Preliminary Draft. Please do not cite, circulate or quote without the authors' permission

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

May 8, 2018

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 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.

*We would like to thank the participants of the 1st EcoSta 2017 conference (Hong Kong, 15–17 June 2017) and the 8th ESOBE 2017 (Maastricht, 26–27 October 2017) for their insightful comments. We are also grateful to Marc Baardman, Paolo Gorgi and Anne Opschoor for providing useful comments and suggestions. 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.

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. (2018b).

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 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. (2014) 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 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 Value-at-Risk and Expected Shortfall 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 use 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, discuss the censored posterior and present a simple toy example to illustrate potential benefits and disadvantages of 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 a GARCH model with Student's t innovations for daily IBM logreturns. Section 5 concludes.

2 Censored posterior and partially censored posterior

2.1 Censored likelihood and 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 $100(1-\alpha)\%$ quantile of the predictive distribution of $\sum_{t=T+1}^{T+H} y_t$ given $y_{1:T}$. We also consider the Expected Shortfall (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:

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

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 scoring (CLS) function, which Gatarek et al. (2014) employ to define the censored likelihood (CL), where the CL is obtained by taking the exponent 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. observation 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. (2014) use the CL to define the censored posterior (CP) density as

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

where $p(\theta)$ is the prior density kernel on the model parameters.

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 $\int p^{cl}(y_{1:T}|\theta)p(\theta)d\theta < \infty$), as long as there are enough observations $y_t \in A_t$ that are not censored.

2.1.1 Censored posterior: advantages and disadvantages in toy application of (split) normal distribution

To illustrate the advantages and disadvantages of estimation based on the censored posterior, we start with a toy simulation study in which we consider as the data generating process (DGP) for y_t an i.i.d. split normal $\mathcal{SN}(\mu, \sigma_1^2, \sigma_2^2)$ model. The split normal density, analysed by e.g. Geweke (1989) and De Roon and Karehnke (2016), is given by

$$p(y_t) = \begin{cases} \phi(y_t; \mu, \sigma_1^2), & y_t > \mu, \\ \phi(y_t; \mu, \sigma_2^2), & y_t \leq \mu, \end{cases}$$

where $\phi(x; m, s)$ denotes the Gaussian density with mean m and variance s evaluated at x.

We consider two cases of the true parameters of the DGP: a symmetric case with $\sigma_1 = 1$ and $\sigma_2 = 1$; and an asymmetric case with $\sigma_1 = 1$ and $\sigma_2 = 2$. In that latter case we set $\mu = \frac{1}{\sqrt{2\pi}}$ 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 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, $A_t = \{y_t : y_t \leq C\}$, and we consider two different values for the threshold C: one equal to the 10% quantile of the generated sample and another one equal to zero, where in both cases all the uncensored observations stem from the left half of the distribution. All the simulations are carried out with $M = 10000$ posterior draws after a burn-in of 1000 using an independence chain Metropolis-Hastings algorithm with target density kernel (2.5) the candidate density being a single Student's t distribution.

Tables 1a and 1b report simulation results for 100 Monte Carlo experiments for the symmetric and asymmetric case, respectively. Figure 2.1 presents kernel density estimates of the asymmetric case for a single simulation for $T = 1000$; we refer to Appendix A for the plots for the remaining cases. 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 since censoring leads to an analysis based on effectively a smaller sample. 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. 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, the VaR is substantially better for the regular posterior than for the censored posterior in case of a small dataset ($T = 100$) where the loss in precision due to censoring has a severe effect. We introduce the Partially Censored Posterior (PCP) in the next subsection, exactly for the reason of limiting this harmful effect of loss of information due to censoring.

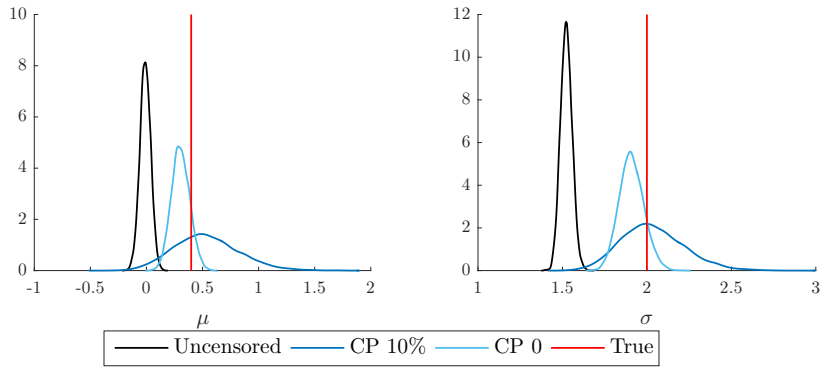


Figure 2.1: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 1000$ observations from DGP of i.i.d. split normal ($\sigma_1 = 1$, $\sigma_2 = 2$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).

2.2 Partially Censored Posterior

The previous subsection illustrated the advantages and disadvantages of the CP with respect to obtaining accurate evaluations of lower quantiles of the predictive posterior distribution: the CP has clearly a better location in case of misspecification, but this comes at the price of a lower precision of the estimates. Moreover, the estimated i.i.d. normal model had only consider 2 parameters whereas obviously most models have many more parameters, so that obtaining precise estimates becomes even harder. However, 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 Partially Censored Posterior: definition and MCMC simulation algorithm *Conditional MitISEM*

Below we formally define the Partially Censored Posterior (PCP) and devise a Markov chain Monte Carlo (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. (2014) 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 distribution so that it may benefit from censoring, while the other parameters, in the subset θ_1, should rather not be censored. 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.,

Value	True	Posterior	CP10%	CP0
$T = 100$				
μ	0.3989	-0.0147	0.9452	0.5321
		(0.1658)	(1.0768)	(0.3171)
σ	2.0000	1.6414	2.5853	2.2724
		(0.1157)	(0.7684)	(0.2935)
99% VaR	-4.2620	-3.6551	-4.5968	-4.4697
		[**0.5082**]	[0.6438]	[0.3506]
95% VaR	-2.8924	-2.5675	-2.8886	-2.9773
		[**0.1984**]	[0.2612]	[0.1402]
$T = 1000$				
μ	0.3989	-0.0103	0.5348	0.3043
		(0.0481)	(0.2895)	(0.0811)
σ	2.0000	1.5229	2.0338	1.9095
		(0.0343)	(0.1872)	(0.0732)
99% VaR	-4.2476	-3.5549	-4.2739	-4.2701
		[0.5063]	[0.0527]	[**0.0293**]
95% VaR	-2.8921	-2.5101	-2.8895	-2.8882
		[0.1540]	[0.0158]	[**0.0145**]
$T = 10000$				
μ	0.3989	0.0334	0.4279	0.4290
		(0.0152)	(0.0901)	(0.0273)
σ	2.0000	1.5125	1.9825	1.9778
		(0.0106)	(0.0568)	(0.0250)
99% VaR	-4.2610	-3.5654	-4.2610	-4.2583
		[0.4787]	[0.0098]	[**0.0091**]
95% VaR	-2.8940	-2.5226	-2.8919	-2.8917
		[0.1369]	[0.0031]	[**0.0029**]

(b) Asymmetric (misspecified) case: $\sigma_2 = 2$.

Value	True	Posterior	CP10%	CP0
$T = 100$				
μ	0.0000	0.0926	1.5715	0.1727
		(0.1014)	(1.5261)	(0.1529)
σ	1.0000	1.0226	1.8779	1.1289
		(0.0741)	(0.9051)	(0.1418)
99% VaR	-2.0976	-2.1245	-2.2322	-2.1519
		[**0.5763**]	[0.6812]	[0.6041]
95% VaR	-1.4804	-1.4899	-1.4668	-1.4951
		[**0.2922**]	[0.3123]	[0.2986]
$T = 1000$				
μ	0.0000	0.0071	0.0196	0.0230
		(0.0304)	(0.1473)	(0.0387)
σ	1.0000	0.9604	0.9446	0.9725
		(0.0215)	(0.0921)	(0.0349)
99% VaR	-2.0927	-2.0998	-2.1020	-2.1074
		[**0.5464**]	[0.5500]	[0.5476]
95% VaR	-1.4804	-1.4816	-1.4823	-1.4858
		[**0.2725**]	[0.2734]	[0.2735]
$T = 10000$				
μ	0.0000	0.0031	0.0300	-0.0049
		(0.0100)	(0.0433)	(0.0123)
σ	1.0000	0.9960	1.0053	0.9865
		(0.0071)	(0.0281)	(0.0111)
99% VaR	-2.1032	-2.0965	-2.0876	-2.0972
		[**0.5427**]	[0.5432]	[0.5428]
95%VaR	-1.4822	-1.4802	-1.4767	-1.4815
		[**0.2712**]	[0.2713]	[0.2713]

(a) Symmetric (correctly specified) case: $\sigma_2 = 1$.

Table 1: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for data from DGP of i.i.d. normal $N(\mu = 0, \sigma = 1)$ and i.i.d. split normal $(\mu, \sigma_1 = 1, \sigma_2 = 2)$. Simulation results for the regular posterior and for the censored posterior with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%). Standard deviations in parentheses. MSEs in brackets, with best MSE in boldface.

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. Whereas we propose to collect in θ_1 the other parameters, such as the coefficients determining the dynamic behavior 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 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 Metropolis-Hastings (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 independence chain Metropolis-Hastings algorithm. To perform the conditional sampling in step 3.2. we use the fact that the conditional distribution corresponding to a joint mixture of Student's t distributions is itself a mixture of Student's t distributions, which we derive in Appendix B. This implies that if we have obtained $q_{mit}(\theta_1, \theta_2)$, a mixture of Student't 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 independence chain Metropolis-Hastings (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 Partially Censored Posterior: *PCP-QERMit* importance sampling method for variance reduction of Value-at-Risk and Expected Shortfall estimators

Putting much effort in obtaining more accurate estimates of risk measures such as Value-at-Risk and Expected Shortfall, using the specific left-tail focus of the Partially Censored Posterior, 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) should be composed of two equally weighted components, one for the high-loss scenarios (corresponding to the tail) and one for remaining realisations of returns. That 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 future innovations from the model and combining these with the posterior draws of model parameters 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 density 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 $p^{cp}(\theta_2|\theta_1, y)$.

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 predicted 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 $\tilde{\varepsilon}^{(i)}_{T+1:T+H}|\theta^{(i)}_1, \theta^{(i)}_2$, $i = 1, \ldots, M$, from its conditional importance density (aimed at high losses) $q_{cmit}(\varepsilon_{T+1:T+H}|\theta^{(i)}_1, \theta^{(i)}_2)$, 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 $(\theta^{(i)}_1, \theta^{(i)}_2, \tilde{\varepsilon}^{(i)}_{T+1:T+H})$, $i = 1, \ldots, M$, as

$$w^{(i)} = \frac{p(\tilde{\varepsilon}^{(i)}_{T+1:T+H}|\theta^{(i)}_1, \theta^{(i)}_2)}{q(\tilde{\varepsilon}^{(i)}_{T+1:T+H}|\theta^{(i)}_1, \theta^{(i)}_2)},$$

where the numerator $p(\varepsilon^{(i)}_{T+1:T+H}|\theta^{(i)}_1, \theta^{(i)}_2)$ 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 $y^{(i)}_{T+1:T+H}$ corresponding to the joint draws $(\theta^{(i)}_1, \theta^{(i)}_2, \tilde{\varepsilon}^{(i)}_{T+1:T+H})$, $i = 1, \ldots, M$, 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

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

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

For the Expected Shortfall 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 fully aim at the high losses – and not at the 50%-50% division between "high-loss" draws and other draws in the regular QERMit method for Bayesian VaR/ES prediction – since we do not use the *scaled* IS weights $w^{(i)}/\sum_{j=1}^{M} w^{(j)}$ that are common in Bayesian IS estimation. 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 matter for "high-loss" draws with indicator $\mathbb{I}\left(\sum_{t=T+1}^{T+H} y^{(i)}_t < C\right) = 1$.

Illustration To illustrate the benefits of the PCP-QERMit method we consider a simple example involving the AR(1) model. Building upon the toy example from subsection 2.1.1, we consider the true DGP of the form

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

with split normally distributed innovations $\varepsilon_t \sim \mathcal{SN}(\delta, \sigma_1^2, \sigma_2^2)$ with $\delta = \frac{\sigma_2 - \sigma_1}{\sqrt{2\pi}}$ so that $E(\varepsilon_t) = 0$. We simulate $T = 1000$ observations from the model with $\mu = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$ and $\rho = 0.8$.

We estimate the AR(1) model with normally distributed innovations $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The priors for μ and σ are the same as in the i.i.d. case, while for ρ we adopt a uniform prior over the stationarity region.

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,

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 2: Estimated 1-step-ahead 99.5%, 99% and 95% VaR and ES (and numerical standard error from 50 MC replications within brackets) for estimated AR(1) model with normally distributed innovations, for $T = 1000$ observations from DGP of AR(1) model with split normally distributed innovations ($\sigma_1 = 1, \sigma_2 = 2$). 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.)

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 $10,000$ draws. (The PCP has a time-constant threshold C_t given by the 10% quantile of the in-sample data.) Table 2 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 Comparison of the Partially Censored Posterior with the regular posterior and the (fully) Censored Posterior: AR(1) model for simulated data

In this subsection we compare the quality of the left-tail density forecasts from the Partially Censored Posterior (PCP) with the regular posterior and the (fully) Censored Posterior (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.

We keep $\mu = 0$, $\rho = 0.8$ and $\sigma_1 = 1$ in the DGP. We do vary the level of misspecification by considering the correctly specified case of $\sigma_2 = 1$ and the misspecified cases of $\sigma_2 = 1.5$ and $\sigma_2 = 2$. Further, we analyze 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 score function of Diks et al. (2011) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns), given by

$$\begin{aligned} S^{cls}(p(y_{T+1}|y_{1:T})) &= \mathbb{I}(y_{T+1} < C_{T+1}) \log p(y_{T+1}|y_{1:T}) \\ &= +\mathbb{I}(y_{T+1} \geq C_{T+1}) \log \left(\int_{C_{T+1}}^{\infty} p(s|y_{1:T}) ds \right). \end{aligned} \tag{2.6}$$

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 a N(0,1) distribution under the null).

Tables 3, 4 and 5 show the results. We observe the following findings. First, as expected, in the case without misspecification ($\sigma_2 = 1$), the regular posterior performs better than the PCP or CP. In this case it is obviously optimal to use all (uncensored) observations. 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.

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
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*** **

Table 3: Left-tail density forecast comparison based on the censored likelihood score function (2.6) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns) *between the regular posterior and the partially censored posterior (PCP)*. This table shows the average of 20 Diebold-Mariano test statistics (with Newey-West standard errors) for 20 simulated data sets. The loss differential (computed for 1000 out-of-sample observations for each simulated dataset) is the difference in the censored likelihood score function (2.6). Positive values indicate superior left-tail forecast performance of the regular Bayesian approach; negative values indicate superior left-tail forecast performance of the PCP. 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).

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
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***

Table 4: Left-tail density forecast comparison based on the censored likelihood score function (2.6) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns) *between the regular posterior and the (fully) censored posterior (CP)*. This table shows the average of 20 Diebold-Mariano test statistics (with Newey-West standard errors) for 20 simulated data sets. The loss differential (computed for 1000 out-of-sample observations for each simulated dataset) is the difference in the censored likelihood score function (2.6). Positive values indicate superior left-tail forecast performance of the regular Bayesian approach; negative values indicate superior left-tail forecast performance of the CP. The significance (in a two-sided test) is indicated by * for $p \leq 0.1$, ** for $p \leq 0.05$ and *** for $p \leq 0.01$.

T	$\sigma_2 = 1$	$\sigma_2 = 1.5$	$\sigma_2 = 2$
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

Table 5: Left-tail density forecast comparison based on the censored likelihood score function (2.6) (with time-constant threshold $C_t = C$ given by the 5% quantile of the returns) *between the (fully) censored posterior and the partially censored posterior (PCP)*. This table shows the average of 20 Diebold-Mariano test statistics (with Newey-West standard errors) for 20 simulated data sets. The loss differential (computed for 1000 out-of-sample observations for each simulated dataset) is the difference in the censored likelihood score function (2.6). Positive values indicate superior left-tail forecast performance of the CP; negative values indicate superior left-tail forecast performance of the PCP. 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).

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. (2014) 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 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, (2.6) can be computed with time-varying C_{T+1}. In our empirical study in Section 4 we consider, next to the standard time-constant threshold for the CSL rule (the 10% percentile of the in-sample data), a time-varying threshold given by the 10% percentile of the MLE-implied conditional distribution.

3.1 Comparison of the Partially Censored Posterior with the regular posterior and the (fully) Censored Posterior: GARCH(1,1) and AGARCH(1,1) models for simulated data

Our aim in this subsection is threefold. First, we investigate the role of the exact model specification on the usefulness of the PCP. There exists an immense amount of models of volatility, including an extensive family of GARCH-type models, cf. Bollerslev (2008) but also recently introduced Generalized Autoregressive Score models (GAS) of Creal et al. (2013). Not all model specifications may be expected to equally benefit from censoring. In other words, we consider the robustness of our results for different model specifications, where for the practical usefulness of the PCP its use should not only be beneficial in certain 'convenient' models. That is, preferably one does not need to particularly adapt the model specification in order to make the PCP useful.

Second, we check what gains can be obtained from censoring with small and large estimation windows.

Third, we analyse how extreme the tails need to be for censoring-based methods to be beneficial. For instance, the Basel requirement involves the 99% VaR, so the 1% percentile. However, for more conservative risk managers the 99.5% VaR may be of interest, while more "risk-seeking" approaches may consider

the 95% VaR. One may expect that the focus on the left tail during the estimation is particularly useful when one is interested in the deep tail. (On the other hand, if one would be interested in the median of the predictive distribution, then one should obviously not focus on the tail during estimation.)

For illustration, consider the DGP of the following GARCH(1,1) model with split normal errors:

$$\begin{aligned} y_t &= \mu + \sqrt{(\kappa^{-1} h_t)}\varepsilon_t, \\ h_t &= \omega(1-\alpha-\beta) + \alpha y_{t-1}^2 + \beta h_{t-1}, \\ \varepsilon_t &\sim \mathcal{SN}(\delta, \sigma_1^2, \sigma_2^2), \end{aligned}$$

where $\delta = \frac{\sigma_2 - \sigma_1}{\sqrt{2\pi}}$ so that $E(\varepsilon_t) = 0$, and where $\kappa = \frac{1}{2}\left(\left(\sigma_1^2 + \sigma_2^2\right) - \frac{(\sigma_2-\sigma_1)^2}{\pi}\right)$ is the variance of ε_t. We set $\mu = 0$, $\omega = 1$, $\alpha = 0.1$, $\beta = 0.8$, and for ε_t we again choose $\sigma_1 = 1$ and $\sigma_2 = 2$. In a single experiment we simulate $T = 1000$ and $T = 2500$ observations from the DGP, and we carry out 50 MC repetitions of such an experiment.

To answer the question about the role of a 'convenient' model specification, we estimate two (misspecified) models: the standard GARCH(1,1) model with normally distributed innovations and the Asymmetric GARCH(1,1) model (AGARCH(1,1)) of Engle and Ng (1993). The latter is characterised by two mean parameters: an actual mean μ_1 and a parameter μ_2 for defining the squared 'demeaned' lagged return $(y_{t-1} - \mu_2)^2$ in the GARCH equation:

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

The GARCH(1,1) model results from the AGARCH(1,1) model by setting $\mu = \mu_1 = \mu_2$. In the PCP for the GARCH(1,1) model we choose the tail related parameters $\theta_2 = \{\mu, \omega\}$ and other (dynamics related) parameters $\theta_1 = \{\alpha, \beta\}$, whereas for the AGARCH(1,1) model we choose $\theta_2 = \{\mu_1, \omega\}$ and $\theta_1 = \{\mu_2, \alpha, \beta\}$. Note that the AGARCH(1,1) model may seem a 'convenient' counterpart of the GARCH(1,1) model for the PCP, as it separates the 'direct' effect of μ on the conditional distribution of y_t (as the mean) and the effect of μ on the dynamics of the GARCH process in two different parameters μ_1 and μ_2.

For both models we take flat priors over the standard domains to impose stationarity and positivity of the volatility process. We analyse 99.5%, 99% and 95% one-step-ahead VaR and ES forecasting over a horizon of 100 days and we carry out 50 independent MC replications. For the estimation we consider, next to the regular posterior, two types of thresholds (for both the CP and PCP): one time-constant threshold (at the 10% data percentile) and the time-varying MLE-based threshold (at the 10% percentile of the estimated conditional distribution).

Tables 6 and 7 show the mean MSEs (i.e., the average of 50 MSEs for the 50 simulated datasets) for the 100 one-step-ahead VaR and ES predictions from the GARCH(1,1) and AGARCH(1,1) models, respectively. We observe the following findings. First, the PCP and CP outperform the regular posterior for the 99.5% and 99% VaR and ES. On the other hand, the regular posterior outperforms the PCP and CP for the 95% VaR and ES, where the 5% quantile is apparently not 'deep enough' in the tail to make the left-tail focus of censoring beneficial. Second, the PCP outperforms the CP for the small estimation window of $T = 1000$ observations (where it is apparently crucial to limit the loss of information due to censoring), whereas the performance of PCP and CP is similar for the large estimation window of $T = 2500$ observations. Notice that it obviously depends on the model whether an estimation window is 'small', where a GARCH(1,1) model requires many more observations than an AR(1) model for accurate estimation, and an AGARCH(1,1) model requires somewhat more observations than a GARCH(1,1) model. Third, typically the PCP (and CP) with time-varying thresholds perform slightly better than

their counterparts with time-constant threshold. Fourth, the use of the PCP is equally or less beneficial in the AGARCH(1,1) model than in the GARCH(1,1) model. Hence, we can say that the PCP approach can perform well even when applied to standard models, so that no specific models need to be used to make the PCP beneficial.

4 Empirical application of a GARCH(1,1)-t model to daily IBM logreturns

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 22nd December 2016 (2512 observations, Figure 4.1).

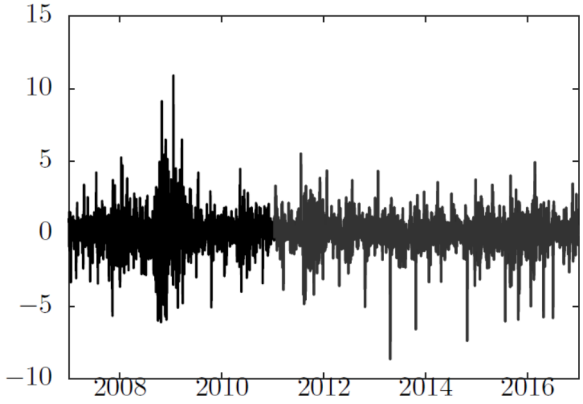


Figure 4.1: The daily logreturns of the IBM stock from the 4th January 2007 to the 22nd December 2016.

In our empirical study we analyse a benchmark model of volatility, commonly employed by practitioners, the Generalized Autoregressive Conditional Heteroscedasticity model (GARCH, Engle, 1982; Bollerslev, 1986) with Student's t innovations. We adopt the following specification

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

and we put flat priors and impose the standard variance positivity and stationarity restrictions (i.e. $\omega > 0$, $\alpha \in (0,1)$, $\beta \in (0,1)$ with $\alpha + \beta < 1$), except for the degrees of freedom, where we use an uninformative yet proper exponential prior (with prior mean 100) for $\nu - 2$.

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 independence chain Metropolis-Hastings 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 $\theta_1 = \{\alpha, \beta\}$ of the dynamics parameters, we conditionally sample $\theta_2 = \{\nu, \mu, \omega\}$ from the conditional censored posterior. For both the CP and PCP we consider two thresholds, a time-constant threshold at the 10% quantile of the in-sample data and a time-varying threshold, the 10% quantile of the MLE-implied conditional distribution.

Table 8 reports the estimation results and Figure 4.2 presents the corresponding kernel density estimates. For the CP leaving α and β to be estimated based on effectively few observations leads to much higher variances. Interestingly, for ν the CP and PCP lead to very different estimation results than the regular

Risk measure	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
			$T = 1000$		
99.5% VaR	0.1556	0.1188	0.1061	0.1156	**0.1040**
99.5% ES	0.2385	0.1531	0.1431	0.1470	**0.1385**
99% VaR	0.1056	0.0953	0.0835	0.0930	**0.0820**
99% ES	0.1784	0.1259	0.1146	0.1219	**0.1118**
95% VaR	**0.0225**	0.0565	0.0491	0.0554	0.0467
95% ES	**0.0645**	0.0770	0.0674	0.0753	0.0655
			$T = 2500$		
99.5% VaR	0.1572	0.0696	**0.0804**	0.0712	0.0807
99.5% ES	0.2447	0.0831	**0.1005**	0.0840	0.1006
99% VaR	0.1056	0.0589	0.0669	0.0602	**0.0659**
99% ES	0.1815	0.0711	0.0843	0.0724	**0.0841**
95% VaR	**0.0194**	0.0411	0.0424	0.0417	0.0420
95% ES	**0.0627**	0.0495	0.0548	0.0508	0.0546

Table 6: Simulation results in estimated (misspecified) GARCH(1,1) model with normally distributed innovations for data from DGP of GARCH(1,1) model with split normally distributed innovations (with $\sigma_1 = 1$ and $\sigma_2 = 2$). The table reports the averages of MSEs (over 50 MC replications) for one-step-ahead VaR and ES prediction over an out-of-sample window of $H = 100$ for standard posterior, censored posterior (CP) and partially censored posterior (PCP) — the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. Bold numbers indicate the lowest average MSE.

Risk measure	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
			$T = 1000$		
99.5% VaR	0.1369	0.1374	0.1238	0.1372	**0.1201**
99.5% ES	0.2110	0.1802	0.1738	0.1793	**0.1654**
99% VaR	0.0924	0.1079	0.0952	0.1078	**0.0923**
99% ES	0.1572	0.1465	0.1361	0.1458	**0.1305**
95% VaR	**0.0188**	0.0593	0.0546	0.0583	0.0505
95% ES	**0.0551**	0.0857	0.0765	0.0853	0.0732
			$T = 2500$		
99.5% VaR	0.1627	0.0772	0.0901	**0.0764**	0.0893
99.5% ES	0.2509	0.0944	0.1160	**0.0937**	0.1140
99% VaR	0.1101	**0.0627**	0.0731	0.0632	0.0728
99% ES	0.1871	0.0791	0.0957	**0.0788**	0.0944
95% VaR	**0.0212**	0.0377	0.0437	0.0388	0.0434
95% ES	0.0660	**0.0503**	0.0592	0.0512	0.0591

Table 7: Simulation results in estimated (misspecified) AGARCH(1,1) model with normally distributed innovations for data from DGP of GARCH(1,1) model with split normally distributed innovations (with $\sigma_1 = 1$ and $\sigma_2 = 2$). The table reports the averages of MSEs (over 50 MC replications) for one-step-ahead VaR and ES prediction over an out-of-sample window of $H = 100$ for standard posterior, censored posterior (CP) and partially censored posterior (PCP) — the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. Bold numbers indicate the lowest average MSE.

Parameter	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
ν	7.0943	45.6427	38.1495	43.2466	33.7796
	(1.4748)	(25.3582)	(26.3139)	(25.3678)	(24.8307)
μ	0.0905	0.6684	0.6287	0.5821	0.4492
	(0.0362)	(0.2725)	(0.3825)	(0.2310)	(0.3131)
ω	16.4548	260.5379	47.3642	288.6082	42.7302
	(18.9873)	(369.5474)	(60.8834)	(423.7474)	(59.0946)
α	0.1260	0.1605	0.1264	0.1683	0.1264
	(0.0271)	(0.0515)	(0.0273)	(0.0562)	(0.0273)
β	0.8652	0.8317	0.8650	0.8234	0.8650
	(0.0280)	(0.0530)	(0.0281)	(0.0572)	(0.0281)

Table 8: Empirical application to daily IBM logreturns: estimation results (means and standard deviations) for the GARCH(1,1)-t model estimated with the regular posterior, the censored posterior (CP) and the partially censored posterior (PCP) — the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively.

posterior. The latter implies a very fat-tailed distribution with a low degrees of freedom parameter, while both the CP and the PCP suggest an almost normal shape of the left tail of the distribution of the innovations (which comes with a high value of ω, suggesting that the left tail may be more like a normal distribution with a higher variance than like a Student's t distribution with a smaller variance). This huge discrepancy between the results from the regular posterior and the (P)CP can be interpreted as evidence of model misspecification.

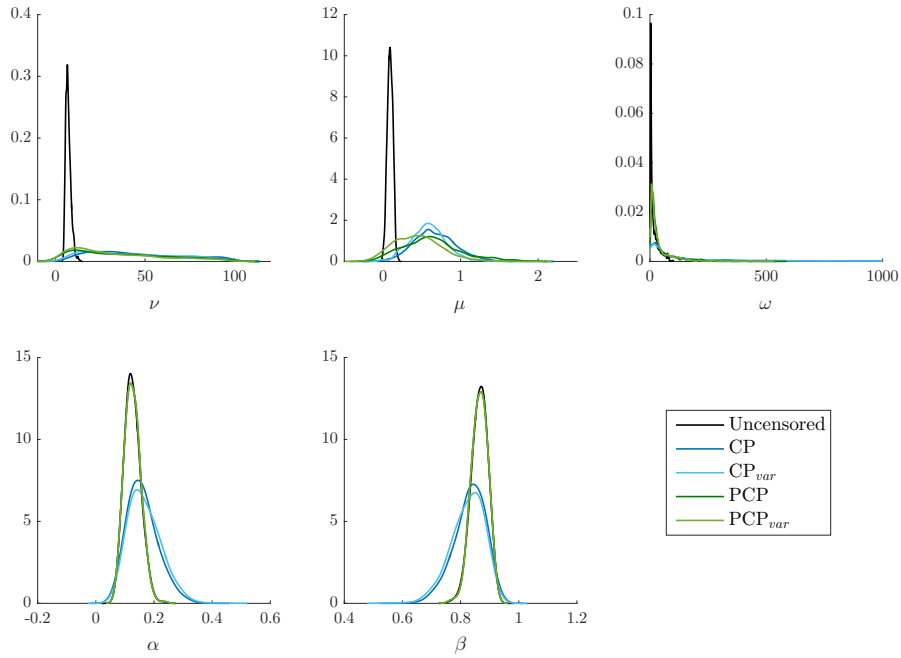


Figure 4.2: Empirical application to daily IBM logreturns: kernel density estimates for the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively.

In our forecasting study we consider $H = 1500$ 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 scoring rule (2.6) 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

	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
Threshold for censored likelihood scoring rule = 0.5% percentile					
Posterior	–	–	–	–	–
CP (const. C)	-2.5013**	–	–	–	–
PCP (const. C)	2.0464**	2.5867***	–	–	–
CP (var. C_t)	-2.9143***	-2.0998**	-2.5866***	–	–
PCP (var. C_t)	2.0369**	2.6460***	-1.8986*	2.6533***	–
Threshold for censored likelihood scoring rule = 1% percentile					
Posterior	–	–	–	–	–
CP (const. C)	-3.8343***	–	–	–	–
PCP (const. C)	1.3922	2.5763***	–	–	–
CP (var. C_t)	-4.0150***	-1.7450*	-2.5415**	–	–
PCP (var. C_t)	1.3609	2.7439***	-1.3752	2.7008***	–
Threshold for censored likelihood scoring rule = 5% percentile					
Posterior	–	–	–	–	–
CP (const. C)	-4.9013***	–	–	–	–
PCP (const. C)	-1.8209*	3.5258***	–	–	–
CP (var. C_t)	-5.0946***	0.0735	-3.2159***	–	–
PCP (var. C_t)	-2.2713**	3.8532***	-0.7073	3.5323***	–

Table 9: Empirical application to daily IBM logreturns: Diebold-Mariano test statistics for pairwise method comparison of forecasting performance in the left tail based on the censored likelihood scoring rule with *time-constant* threshold for evaluation (i.e., for computing the censored likelihood scoring rule), for $H = 1500$ out-of-sample observations, between the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. *, **, *** indicate significance at 10%, 5%, 1% level, respectively.

literature, which we set at the 10% percentile of the in-sample data.

Tables 9 and 10 present the results of the Diebold-Mariano test based on the censored scoring rule with time-constant and time-varying threshold, respectively. A positive number indicates that the corresponding row method provides better left-tail density forecasts than the corresponding column method. The plots of the loss differentials used in the Diebold-Mariano tests, provided in Figure C.2 in Appendix C, show that the PCP provides substantially better left-tail density predictions than the CP and the regular posterior on multiple days, whereas it is never (or hardly ever) substantially outperformed.

We can see that the censored likelihood scoring rule with time-constant threshold for evaluation clearly prefers the PCP over the CP: for all the quantile levels considered the PCP significantly outperforms the fully censored approach (at 1% significance level). Moreover, the PCP performs significantly better for the extreme left tail than the regular posterior (at 5% significance level). In this application full censoring is only harmful compared to the regular posterior, which stresses the merit of the introduced *partial* censoring.

Also the conclusions from the evaluations based on the time-varying threshold are supportive for the PCP approach. For the extreme left tail the regular posterior is significantly outperformed by PCP (at 5% significance level).

We can conclude that for more complex models, usually applied in empirical practice, the role of *partial* censoring becomes crucial. With multiple parameters to be estimated based on effectively few observations, it might be hard for the fully censored posterior to provide accurate left-tail density forecasts, so that it may be more beneficial to use the regular posterior. On the other hand, with an 'appropriately' chosen subset of parameters to apply censoring, we can achieve better left-tail density forecasts than with the standard posterior.

	Posterior	CP (const. C)	PCP (const. C)	CP (var. C_t)	PCP (var. C_t)
Threshold for censored likelihood scoring rule = 0.5% percentile					
Posterior	–	–	–	–	–
CP (const. C)	0.9642	–	–	–	–
PCP (const. C)	2.1526**	2.2572**	–	–	–
CP (var. C_t)	0.9848	-0.2301	-2.1132**	–	–
PCP (var. C_t)	2.3944**	2.4226**	-0.3568	2.3016**	–
Threshold for censored likelihood scoring rule = 1% percentile					
Posterior	–	–	–	–	–
CP (const. C)	-0.1137	–	–	–	–
PCP (const. C)	1.3287	2.8183***	–	–	–
CP (var. C_t)	-0.0889	0.4013	-2.5591**	–	–
PCP (var. C_t)	1.5510	3.2282***	0.3566	2.9846***	–
Threshold for censored likelihood scoring rule = 5% percentile					
Posterior	–	–	–	–	–
CP (const. C)	0.6637	–	–	–	–
PCP (const. C)	2.2503**	3.5778***	–	–	–
CP (var. C_t)	0.5552	-2.2570**	-3.6174***	–	–
PCP (var. C_t)	2.1086**	3.1398***	-2.5594**	3.2996***	–

Table 10: Empirical application to daily IBM logreturns: Diebold-Mariano test statistics for pairwise method comparison of forecasting performance in the left tail based on the censored likelihood scoring rule with *time-varying* threshold for evaluation (i.e., for computing the censored likelihood scoring rule), for $H = 1500$ out-of-sample observations, between the regular posterior, censored posterior (CP) and partially censored posterior (PCP) – the latter two with time-constant threshold (const. C) and time-varying threshold (var. C_t), at the 10% percentile of the empirical distribution and the 10% percentile of the MLE-implied conditional distribution, respectively. *, **, *** indicate significance at 10%, 5%, 1% level, respectively.

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. (2014) 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 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. Also extensions of the classical approach of Opschoor et al. (2016) based on so-called pooling are relevant in this regard. The Bayesian approach of Aastveit et al. (2018a) can be used in this context. 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. Finally, an interesting extension would be the analysis of credit risk and defaults.

Bibliography

Aastveit, K. A., L. F. Hoogerheide, J. Mitchell, and H.K. van Dijk (2018a), "Structure and workings of density forecast combinations in economics", Unpublished manuscript.

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and Herman K. van Dijk (2018b), "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 and 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.

Bollerslev, T. (1986), "Generalised Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 51, 307–327.

Bollerslev, T. (2008), "Glossary to ARCH (GARCH)." Technical Report 2008-49, CREATES Research Paper.

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 and Economic Statistic*, 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 (2014), "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.

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.

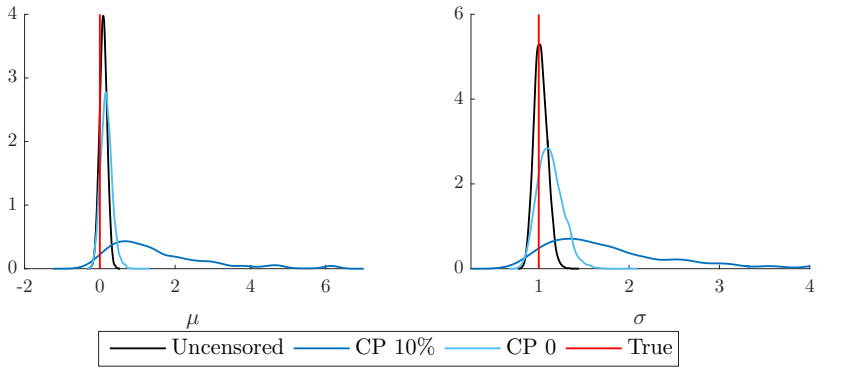
Opschoor, A., D. van Dijk, and M. van der Wel (2016), "Combining Density Forecasts using Focused Scoring Rules." *Tinbergen Institute Discussion Paper*, 14-090/III.

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.

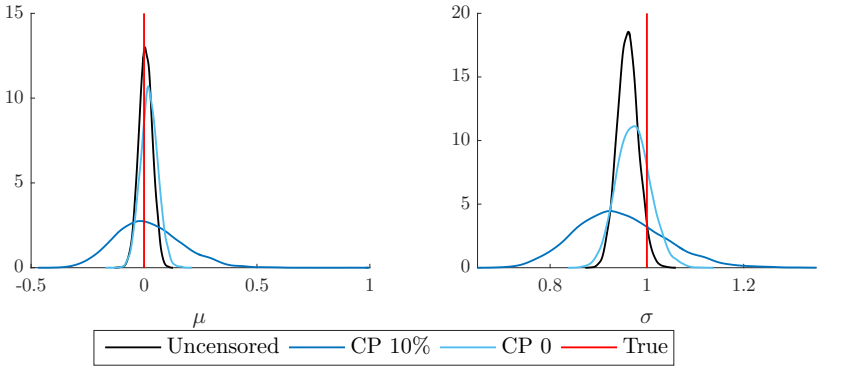
Roth, M. (2013), "On the Multivariate t Distribution." Technical Report LiTH-ISY-R-3059, Automatic Control Group at Linköpings Universitet.

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

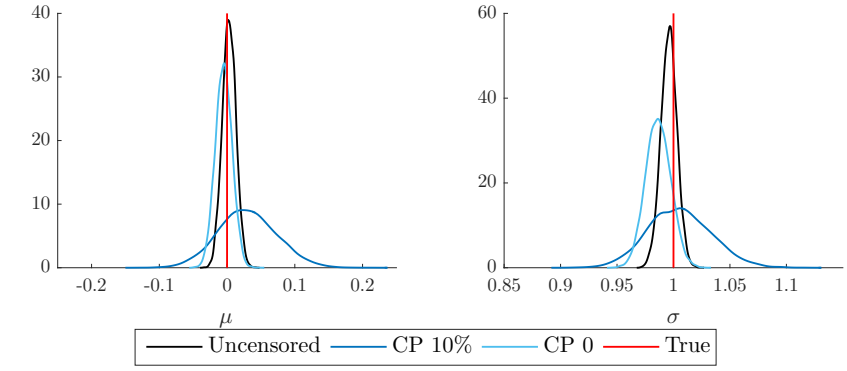
A Estimated posterior densities and censored posterior densities in estimated i.i.d. normal model



(a) $T = 100$

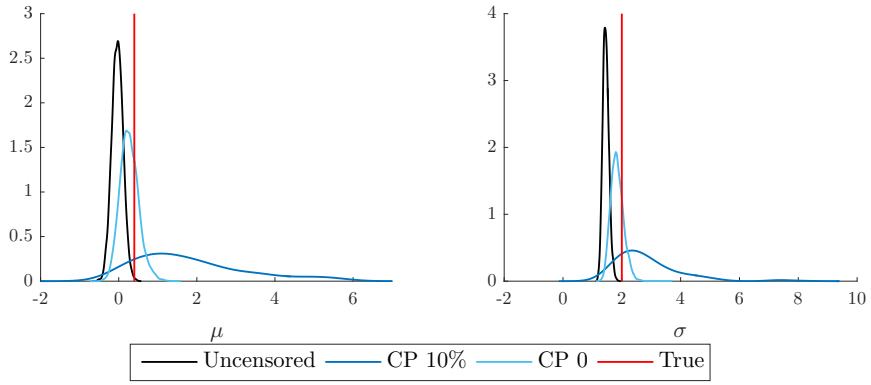


(b) $T = 1000$

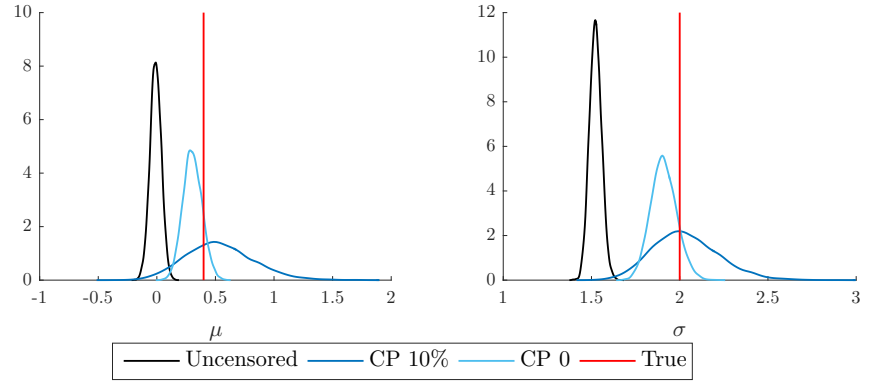


(c) $T = 10000$

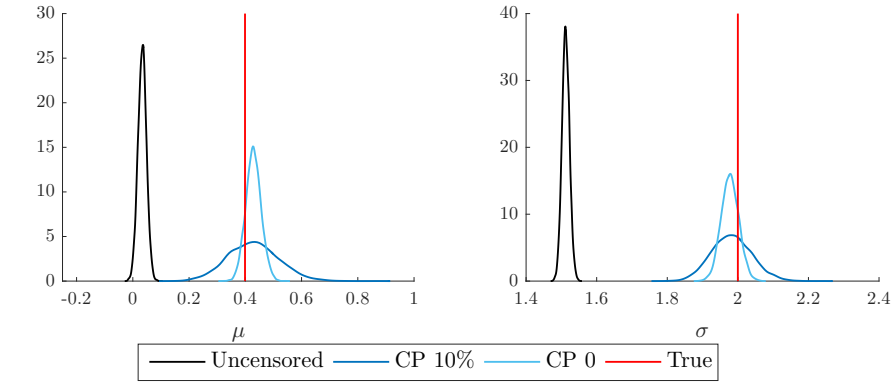
Figure A.1: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. normal ($\sigma = 1$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).



(a) $T = 100$



(b) $T = 1000$



(c) $T = 10000$

Figure A.2: Estimation results in i.i.d. normal $N(\mu, \sigma^2)$ model for $T = 100, 1000, 10000$ observations from DGP of i.i.d. split normal ($\sigma_1 = 1$, $\sigma_2 = 2$). Kernel density estimates of regular posterior and censored posterior (CP) with two different thresholds, at 0 (CP0) and at the 10% data percentile (CP10%) together with the true parameter values (corresponding to left tail).

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 Students' 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

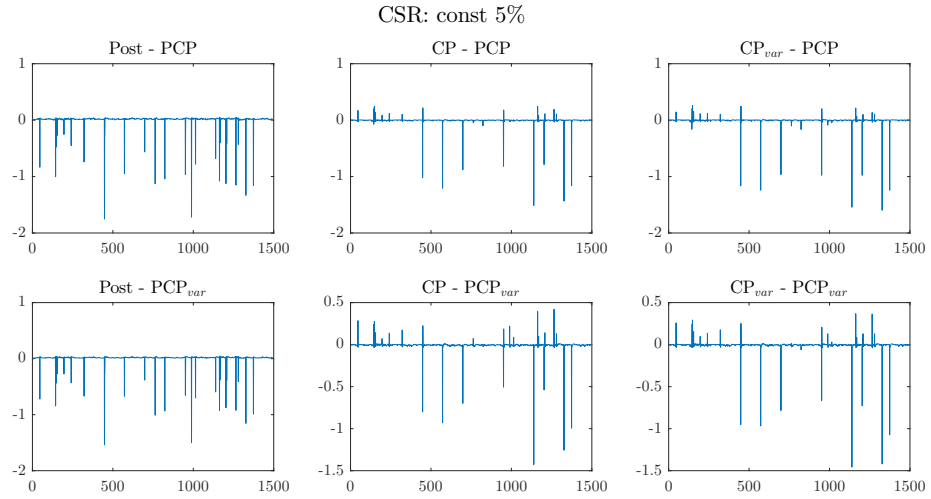
$$\begin{aligned}
\mu_{h,2|1} &= \mu_{h,2} + \Sigma_{h,21}\Sigma_{h,11}^{-1}(x_1-\mu_{h,1}),\\
\Sigma_{h,2|1} &= \frac{\nu_h + (x_1-\mu_{h,1})'\Sigma_{h,11}^{-1}(x_1-\mu_{h,1})}{\nu_h+d_1}\left(\Sigma_{h,22}-\Sigma_{h,21}\Sigma_{h,h,11}^{-1}\Sigma_{h,12}\right),\\
\nu_{h,2|1} &= \nu_h + d_1,
\end{aligned}$$

and with adjusted component probabilities

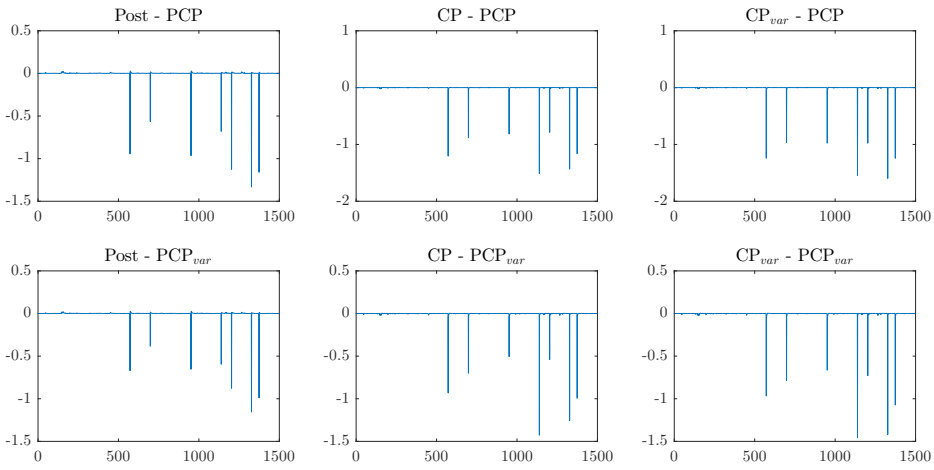
$$\eta_{h,2|1} = \mathbb{P}[z=e_h|x] = \frac{\eta_h t(x_1;\mu_{h,1},\Sigma_{h,11},\nu_h)}{\sum_{m=1}^{H}\eta_m t(x_1;\mu_{m,1},\Sigma_{m,11},\nu_m)}.$$

This implies that if we have obtained $q_{mit}(\theta_1,\theta_2)$, a mixture of Student't 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 our proposed *Conditional MitISEM* method.

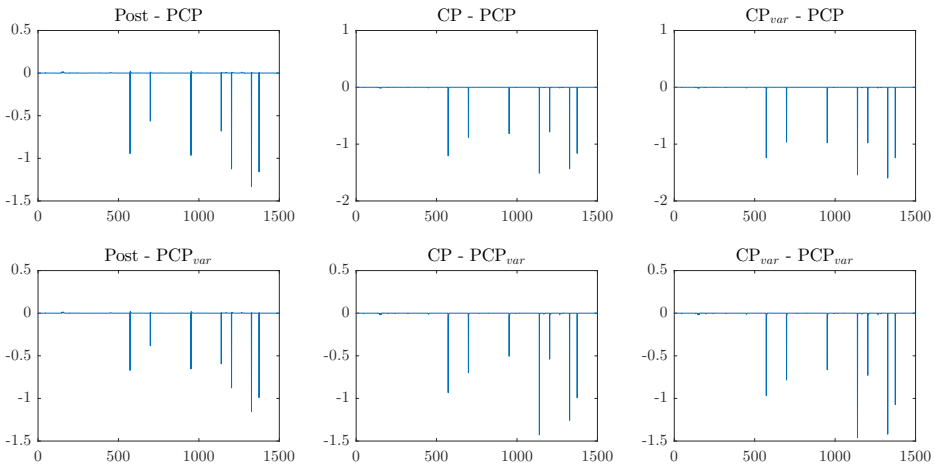
C Loss differential plots



(a) Threshold for censored likelihood scoring rule = 5% percentile.

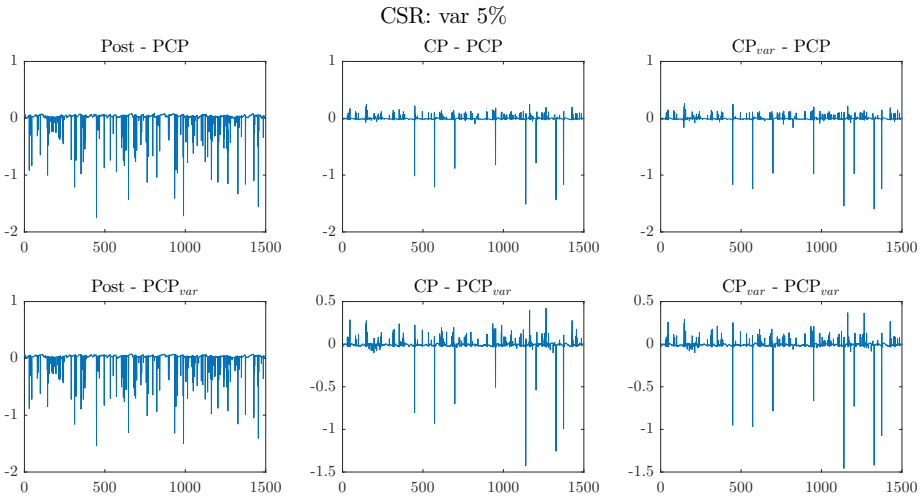


(b) Threshold for censored likelihood scoring rule = 1% percentile.

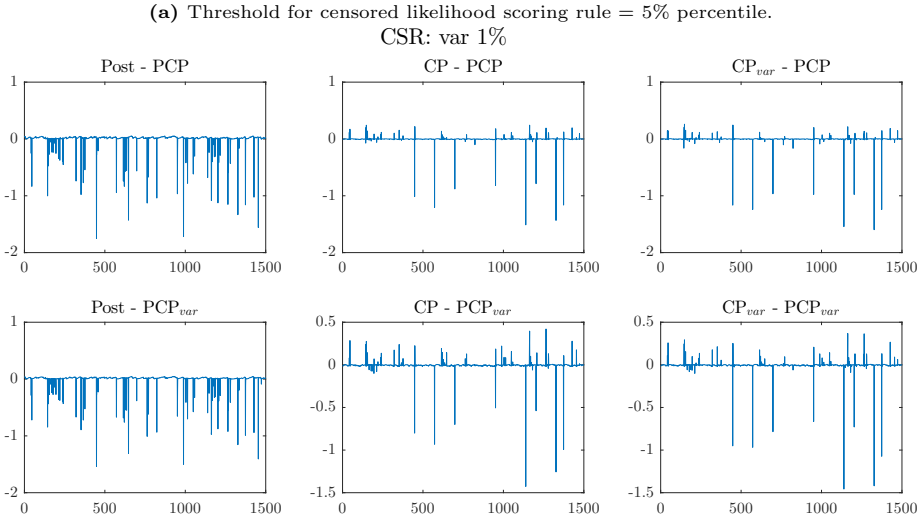


(c) Threshold for censored likelihood scoring rule = 0.5% percentile.

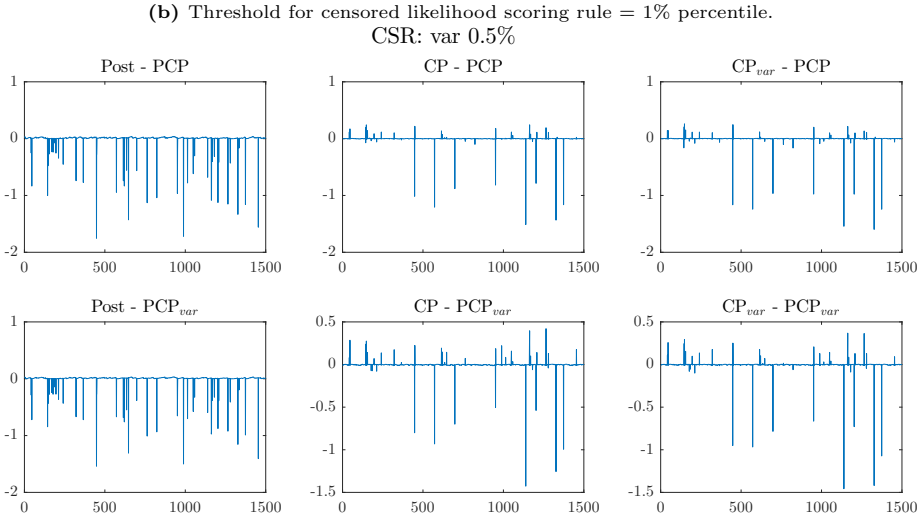
Figure C.1: Empirical application to daily IBM logreturns: loss differentials based on the censored likelihood scoring rule with *time-constant* evaluation threshold (computed as the percentile of the empirical distribution). Negative values of the loss differential indicate that the Partially Censored Posterior (PCP) performs better than the alternative.



(a) Threshold for censored likelihood scoring rule = 5% percentile.



(b) Threshold for censored likelihood scoring rule = 1% percentile.



(c) Threshold for censored likelihood scoring rule = 0.5% percentile.

Figure C.2: Empirical application to daily IBM logreturns: loss differentials based on the censored likelihood scoring rule with *time-varying* evaluation threshold (computed as the percentile of the estimated conditional distribution based on the ML estimator). Negative values of the loss differential indicate that the Partially Censored Posterior (PCP) performs better than the alternative.