for multiple thresholds C. Table 11 shows that the best threshold for estimating the censored posterior depends on the quantile that we are interested in. The deeper part of the tail we are interested in, the deeper in the tail lies the optimal threshold for censoring. For the 99.5% VaR the 5% quantile is typically the best of the considered thresholds, whereas for the 95% VaR the 20% quantile performs the best among the considered thresholds.

In practice one can perform a sensitivity analysis, where one compares the quality of forecasts using different threshold values for a hold-out sample of out-of-sample observations.

	λ	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60
T=100														
	Posterior - CP0	20.68	15.68	13.62	9.32	6.73	5.62	-2.01	-4.92	-2.39	1.12	12.26	17.98	17.19
99.5% VaR	Posterior - CP10%	5.12	2.95	2.64	0.39	-0.15	0.59	-3.61	-5.78	-5.38	-4.58	4.38	12.50	15.04
	CP0 - CP10%	-1.35	-1.72	-2.07	-2.95	-2.50	-0.97	-3.12	-3.70	-3.79	-4.89	-6.70	-5.93	0.48
	Posterior - CP0	14.09	10.11	8.92	5.90	3.92	3.36	-2.87	-3.82	0.32	4.70	15.68	18.29	17.33
99% VaR	Posterior - CP10%	3.11	0.94	0.72	-1.31	-1.45	-1.51	-4.86	-6.96	-4.13	-2.06	7.92	14.89	16.53
	CP0 - CP10%	-2.92	-3.31	-3.74	-4.71	-3.63	-2.69	-4.20	-5.24	-5.16	-6.05	-6.75	-2.29	3.09
	Posterior - CP0	0.38	-1.80	-2.07	-3.71	-2.35	-3.14	-2.10	1.71	5.07	9.49	17.01	18.18	15.24
95% VaR	Posterior - CP10%	0.44	-2.18	-2.00	-4.22	-3.80	-4.50	-3.31	0.80	4.43	8.88	15.87	19.34	16.27
	CP0 - CP10%	0.23	-1.84	0.23	-1.65	-4.44	-3.01	-3.10	-2.16	-2.47	-0.15	1.77	5.74	6.94
T=1000														
	Posterior - CP0	56.58	60.56	49.65	34.70	27.97	14.15	1.04	-12.99	-18.53	0.28	25.59	50.55	53.59
99.5% VaR	Posterior - CP10%	43.83	48.95	34.82	32.45	23.80	22.19	13.80	4.57	-6.24	3.98	24.53	40.51	57.55
	CP0 - CP10%	15.05	18.76	14.21	13.91	11.26	14.25	12.23	10.28	8.68	3.94	-1.43	-1.79	16.59
	Posterior - CP0	62.96	60.64	56.38	39.03	28.48	14.74	0.30	-11.26	-3.75	15.56	41.50	51.40	57.33
99% VaR	Posterior - CP10%	23.91	25.01	18.43	17.87	12.38	9.60	4.28	-1.87	-1.86	16.23	31.22	49.23	61.14
	CP0 - CP10%	3.77	5.37	3.71	4.93	4.01	4.59	3.93	3.43	1.98	-0.39	-6.10	5.31	23.86
	Posterior - CP0	1.33	0.43	-0.10	-0.49	-3.57	-6.10	-1.33	12.44	24.95	42.01	48.99	45.79	52.47
95% VaR	Posterior - CP10%	0.12	-0.88	-1.35	-2.13	-5.01	-8.22	-2.36	8.67	23.38	42.73	49.70	50.93	61.18
	CP0 - CP10%	-3.40	-3.88	-3.74	-4.70	-5.53	-7.01	-1.16	-3.88	-3.76	1.82	9.13	16.30	30.72
T=10000														
	Posterior - CP0	66.59	44.53	40.64	33.39	23.13	15.24	-0.24	-15.70	-31.65	2.32	49.13	99.48	114.94
99.5% VaR	Posterior - CP10%	85.49	43.86	66.49	76.87	61.54	46.08	29.02	13.41	-2.72	20.62	50.60	99.67	138.47
	CP0 - CP10%	38.39	18.69	30.45	39.65	41.56	33.43	33.92	41.92	24.95	20.10	5.80	3.56	36.52
	Posterior - CP0	67.20	49.51	49.94	38.70	30.12	16.97	-0.50	-18.90	-6.28	37.52	82.05	124.25	113.12
99% VaR	Posterior - CP10%	48.44	26.54	39.16	37.49	37.58	24.87	13.16	2.56	7.49	39.06	86.41	129.91	144.35
	CP0 - CP10%	13.52	7.99	11.03	14.04	15.26	15.25	13.76	16.08	15.45	3.19	-4.81	20.03	54.67
	Posterior - CP0	3.34	1.90	-0.30	-2.06	-4.94	-7.40	3.03	22.07	36.49	61.36	89.64	123.90	125.18
95% VaR	Posterior - CP10%	1.78	0.03	-2.22	-6.25	-10.27	-9.75	-2.93	12.73	29.94	57.40	91.59	150.92	146.80
	CP0 - CP10%	-3.46	-3.68	-2.87	-5.60	-8.38	-4.87	-6.38	-5.44	-2.40	1.09	13.73	28.41	48.75

Table 10: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. y_t from the skewed-t distribution $\mathcal{SKT}(0, 1, \nu = 5, \lambda)$ of Hansen (1994), where λ further from 0 implies a larger asymmetry (and larger deviation from the estimated normal distribution). Results of t-statistic in Diebold-Mariano test with loss differentials given by the 100 differences in absolute errors of the estimated VaR for 100 simulated datasets. A value ≥ 1.96 (≤ -1.96) means that the mean of the absolute errors is significantly larger (smaller) for the first approach than for the second approach (at 5% significance level).

	λ	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50
99.5% VaR	CP 5%	**0.41**	**0.27**	**0.36**	**0.24**	**0.33**	**0.16**	**0.17**	0.09	0.08	0.04	0.04
	CP 10%	0.46	0.33	0.46	0.26	0.34	0.21	0.23	0.10	0.08	0.06	0.03
	CP 15%	0.45	0.31	0.43	0.27	0.36	0.18	0.20	**0.09**	0.08	0.05	0.03
	CP 20%	0.73	0.68	0.58	0.48	0.39	0.32	0.21	0.14	**0.05**	**0.01**	**0.01**
99% VaR	CP 5%	0.13	0.14	**0.09**	0.08	0.12	**0.06**	**0.06**	0.04	0.03	0.02	0.01
	CP 10%	**0.11**	**0.09**	0.11	**0.06**	**0.09**	0.06	0.07	**0.03**	0.02	0.02	**0.01**
	CP 15%	0.17	0.11	0.16	0.10	0.13	0.07	0.07	0.04	0.03	0.02	0.01
	CP 20%	0.22	0.21	0.18	0.14	0.12	0.09	0.06	0.04	**0.01**	**0.00**	0.01
95% VaR	CP 5%	2.92	2.41	2.67	2.05	2.54	1.39	1.32	0.57	0.66	0.42	0.21
	CP 10%	1.15	0.91	1.04	0.85	0.88	0.49	0.48	0.20	0.24	0.16	0.08
	CP 15%	0.47	0.41	0.46	0.36	0.39	0.25	0.21	0.09	0.11	0.08	0.04
	CP 20%	**0.03**	**0.02**	**0.03**	**0.02**	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**

Table 11: Estimation results in i.i.d. normal $\mathcal{N}(\mu, \sigma^2)$ model for $T = 1000$ observations from DGP of i.i.d. y_t from the skewed-t distribution $\mathcal{SKT}(0, 1, \nu = 5, \lambda)$ of Hansen (1994), where λ further from 0 implies a larger asymmetry (and larger deviation from the estimated normal distribution). Results for the Mean Squared Error (MSE) of the estimated VaR for 100 simulated datasets, where different thresholds C have been used for estimation of the censored posterior. The lowest MSE value is reported in boldface.

B Conditional density of (mixture of) multivariate Student's t distributions

Student's t distribution Let $x \in \mathbb{R}^d$ follow the Student's t distribution with mode μ, scale matrix Σ and ν degrees of freedom, denoted $t(x; \mu, \Sigma, \nu)$, where we assume $\nu > 2$ so that $\text{var}(x) = \frac{\nu}{\nu-2}\Sigma$. Then, the probability density function (pdf) of x is given by (cf. Zellner, 1996; Roth, 2013)

$$p(x) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)(\pi\nu)^{\frac{d}{2}}}|\Sigma|^{-\frac{1}{2}}\left(1 + \frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{d+\nu}{2}}.$$

Next, consider a partitioning of x into $x = (x_1', x_2')'$ with x_1 and x_2 of dimensions d_1 and d_2, respectively. The corresponding parameter partitionings are then

$$\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \qquad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}.$$

Then, the conditional density of x_2 given x_1 is also a Student's t density, which is given by

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p(x_1)} = t(x_2; \mu_{2|1}, \Sigma_{2|1}, \nu_{2|1}),$$

with

$$\begin{aligned}
\mu_{2|1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \\
\Sigma_{2|1} &= \frac{\nu + (x_1 - \mu_1)'\Sigma_{11}^{-1}(x_1 - \mu_1)}{\nu + d_1}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right), \\
\nu_{2|1} &= \nu + d_1.
\end{aligned}$$

Mixture of Student's t distributions The above result extends to mixtures of Student's t distributions. Now let x follow an H component mixture of Student's t distributions $t(x; \mu_h, \Sigma_h, \nu_h)$, with component probabilities η_h, $h = 1, \ldots, H$, so that its pdf is given by

$$p(x) = \sum_{h=1}^{H} \eta_h t(x; \mu_h, \Sigma_h, \nu_h).$$

Let z denote a (latent) H-dimensional vector indicating from which component the observation x stems: if x stems from the hth component then $z = e_h$, the hth vector of the standard basis of $\mathbb{R}^H$, i.e. $z_h = 1$ and $z_l = 0$ for $l \neq h$. Obviously, unconditionally $\mathbb{P}[z = e_h] = \eta_h$. The conditional probability of x stemming from the hth component is

$$\begin{aligned}
\mathbb{P}[z = e_h|x] &= \frac{p(z = e_h, x)}{p(x)} \\
&= \frac{\mathbb{P}[z = e_h]p(x|z = e_h)}{\sum_{m=1}^{H}\mathbb{P}[z = e_m]p(x|z = e_m)} \\
&= \frac{\eta_h t(x; \mu_h, \Sigma_h, \nu_h)}{\sum_{m=1}^{H}\eta_m t(x; \mu_m, \Sigma_m, \nu_m)}.
\end{aligned}$$

Then, the conditional density of x_2 given x_1 is given by

$$p(x_2|x_1) = \frac{p(x_1,x_2)}{p(x_1)} = \frac{\sum_{h=1}^{H} \eta_h t(x;\mu_h,\Sigma_h,\nu_h)}{\sum_{h=1}^{H} \eta_h t(x_1;\mu_{h,1},\Sigma_{h,1},\nu_h)} = \sum_{h=1}^{H} \eta_{h,2|1} t(x_2;\mu_{h,2|1},\Sigma_{h,2|1},\nu_{h,2|1}),$$

with

$$\begin{aligned}
\mu_{h,2|1} &= \mu_{h,2} + \Sigma_{h,21}\Sigma_{h,11}^{-1}(x_1 - \mu_{h,1}), \\
\Sigma_{h,2|1} &= \frac{\nu_h + (x_1 - \mu_{h,1})'\Sigma_{h,11}^{-1}(x_1 - \mu_{h,1})}{\nu_h + d_1} \left(\Sigma_{h,22} - \Sigma_{h,21}\Sigma_{h,11}^{-1}\Sigma_{h,12}\right), \\
\nu_{h,2|1} &= \nu_h + d_1,
\end{aligned}$$

and with adjusted component probabilities

$$\eta_{h,2|1} = \mathbb{P}[z = e_h|x_1] = \frac{\eta_h t(x_1;\mu_{h,1},\Sigma_{h,11},\nu_h)}{\sum_{m=1}^{H} \eta_m t(x_1;\mu_{m,1},\Sigma_{m,11},\nu_m)}.$$

This implies that if we have obtained $q_{mit}(\theta_1,\theta_2)$, a mixture of Student's t densities that approximates the joint censored posterior $p^{cp}(\theta_1,\theta_2|y)$, then we can use the M implied conditional mixtures of Student's t densities $q_{cmit}(\theta_2|\theta_1 = \theta_1^{(i)})$ $(i = 1,\ldots,M)$ as candidate densities for the conditional censored posterior densities $p^{cp}(\theta_2|\theta_1^{(i)},y)$ $(i = 1,\ldots,M)$. Hence, we only need one MitISEM approximation to obtain all the conditional candidate densities in our proposed *Conditional MitISEM* method.