Probabilistic Forecasting and Comparative Model Assessment Based on Markov Chain Monte Carlo Output

Fabian Krüger[a] Sebastian Lerch[b,c] Thordis Thorarinsdottir[d]
Tilmann Gneiting[b,c]

[a]Heidelberg University
[b]Heidelberg Institute for Theoretical Studies
[c]Karlsruhe Institute of Technology
[d]Norwegian Computing Center

March 30, 2017

Abstract

In Bayesian inference, predictive distributions are typically available only through a sample generated via Markov chain Monte Carlo (MCMC) or related algorithms. In this paper, we conduct a systematic analysis of how to make and evaluate probabilistic forecasts from such simulation output. Based on proper scoring rules, we develop a notion of consistency that allows to assess the adequacy of methods for estimating the stationary distribution underlying the simulation output. We then provide asymptotic results that account for the salient features of Bayesian posterior simulators, and derive conditions under which choices from the literature satisfy our notion of consistency. Importantly, these conditions depend on the scoring rule being used, such that the choices of approximation method and scoring rule are intertwined. While the logarithmic rule requires fairly stringent conditions, the continuous ranked probability score (CRPS) yields consistent approximations under minimal assumptions. These results are illustrated in a simulation study and an economic data example. Overall, we find that mixture-of-parameters approximations which exploit the parametric structure of Bayesian models perform particularly well.

1 Introduction

Probabilistic forecasts take the form of predictive probability distributions over quantities or events of interest (Gneiting and Katzfuss, 2014). In this context, a rapidly growing transdisciplinary literature uses Bayesian inference to produce posterior predictive distributions in a wide range of applications, including economic, financial, ecological, and meteorological problems, among many others. Prompted by these developments, as recently noted by Broms et al. (2016, p. 1759),

> "a quiet regime shift is occurring in Bayesian statistics where predictive model comparison approaches are experiencing a resurgence."

The shift of emphasis has been particularly prominent in economic applications, as evidenced by the dominance of econometrically oriented references in Table 1.

In Bayesian statistics, posterior predictive distributions arise as mixture distributions with respect to the posterior distribution of the parameter vector. For a real-valued continuous quantity of interest, the posterior predictive distribution, F_0, can be represented by its cumulative distribution function (CDF) or the respective density. The posterior predictive CDF is then of the generic form

$$F_0(x) = \int_\Theta F_c(x|\theta)\,\mathrm{d}P_{\text{post}}(\theta) \tag{1}$$

for $x \in \mathbb{R}$, where P_{post} is the posterior distribution of the parameter, θ, over some parameter space, Θ, and $F_c(\cdot|\theta)$ is the conditional predictive CDF when $\theta \in \Theta$ is the true parameter. Harris (1989) argues that predictive distributions of this form have appeal in frequentist settings as well. Frequently, the integral in (1) does not admit a solution in closed form, and so the posterior predictive CDF must be approximated or estimated in some way, typically using some form of Markov chain Monte Carlo (MCMC); see, e.g., Gelfand and Smith (1990) and Gilks et al. (1996).

Given a simulated sequence $(\theta_i)_{i=1}^m$ of parameter values from P_{post}, one approach, which we call the *mixture-of-parameters* (MP) technique, is to approximate F_0 by

$$\hat{F}_m^{\text{MP}}(x) = \frac{1}{m}\sum_{i=1}^m F_c(x|\theta_i). \tag{2}$$

However, this method can be used only when the conditional distributions $F_c(\cdot|\theta)$ are available in closed form. An alternative route is to simulate a sequence $(X_i)_{i=1}^m$ where $X_i \sim F_c(\cdot|\theta_i)$, and to approximate F_0 based on this sample, using either nonparametric or parametric techniques. The most straightforward option is to estimate F_0 by the *empirical CDF* (ECDF),

$$\hat{F}_m^{\text{ECDF}}(x) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{x \geq X_i\}. \tag{3}$$

Alternatively, one might employ a *kernel density* (KD) estimate,

$$\hat{f}_m^{\text{KD}}(x) = \frac{1}{mh_m}\sum_{i=1}^m K\left(\frac{x - X_i}{h_m}\right), \tag{4}$$

of the posterior predictive density, where K is a kernel function, i.e., a symmetric, bounded, and square-integrable probability density, such as the Gaussian or the Epanechnikov kernel, and h_m is a suitable bandwidth (Rosenblatt, 1956; Silverman, 1986). Finally, much extant work employs a *Gaussian approximation* (GA) to F_0, namely

$$\hat{F}_m^{\text{GA}}(x) = \Phi\left(\frac{x - \hat{\mu}_m}{\hat{\sigma}_m}\right), \tag{5}$$

where Φ is the CDF of the standard normal distribution, and $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the empirical mean and standard deviation of the sample $(X_i)_{i=1}^m$.

Following Rubin (1984) and Little (2006), it is now widely accepted that posterior predictive inference should be evaluated using frequentist principles, without prior information entering the model evaluation stage. For the comparison and ranking of probabilistic forecasting methods one typically uses a proper scoring rule (Gneiting and Raftery, 2007) that assigns a numerical score or penalty based on the predictive CDF, F, or its density, f, and the corresponding realization, y, such as the logarithmic score (LogS; Good, 1952),

$$\text{LogS}(F, y) = -\log f(y), \tag{6}$$

or the continuous ranked probability score (CRPS; Matheson and Winkler, 1976),

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{z \geq y\}\right)^2 \mathrm{d}z. \tag{7}$$

In practice, one finds and compares the mean score over an out-of-sample test set, and the forecasting method with the smaller mean score is preferred. Formal tests of the null hypothesis of equal predictive performance can be employed as well (Diebold and Mariano, 1995; Giacomini and White, 2006; Clark and McCracken, 2013).

Table 1 summarizes the use of evaluation techniques in recently published comparative studies of probabilistic forecasting methods that use Bayesian inference via MCMC. The listing highlights that the choices of approximation method and scoring rule are heavily intertwined. For example, the mixture-of-parameters technique has mainly been applied in concert with the logarithmic score, whereas the empirical CDF method can be used in conjunction with the CRPS only. However, to this date, there are few, if any, guidelines to support choices in the table, and it is not clear how they affect practical model comparisons. The present paper provides a systematic analysis of this topic. We focus on the following questions. First, what defines reasonable choices of the approximation method and scoring rule? Second, under what conditions do extant choices from the literature satisfy this definition? Third, for a given scoring rule, how accurate are alternative approximation methods in practically relevant scenarios?

To answer the first question, we propose the notion of a consistent approximation method. This formalizes the idea that, as the size of the simulated sample becomes larger and larger, and in terms of a given scoring rule, the approximation ought to perform as well as the unknown true forecast distribution. To tackle the second question, we evaluate methodological choices such as those in Table 1. The results account for the salient features of Bayesian posterior simulators, in particular time series type dependence among the MCMC draws. Regarding the third question, we provide a simulation study that is motivated by practical applications of MCMC. Furthermore, we consider the behavior of various approximation methods and scoring rules in an economic data example. Overall, our findings support the use of the mixture-of-parameters estimator in order to approximate the posterior predictive distribution of interest.

We emphasize that the present study – and the use of scoring rules in general – concern the *comparative* assessment of two or more predictive models: The model with the smallest mean score is considered the most appropriate. Comparative assessment is essential in order to choose among a large number of specifications typically available in practice. This task is different from *absolute* assessment, which amounts to diagnosing possible misspecification, using the probability integral transform (e.g. Diebold et al., 1998; Gneiting et al., 2007; Held et al., 2010), posterior predictive checks (Gelman et al., 1996; Gelman et al., 2014a, Chapter 6) and related methods.

Table 1: Approximation methods and scoring rules in recent studies using probabilistic forecasts based on MCMC output. For details see Appendix A.

Approximation method	Logarithmic score	CRPS
Based on parameter draws		
Mixture-of-parameters	Amisano and Giacomini (2007) Gschlößl and Czado (2007) Lopes et al. (2008) Maheu and Gordon (2008) Liu and Maheu (2009) Maheu and McCurdy (2009) Geweke and Amisano (2010, 2011) Jensen and Maheu (2010, 2013, 2014) Jochmann et al. (2010) Kallache et al. (2010) Li et al. (2010) Delatola and Griffin (2011) Maheu et al. (2012) Maneesoonthorn et al. (2012) Jin and Maheu (2013, 2016) Baştürk et al. (2014) Maheu and Song (2014) Risser and Calder (2015) Zhou et al. (2015) Kastner (2016) Warne et al. (2016)	Kallache et al. (2010) Trombe et al. (2012) Risser and Calder (2015)
Based on a sample		
Empirical CDF		Gschlößl and Czado (2007) Lopes et al. (2008) Panagiotelis and Smith (2008) De la Cruz and Branco (2009) Salazar et al. (2011) Friederichs and Thorarinsdottir (2012) Sigrist et al. (2012) Groen et al. (2013) Leininger et al. (2013) Berrocal et al. (2014) Clark and Ravazzolo (2015) Krüger et al. (2015) Sahu et al. (2015) Sigrist et al. (2015) Smith and Vahey (2016) Berg (2017)
Kernel density estimation	Belmonte et al. (2014) Bauwens et al. (2015) Berg and Henzel (2015) Carriero et al. (2015b,c) Berg (2017)	Krüger and Nolte (2016)
Gaussian approximation	Adolfson et al. (2007) Clark (2011) Carriero et al. (2015a, 2016) Clark and Ravazzolo (2015) Giannone et al. (2015) Warne et al. (2016)	Brandt et al. (2014) Rodrigues et al. (2014)

The remainder of this paper is organized as follows. Section 2 introduces the notion of a consistent approximation method. In Section 3 we provide theoretical justifications of approximation methods encountered in the literature. Sections 4 and 5 present simulation and empirical evidence on the performance of these methods, and Section 6 concludes with a discussion. Technical material is deferred to Appendices, and software is available within the scoringRules package for R.

2 Formal setting

In this section, we discuss the posterior predictive distribution in Bayesian forecasting, give a brief review of proper scoring rules and score divergences, and introduce the concept of a consistent approximation method based on MCMC output.

As discussed earlier, the posterior predictive cumulative distribution function (CDF) of a Bayesian forecasting model is given by

$$F_0(x) = \int_\Theta F_c(x|\theta)\,\mathrm{d}P_{\text{post}}(\theta)$$

where $\theta \in \Theta$ is the parameter, P_{post} is the posterior distribution of the parameter, and $F_c(\cdot|\theta)$ is the predictive distribution *conditional* on a parameter value θ; see, e.g., Greenberg (2013, p. 33) or Gelman et al. (2014a, p. 7). A generic Markov chain Monte Carlo (MCMC) algorithm designed to sample from F_0 can be sketched as follows.

- Fix $\theta_0 \in \Theta$ at some arbitrary value.
- For $i = 1, 2, \ldots$ iterate as follows:
  - Draw $\theta_i \sim \mathcal{K}(\theta_i|\theta_{i-1})$, where $\mathcal{K}$ is a transition kernel that specifies the conditional distribution of θ_i given θ_{i-1}.
  - Draw $X_i \sim F_c(\cdot|\theta_i)$.

We assume throughout that the transition kernel $\mathcal{K}$ is such that the sequence $(\theta_i)_{i=1,2,\ldots}$ is stationary and ergodic in the sense of Geweke (2005, Definition 4.5.5) with invariant distribution P_{post}, as holds widely in practice (Craiu and Rosenthal, 2014). Importantly, stationarity and ergodicity of $(\theta_i)_{i=1,2,\ldots}$ with invariant distribution P_{post} imply that $(X_i)_{i=1,2,\ldots}$ is stationary and ergodic with invariant distribution F_0 (Genon-Catalot et al., 2000, Proposition 3.1).

This generic MCMC algorithm allows for two general options for estimating the posterior predictive distribution F_0 in (1), namely,

- Option A: Based on parameter draws $(\theta_i)_{i=1}^m$,
- Option B: Based on a sample $(X_i)_{i=1}^m$,

where m typically is on the order of a few thousands or ten thousands. Alternatively, some authors, such as Krüger et al. (2015), generate, for each $i = 1, \ldots, m$, independent draws $X_{ij} \sim F_c(\cdot|\theta_i)$, where $j = 1, \ldots, J$. The considerations below apply in this more general setting as well.

2.1 Approximation methods

In the case of Option A, the sequence $(\theta_i)_{i=1}^m$ of parameter draws is used to approximate the posterior predictive distribution, F_0, by the mixture-of-parameters estimator $\hat{F}_m^{\mathrm{MP}}$ in (2). Under the assumption of ergodicity,

$$\hat{F}_m^{\mathrm{MP}}(x) = \frac{1}{m}\sum_{i=1}^{m} F_c(x|\theta_i) \longrightarrow \int_\Theta F_c(x|\theta)\,\mathrm{d}P_{post}(\theta) = F_0(x)$$

for $x \in \mathbb{R}$. This estimator was popularized by Gelfand and Smith (1990, Section 2.2), based on earlier work by Tanner and Wong (1987), and is often called a *conditional* or *Rao-Blackwellized* estimator. For independent samples, the Rao-Blackwell theorem implies that exploiting the conditional distributions reduces the variance in the estimation; see, e.g., Geweke (2005, Section 4.4.1). However, this theoretical motivation does not readily extend to the more realistic case of dependent samples. Therefore, we follow suggestions of Geyer (1995, p. 152) and avoid the term Rao-Blackwellization. Instead, we refer to $\hat{F}_m^{\mathrm{MP}}$ as the *mixture-of-parameters* (MP) estimator.

In the case of Option B, the sample $(X_i)_{i=1}^m$ is employed to approximate the posterior predictive distribution F_0. Various methods for doing this have been proposed and used, including the *empirical CDF* of the sample, denoted $\hat{F}_m^{\mathrm{ECDF}}$ in (3), the *kernel density* estimator $\hat{f}_m^{\mathrm{KD}}$ in (4), and the *Gaussian* approximation $\hat{F}_m^{\mathrm{GA}}$ in (5). Approaches of this type imply 'more randomness than necessary', in that the simulation step to draw $(X_i)_{i=1}^m$ can be avoided if Option A is used. That said, Option A requires full knowledge of the model specification, as the conditional distributions must be known in closed form in order to compute $\hat{F}_m^{\mathrm{MP}}$. There are situations where this is restrictive, e.g., when the task is to predict a complex transformation of the original, possibly vector-valued predictand. The implementation of the approximation methods is straightforward, except for the case of kernel density estimation, in which we discuss implementation choices in Section 3.3.

2.2 Proper scoring rules and score divergences

Let $\Omega \subseteq \mathbb{R}$ denote the set of possible values of the quantity of interest, and let $\mathcal{F}$ denote a convex class of probability distributions on Ω. A *scoring rule* is a function

$$\mathrm{S} : \mathcal{F} \times \Omega \longrightarrow \mathbb{R} \cup \{\infty\}$$

that assigns numerical values to pairs of forecasts $F \in \mathcal{F}$ and observations $y \in \Omega$. We typically set $\Omega = \mathbb{R}$, but will occasionally restrict attention to compact subsets.

Throughout this paper, we define scoring rules to be negatively oriented, i.e., a lower score indicates a better forecast. A scoring rule is *proper* relative to $\mathcal{F}$ if the expected score

$$\mathrm{S}(F, G) = \mathbb{E}_{Y\sim G}\,\mathrm{S}(F, Y)$$

is minimized for $F = G$, i.e.,

$$\mathbb{E}_{Y\sim G}\,\mathrm{S}(G, Y) = \mathrm{S}(G, G) \leq \mathrm{S}(F, G) = \mathbb{E}_{Y\sim G}\,\mathrm{S}(F, Y)$$

for all probability distributions $F, G \in \mathcal{F}$. It is *strictly proper* relative to the class $\mathcal{F}$ if, furthermore, equality implies that $F = G$. The *score divergence* associated with the scoring rule S is given by

$$d_{\mathrm{S}}(F, G) = \mathrm{S}(F, G) - \mathrm{S}(G, G).$$

Table 2: Examples of proper scoring rules, along with the associated score divergence and natural domain, $\mathcal{F}$. For a probability distribution with CDF F, we write μ_F for its mean, σ_F for its standard deviation, and f for its density.

Scoring rule	$\mathrm{S}(F, y)$	$d_{\mathrm{S}}(F, G)$	$\mathcal{F}$
Logarithmic score	$-\log f(y)$	$\int g(z) \log \frac{g(z)}{f(z)} \, \mathrm{d}z$	$\mathcal{L}_1$
Continuous ranked probability score	$\int (F(z) - \mathbb{1}\{z \geq y\})^2 \, \mathrm{d}z$	$\int (F(z) - G(z))^2 \, \mathrm{d}z$	$\mathcal{M}_1$
Dawid–Sebastiani score	$\log \sigma_F^2 + \frac{(y-\mu_F)^2}{\sigma_F^2}$	$\frac{\sigma_G^2}{\sigma_F^2} - \log \frac{\sigma_G^2}{\sigma_F^2} + \frac{(\mu_F - \mu_G)^2}{\sigma_F^2} - 1$	$\mathcal{M}_2$

Clearly, $d_{\mathrm{S}}(F, G) \geq 0$ for all $F, G \in \mathcal{F}$ is equivalent to propriety of the scoring rule S, which is a critically important property in practice (Gneiting and Raftery, 2007).

Table 2 shows frequently used proper scoring rules, along with the associated score divergences and the natural domain. For any given scoring rule S, the associated *natural domain* is the largest convex class of probability distributions F such that $\mathrm{S}(F, y)$ is well-defined and finite almost surely under F. Specifically, the natural domain for the popular *logarithmic score* (LogS; eq. (6)) is the class $\mathcal{L}_1$ of the probability distribution with densities, and the respective score divergence is the Kullback-Leibler divergence; see, e.g., Gneiting and Raftery (2007) and Thorarinsdottir et al. (2013). For the *continuous ranked probability score* (CRPS; eq. (7)), the natural domain is the class $\mathcal{M}_1$ of the probability distributions with finite mean. These two scores are strictly proper relative to their respective natural domains. Finally, the natural domain for the *Dawid–Sebastiani score* (DSS; Dawid and Sebastiani, 1999) is the class $\mathcal{M}_2$ of the probability distributions with strictly positive, finite variance. This score is proper, but not strictly proper, relative to $\mathcal{M}_2$.

2.3 Consistent approximations

To study the combined effects of the choices of approximation method and scoring rule in the evaluation of Bayesian predictive distributions, we introduce the notion of a *consistent* approximation procedure.

Specifically, let $(\theta_i)_{i=1,2,\ldots}$ or $(X_i)_{i=1,2,\ldots}$, where $X_i \sim F_c(\cdot|\theta_i)$, be output from a generic MCMC algorithm with the following property.

(A) The process $(\theta_i)_{i=1,2,\ldots}$ is stationary and ergodic with invariant distribution P_{post}.

As noted, assumption (A) implies that $(X_i)_{i=1,2,\ldots}$ is stationary and ergodic with invariant distribution F_0. Consider an approximation method that produces, for all sufficiently large positive integers m, an estimate $\hat{F}_m$ that is based on $(\theta_i)_{i=1}^m$ or $(X_i)_{i=1}^m$, respectively. Let S be a proper scoring rule, and let $\mathcal{F}$ be the associated natural domain. Then the approximation method is *consistent relative to the scoring rule* S *at the distribution* $F_0 \in \mathcal{F}$ if $\hat{F}_m \in \mathcal{F}$ for all sufficiently large m, and

$$d_{\mathrm{S}}(\hat{F}_m, F_0) \longrightarrow 0$$

or, equivalently, $\mathrm{S}(\hat{F}_m, F_0) \to \mathrm{S}(F_0, F_0)$ almost surely as $m \to \infty$. This formalizes the idea that under continued MCMC sampling, the approximation ought to perform as well as the

Table 3: Upper bounds on the computational complexity of approximation methods in terms of the size m of the MCMC sample $(\theta_i)_{i=1}^m$ or $(X_i)_{i=1}^m$, respectively, for pre-processing, and for the exact computation of the CRPS, Dawid–Sebastiani score (DSS) and logarithmic score (LogS).

Approximation method	Pre-processing	CRPS	DSS	LogS
MP	$\mathcal{O}(1)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m)$
ECDF	$\mathcal{O}(1)$	$\mathcal{O}(m \log m)$	$\mathcal{O}(m)$	
KD	$\mathcal{O}(m)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m)$	$\mathcal{O}(m)$
Gaussian	$\mathcal{O}(m)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$

unknown true posterior predictive distribution. We contend that this is a highly desirable property in practical work.

Note that $\hat{F}_m$ is a random quantity that depends on the sample $(\theta_i)_{i=1}^m$ or $(X_i)_{i=1}^m$. The specific form of the divergence stems from the scoring rule, which mandates convergence of a certain functional of the estimator or approximation, $\hat{F}_m$, and the theoretical posterior predictive distribution, F_0. As we will argue, this aspect has important implications for the choice of scoring rule and approximation method.

Our concept of a consistent approximation procedure is independent of the question of how well a forecast model represents the true uncertainty. The definition thus allows to separate the problem of interest, namely, to find a good approximation $\hat{F}_m$ to F_0, from the distinct task of finding a good probabilistic forecast F_0.[1] We further emphasize that we study convergence in the sample size, m, of MCMC output, given a fixed number of observations, say, T, used to fit the model. Our analysis is thus distinct from traditional Bayesian asymptotic analyses that study convergence of the posterior distribution as T becomes larger and larger (see, e.g., Gelman et al., 2014a, Section 4), thereby calling for markedly different technical tools.

3 Consistency results and computational complexity

We now investigate sufficient conditions for consistency of the aforementioned approximation methods, namely, the mixture-of-parameters (MP) estimator $\hat{F}_m^{\mathrm{MP}}$ in (2), the empirical CDF (ECDF) method $\hat{F}_m^{\mathrm{ECDF}}$ in (3), the kernel density (KD) estimate $\hat{f}_m^{\mathrm{KD}}$ in (4), and the Gaussian approximation (GA) $\hat{F}_m^{\mathrm{GA}}$ in (5). Table 3 summarizes upper bounds on the computational cost of pre-processing and computing the CRPS, Dawid–Sebastiani score (DSS) and logarithmic score (LogS) under these methods in terms of the size m of the MCMC sample $(\theta_i)_{i=1}^m$ or $(X_i)_{i=1}^m$, respectively.

Consistency requires the convergence of some functional of the approximation, $\hat{F}_m$, and the true posterior predictive distribution, F_0. The conditions to be placed on the Bayesian

[1] In fact, it is possible for an inconsistent approximation to a misspecified F_0 to approach the data generating model, whereas any consistent approximation will approach the misguided F_0. However, the misspecification can be detected by diagnostic tools such as probability integral transform (PIT) histograms; see, e.g., Diebold et al. (1998) and Gneiting et al. (2007). The appropriate remedy is to improve the model specification, rather than attempting to save a flawed model by using an inconsistent approximation.

model F_0, the estimator $\hat{F}_m$, and the dependence structure of the MCMC output depend on the scoring rule at hand.

3.1 Mixture-of-parameters estimator

We now establish consistency of the mixture-of-parameters estimator $\hat{F}_m^{\mathrm{MP}}$ in (2) relative to the CRPS, DSS and logarithmic score. The proofs are deferred to Appendix C.

Theorem 1 (consistency of mixture-of-parameters approximations relative to the CRPS and DSS). *Under assumption (A), the mixture-of-parameters approximation is consistent relative to the CRPS at every distribution F_0 with finite mean, and consistent relative to the DSS at every distribution F_0 with strictly positive, finite variance.*

Theorem 1 is the best possible result of its kind: It applies to every distribution in the natural domain and does not invoke any assumptions on the Bayesian model. In contrast, Theorem 2 hinges on rather stringent further conditions on the distribution F_0 and the Bayesian model (1), as follows.

(B) The distribution F_0 is supported on some bounded interval Ω. It admits a density, f_0, that is continuous and strictly positive on Ω. Furthermore, the density $f_c(\cdot|\theta)$ is continuous for every $\theta \in \Theta$.

Theorem 2 (consistency of mixture-of-parameters approximations relative to the logarithmic score). *Under assumptions (A) and (B), the mixture-of-parameters approximation is consistent relative to the logarithmic score at the distribution F_0.*

While we believe that the mixture-of-parameters technique is consistent under weaker assumptions, this is the strongest result that we have been able to prove. In particular, condition (B) does not allow for the case $\Omega = \mathbb{R}$. However, practical applications often involve a truncation of the support for numerical reasons, as exemplified in Section 4, and in this sense the assumption may not be overly restrictive.

As regards the computation of the CRPS, the logarithmic score and the DSS for a predictive distribution $\hat{F}_m^{\mathrm{MP}}$ of the form (2), the latter two are straightforward. To compute the CRPS, we note from eq. (21) of Gneiting and Raftery (2007) that

$$\mathrm{CRPS}\left(\hat{F}_m^{\mathrm{MP}}, y\right) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}|Z_i - y| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}|Z_i - Z_j|, \tag{8}$$

where Z_i and Z_j are independent random variables with distribution $F_c(\cdot|\theta_i)$ and $F_c(\cdot|\theta_j)$, respectively. The expectations on the right-hand side of (8) often admit closed form expressions that can be derived with techniques described by Jordan (2016) and Taillardat et al. (2016), including but not limited to the ubiquitous case of Gaussian variables. Then the evaluation requires $\mathcal{O}(m^2)$ operations, as reported in Table 3. In Appendix B, we provide details and investigate the use of numerical integration in (7), which provides an attractive, computationally efficient alternative.

3.2 Empirical CDF-based approximation

The empirical CDF-based approximation $\hat{F}_m^{\mathrm{ECDF}}$ in (3), which builds on a simulated sample $(X_i)_{i=1}^m$, is consistent relative to the CRPS and DSS under minimal assumptions. We prove

the following result in Appendix D, which is the best possible of its kind, as it applies to every distribution in the natural domain and does not invoke any assumptions on the Bayesian model.

Theorem 3 (consistency of empirical CDF-based approximations relative to the CRPS and DSS). *Under assumption (A), the empirical CDF technique is consistent relative to the CRPS at every distribution F_0 with finite mean, and consistent relative to the DSS at every distribution F_0 with strictly positive, finite variance.*

The computation of the CRPS under $\hat{F}_m^{\mathrm{ECDF}}$ requires $\mathcal{O}(m \log m)$ operations only (Hersbach, 2000). Specifically, let $X_{(1)} \leq \cdots \leq X_{(m)}$ denote the order statistics of $X_1, \ldots, X_m$, which can be obtained in $\mathcal{O}(m \log m)$ operations. Then

$$\mathrm{CRPS}\left(\hat{F}_m^{\mathrm{ECDF}}, y\right) = \sum_{i=0}^{m} \left(\alpha_i(y) \left(\frac{i}{m}\right)^2 + \beta_i(y) \left(1 - \frac{i}{m}\right)^2 \right), \tag{9}$$

where

$$\alpha_i(y) = \begin{cases} 0, & y < X_{(i)}, \\ y - X_{(i)}, & X_{(i)} \leq y \leq X_{(i+1)}, \\ X_{(i+1)} - X_{(i)}, & X_{(i+1)} < y, \end{cases} \quad \beta_i(y) = \begin{cases} X_{(i+1)} - X_{(i)}, & y < X_{(i)}, \\ X_{(i+1)} - y, & X_{(i)} \leq y \leq X_{(i+1)}, \\ 0, & X_{(i+1)} < y, \end{cases}$$

for $i = 1, \ldots, m-1$, and

$$\alpha_i(y) = \begin{cases} 0, & y < X_{(m)}, \\ y - X_{(m)}, & X_{(m)} \leq y, \end{cases} \quad \beta_i(y) = \begin{cases} X_{(1)} - y, & y \leq X_{(1)}, \\ 0, & X_{(1)} < y, \end{cases}$$

for $i = 0$ and $i = m$. A special case of eq. (8) suggests another way of computing the CRPS, in that

$$\mathrm{CRPS}\left(\hat{F}_m^{\mathrm{ECDF}}, y\right) = \frac{1}{m} \sum_{i=1}^{m} |X_i - y| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} |X_i - X_j|. \tag{10}$$

While the representations in (9) and (10) are algebraically equivalent, the latter requires $\mathcal{O}(m^2)$ operations and thus is inefficient.

3.3 Kernel density estimator

We now discuss conditions for the consistency of the kernel density estimator $\hat{f}_m^{\mathrm{KD}}$. In the present case of dependent samples $(X_i)_{i=1}^m$, judicious choices of the bandwidth h_m in (4) require knowledge of dependence properties of the sample, and the respective conditions are difficult to verify in practice. In our simulation and data examples, we use a simple implementation based on Gaussian kernel and the Silverman (1986, Section 3.4) plug-in rule for bandwidth selection. This leads to the specific form

$$\hat{F}_m^{\mathrm{KD}}(x) = \frac{1}{m} \sum_{i=1}^{m} \Phi\left(\frac{x - X_i}{h_m}\right), \tag{11}$$

where Φ denotes the CDF of the standard normal distribution, and

$$h_m = \left(\frac{4}{3}\frac{\hat{\sigma}_m^5}{m}\right)^{1/5} \approx 1.06\,\hat{\sigma}_m\, m^{-1/5}, \tag{12}$$

with $\hat{\sigma}_m$ being the standard deviation of $(X_i)_{i=1}^m$, for pre-processing costs of $\mathcal{O}(m)$, as shown in Table 3. This choice is motivated by simulation evidence in Hall et al. (1995). Using the Sheather and Jones (1991) rule or cross-validation based methods yields slightly inferior results in our experience. For a recent perspective on bandwidth selection for kernel density estimation based on MCMC samples, see Kim et al. (2016).

The score divergence associated with the logarithmic score is the Kullback-Leibler divergence, which is highly sensitive to tail behavior. Therefore, consistency of $\hat{f}_m^{\mathrm{KD}}$ requires that the tail properties of the kernel K in (4) and the true posterior predictive density f_0 be carefully matched, and any results tend to be technical (Hall, 1987; Wasserman, 2006, p. 57). Therefore, we give a brief discussion only, in which we sketch sufficient conditions for almost sure strong uniform consistency of $\hat{f}_m^{\mathrm{KD}}$, which, as noted in Appendix C, implies consistency relative to the logarithmic score under assumption (B). Roussas (1988) and Györfi et al. (1989) establish almost sure strong uniform consistency under α-mixing and other conditions. Yu (1993) provides generalizations and optimal minimax convergence rates under β-mixing conditions. The respective regularity conditions are stringent and difficult to check in practice.

Consistency of kernel density estimation approximations relative to the CRPS can be established under somewhat weaker conditions. For example, $\sup_{x\in\mathbb{R}} |\hat{F}_m^{\mathrm{KD}}(x) - F_0(x)| \leq \int_{\mathbb{R}} |\hat{f}_m^{\mathrm{KD}}(z) - f_0(z)|\,\mathrm{d}z$, and Tran (1989) studies convergence of the latter quantity under α-mixing. However, the empirical CDF-based approximation provides a simpler and computationally more efficient alternative, for which Theorem 3 guarantees consistency relative to the CRPS under minimal assumptions.

Kernel density estimation approximations are generally not consistent relative to the DSS due to the variance inflation induced by typical choices of the bandwidth. However, adaptations based on rescaling or weighting allow for kernel density estimation under moment constraints, see, e.g., Jones (1991) and Hall and Presnell (1999).

3.4 Gaussian approximation

A natural approximation method fits a member of a fixed parametric family, say $\mathcal{F}_\Gamma$, of probability distributions to the MCMC sample $(X_i)_{i=1}^m$. The problem of estimating the unknown distribution F_0 is then reduced to a finite-dimensional parameter estimation problem. However, approximation methods of this type generally fail to be consistent relative to a strictly proper scoring rule at the distribution F_0, unless F_0 belongs to $\mathcal{F}_\Gamma$.

The most important case is the *quadratic* or *Gaussian approximation*, which takes $\mathcal{F}_\Gamma$ to be the Gaussian family, so that

$$\hat{F}_m^{\mathrm{GA}}(x) = \Phi\left(\frac{x - \hat{\mu}_m}{\hat{\sigma}_m}\right),$$

where $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the empirical mean and standard deviation of $(X_i)_{i=1}^m$. If F_0 has a density f_0 that is unimodal and symmetric, the approximation can be motivated by a Taylor series expansion of the log predictive density at the mode, similar to Gaussian

approximations of posterior distributions in large-sample Bayesian inference (e.g. Kass and Raftery, 1995; Gelman et al., 2014a, Chapter 4).

The Ergodic Theorem implies that the Gaussian approximation is consistent relative to the Dawid–Sebastiani score under minimal conditions.

Theorem 4 (consistency of Gaussian approximations relative to the DSS). *Under assumption (A), the Gaussian approximation technique is consistent relative to the DSS at every distribution F_0 with strictly positive, finite variance.*

As noted, the Gaussian approximation fails to to be consistent relative to the CRPS and logarithmic score. However, the logarithmic score for the Gaussian approximation $\hat{F}_m^{\mathrm{GA}}$ corresponds to the Dawid–Sebastiani score for the empirical CDF-based approximation $\hat{F}_m^{\mathrm{ECDF}}$, in that

$$\mathrm{LogS}\Big(\hat{F}_m^{\mathrm{GA}}, y\Big) = \frac{1}{2}\left(\log 2\pi + \mathrm{DSS}\Big(\hat{F}_m^{\mathrm{ECDF}}, y\Big)\right)$$

for $y \in \mathbb{R}$. Therefore, the Gaussian approximation under the logarithmic score yields model rankings that are identical to those for the empirical CDF technique under the Dawid–Sebastiani score. This suggests that the Gaussian approximation may not be overly restrictive when used in concert with the logarithmic score, which is well in line with the findings in recent empirical work by Warne et al. (2016). However any such use is subject to the caveat that a proper, but not strictly proper, scoring rule (namely the Dawid–Sebastiani score) is employed.

4 Simulation study

We now investigate the various approximation methods in a simulation study that is designed to emulate realistic MCMC behavior with dependent samples. Here, the posterior predictive distribution F_0 is known by construction, and so we can compare the different approximations to the true forecast distribution.

In order to judge the quality of an approximation $\hat{F}_m$ of F_0 we consider the score divergence $d_{\mathrm{S}}(\hat{F}_m, F_0)$. Note that $d_{\mathrm{S}}(\hat{F}_m, F_0)$ is a random variable, since $\hat{F}_m$ depends on the particular MCMC sample $(\theta_i)_{i=1}^m$ or $(X_i)_{i=1}^m$. In our results below, we therefore consider the distribution of $d_{\mathrm{S}}(\hat{F}_m, F_0)$ across repeated simulation runs. For a generic approximation method producing an estimate $\hat{F}_m$, we proceed as follows:

- For simulation run $k = 1, \ldots, K$:
  - Draw MCMC samples $(\theta_i^{(k)})_{i=1}^m$ and $(X_i^{(k)})_{i=1}^m$
  - Compute the approximation $\hat{F}_m^{(k)}$ and the divergence $d_{\mathrm{S}}(\hat{F}_m^{(k)}, F_0)$ for the approximation methods and scoring rules under consideration.
- For each approximation method and scoring rule, summarize the distribution of $d_{\mathrm{S}}(\hat{F}_m^{(1)}, F_0), \ldots, d_{\mathrm{S}}(\hat{F}_m^{(K)}, F_0)$.

In order to simplify notation, we typically suppress the superscript that identifies the Monte Carlo iteration. The results below are based on $K = 1\,000$ replicates.

Table 4: Hyper-parameters for the data generating process in the simulation setting of equations (14) to (17).

Parameter	Main role	Value(s) considered
α	Persistence of θ_i^2	$\{0.1, 0.5, 0.9\}$
s	Unconditional mean of θ_i^2	2
n	Unconditional variance of θ_i^2	$\{12, 20\}$

4.1 Data generating process

We generate sequences $(\theta_i)_{i=1}^m$ and $(X_i)_{i=1}^m$ in such a way that the invariant distribution,

$$F_0(x) = \int_{(0,\infty)} \Phi\Big(\frac{x}{\theta}\Big)\, \mathrm{d}H_0(\theta^2),$$

where Φ denotes the standard normal CDF, is a compound Gaussian distribution or normal scale mixture. Depending on the measure H_0, which assumes the role of the posterior distribution P_{post} in the general Bayesian model (1), F_0 can be modeled flexibly, including many well-known parametric distributions (Gneiting, 1997). As detailed below, our choice of H_0 implies that

$$F_0(x) = \mathsf{T}\left(x \,\middle|\, 0, \frac{ns}{n+2}, n+2\right) \tag{13}$$

where $\mathsf{T}(\cdot|a,b,c)$ denotes the CDF of a variable Z with the property that $(Z-a)/\sqrt{b}$ is standard Student t distributed with c degrees of freedom. To mimic a realistic MCMC scenario with dependent draws, we proceed as proposed by Fox and West (2011). Given parameter values $n > 0$, $s > 0$ and $\alpha \in (-1, 1)$, let

$$\psi_i \sim \mathrm{IG}\left(\frac{1}{2}(n+3), \frac{1}{2}ns(1-\alpha^2)\right), \tag{14}$$

$$\upsilon_i \,|\, \psi_i \sim \mathcal{N}\left(\alpha, \frac{\psi_i}{ns}\right), \tag{15}$$

$$\theta_i^2 = \psi_i + \upsilon_i^2\, \theta_{i-1}^2, \tag{16}$$

$$X_i \,|\, \theta_i^2 \sim \mathcal{N}(0, \theta_i^2) \tag{17}$$

where IG is the Inverse Gamma distribution, parametrized such that $Z \sim \mathrm{IG}(a,b)$ when $1/Z \sim \mathrm{G}(a,b)$, with G being the Gamma distribution with shape $a \geq 0$ and rate $b > 0$.

Table 4 summarizes our choices for the parameter configurations of the data generating process. The parameter α determines the persistence of the chain, in that the unconditional mean of υ_i^2, which can be viewed as an average autoregressive coefficient (Fox and West, 2011, Section 2.3), is given by $(n\alpha^2+1)/(n+1)$. We consider three values, aiming to mimic MCMC chains with different persistence properties. The parameter s represents a scale effect that is not of importance here. Finally, the parameter n governs the tail thickness of the unconditional Student t distribution in (13). We consider values of 12 and 20 that seem realistic for macroeconomic variables, such as the growth rate of the gross domestic product (GDP), that feature prominently in the empirical literature.

4.2 Approximation methods

We consider the following approximation methods, which have been discussed in detail in Section 3. The first approximation uses a sequence $(\theta_i)_{i=1}^m$ of parameter draws, and the other three employ an MCMC sample $(X_i)_{i=1}^m$.

- Mixture-of-parameters estimator $\hat{F}_m^{\mathrm{MP}}$ in (2), which here is of the form

$$\hat{F}_m^{\mathrm{MP}}(x) = \frac{1}{m}\sum_{i=1}^{m} \Phi\left(\frac{x}{\theta_i}\right),$$

  where θ_i is the predictive standard deviation drawn in MCMC iteration i.

- Empirical CDF-based approximation $\hat{F}_m^{\mathrm{ECDF}}$ in (3).

- The nonparametric kernel density estimator $\hat{f}_m^{\mathrm{KD}}$ using a Gaussian kernel and the Silverman rule for bandwidth selection, with predictive CDF $\hat{F}_m^{\mathrm{KD}}$ of the form (11).

- Gaussian approximation $\hat{F}_m^{\mathrm{GA}}$ in (5).

It is interesting to observe that here $\hat{F}_m^{\mathrm{MP}}$ is a scale mixture of centered Gaussian distributions, and $\hat{F}_m^{\mathrm{KD}}$ is a location mixture of normal distributions, whereas the quadratic approximation $\hat{F}_m^{\mathrm{GA}}$ is a single Gaussian.

The conditions for consistency of the mixture-of-parameters and empirical CDF approximations relative to the CRPS in Theorems 1 and 3 are satisfied. Furthermore, one might argue that numerically the support of F_0 and $\hat{F}_m^{\mathrm{MP}}$ is bounded (see below), and then the assumptions of Theorem 2 also are satisfied. Clearly, the Gaussian approximation fails to be consistent relative to the CRPS or the logarithmic score, as F_0 is not Gaussian.

For each approximation method, scoring rule S, sample size m and replicate k, we evaluate the score divergence $d_{\mathrm{S}}(\hat{F}_m^{(k)}, F_0)$. The divergence takes the form of a univariate integral (cf. Table 2) that is not available in closed form. Therefore, we compute $d_{\mathrm{S}}(\hat{F}_m^{(k)}, F_0)$ by numerical integration as implemented in the R function integrate. This is unproblematic if the scoring rule is the CRPS. For the logarithmic score, the integration is numerically challenging, as the logarithm of the densities needs to be evaluated in their tails. We therefore truncate the support of the integral to the minimal and maximal values that yield numerically finite values of the integrand.

4.3 Main results

In the interest of brevity, we restrict attention to results for a single set of parameters of the data generating process, namely $(\alpha, s, n) = (0.5, 2, 12)$. This implies an unconditional Student t distribution with 14 degrees of freedom, and intermediate autocorrelation of the MCMC draws. The results for the other parameter constellations in Table 4 are similar and available in the Online Supplement.

Figure 1 illustrates the performance of the approximation methods under the logarithmic score and the CRPS, by showing the distribution of the score divergence $d_{\mathrm{S}}(\hat{F}_m, F_0)$ as the sample size m grows. The mixture-of-parameters estimator dominates the other methods by a wide margin: Its divergences are very close to zero, and show little variation

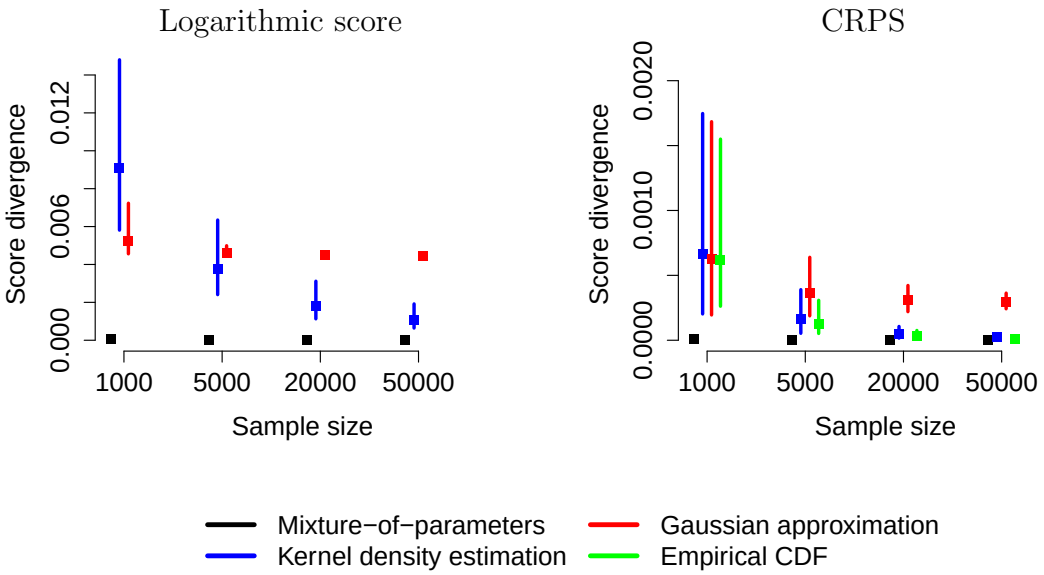

Figure 1: Score divergences in the simulation study with $(\alpha, s, n) = (0.5, 2, 12)$. For a given method and MCMC sample size, the bars range from the 10th to the 90th percentile of the score divergences across 1 000 replicates. The squares mark the respective medians.

across replicates. Under the logarithmic score, the performance of the kernel density estimator is highly variable across the replicates, even for large sample sizes. The variability is less under the CRPS, where the kernel density approach using the Silverman (1986) rule of thumb for bandwidth selection performs similar to the empirical CDF-based approximation. Other bandwidth selection rules we have experimented with tend to be inferior, as indicated by slower convergence and higher variability across replicates. Finally, we observe the lack of consistency of the Gaussian approximation.

Figure 2 provides insight into the performance of the mixture-of-parameters approximation for small MCMC samples. Using as few as 150 draws, the method attains a lower median CRPS divergence than the kernel density estimator based on 20 000 draws. The superiority of the mixture-of-parameters estimator is even more pronounced under the logarithmic score, where only 50 draws are required to outperform the kernel density estimator based on 20 000 draws.

4.4 Thinning the MCMC sample

We further investigate the effect of thinning the Markov chains. Thinning a chain by a factor of τ means that only every τth simulated value is retained, and the rest is discarded. Thinning is often applied routinely with the goal of reducing autocorrelation in the draws. Of the articles listed in Table 1, about one in four explicitly reports thinning of the simulation output, with thinning factors ranging from 2 to 100. Here we compare three sampling approaches:

(S1) 4 000 MCMC draws, without thinning

(S2) 4 000 MCMC draws, retaining every 10th draw from a sequence of 40 000 draws

(S3) 40 000 MCMC draws, without thinning

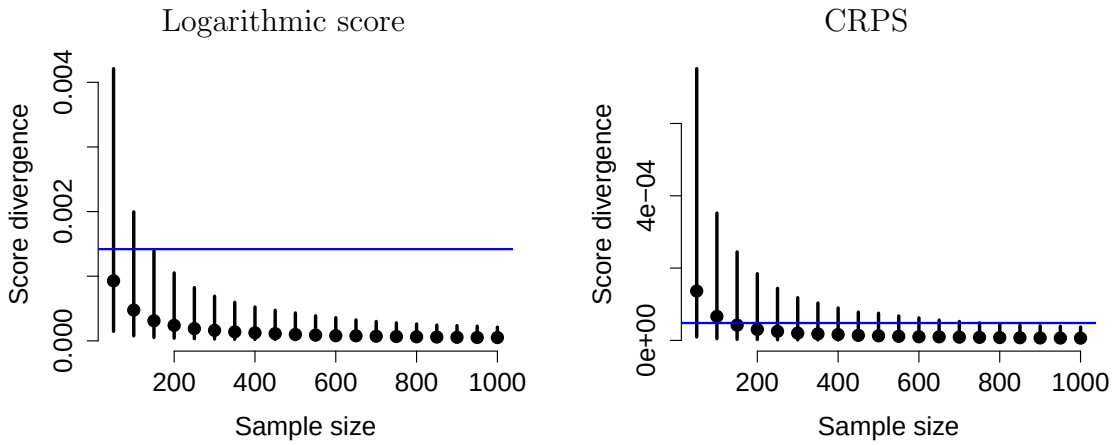

Figure 2: Performance of the mixture-of-parameters estimator. The design is as in Figure 1, but for smaller sample sizes. For comparison, the blue horizontal line marks the median divergence of the kernel density estimator based on 20 000 draws.

Note that the samples in S1 and S3 have the same dynamic properties, whereas S2 will typically produce a chain with less autocorrelation. Furthermore, S2 and S3 require the same computing time, which exceeds that of S1 by a factor of ten. Figure 3 summarizes the corresponding simulation results, using parameter values $s = 2$ and $n = 12$, and varying values of the persistence parameter α. We report results for four popular combinations of scoring rules and approximation methods.

As expected, S2 tends to outperform S1: When the sample size is held fixed, less autocorrelation entails more precise estimators. While the difference in performance is modest in most cases, S2 attains large (relative) gains over S1 when the mixture-of-parameters estimator is applied to a very persistent sample with $\alpha = 0.9$. This can be explained by the direct effect of the persistence parameter α on the parameter draws $(\theta_i)_{i=1}^m$, whereas the influence is less immediate for the KDE and ECDF approximation methods, which are based on the sequence $(X_i)_{i=1}^m$ obtained in an additional sampling step. Furthermore, S3 outperforms S2 in all cases covered. While the effects of thinning have not been studied in the context of predictive distributions before, this observation is in line with extant reports of the greater precision of unthinned chains (Geyer, 1992; MacEachern and Berliner, 1994; Link and Eaton, 2012). The performance gap between S3 and S2 is modest for the mixture-of-parameters estimator (top row of Figure 3), but very pronounced for the other estimators.

From a practical perspective, thinning thus seems justified if, and only if, a small MCMC sample is desired and the mixture-of-parameter estimator is applied. Two arguments in favor of a small sample appear particularly relevant even today. First, storing large amounts of data on public servers (as is often done for replication purposes) may be costly or inconvenient. Second, post-processing procedures such as score computations applied to the MCMC sample may be computationally demanding (cf. Table 3), and therefore may encourage and justify thinning.

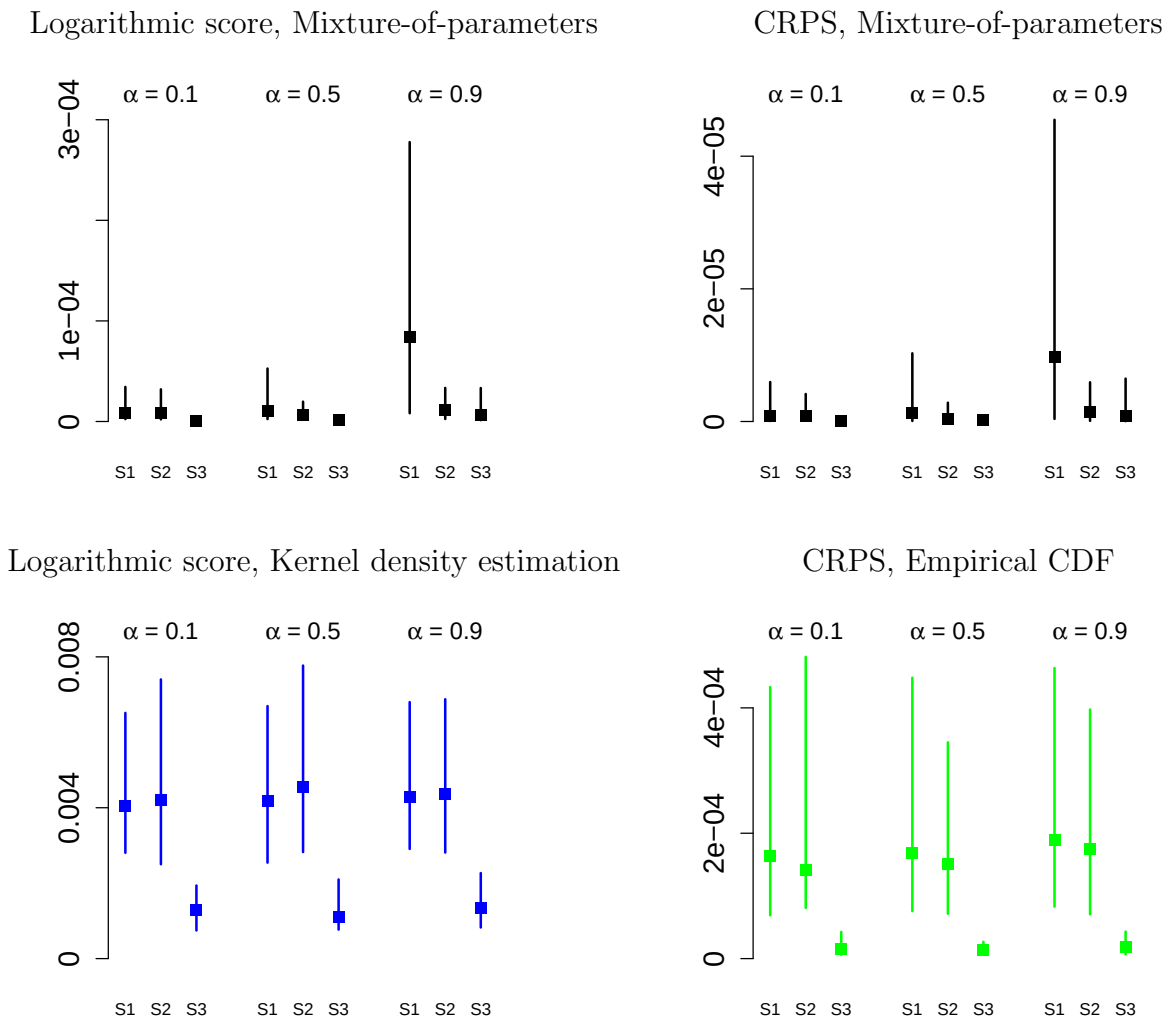

Figure 3: Performance of three sampling strategies: S1: 4 000 draws, without thinning; S2: 4 000 MCMC draws, retaining every 10th draw from a sequence of 40 000 draws; S3: 40 000 draws, without thinning. Bars range from the 10th to the 90th percentile of the score divergences across 1 000 replicates. The squares mark the respective medians.

5 Economic data example

In real-world uses of Bayesian forecasting methods, the posterior predictive distribution F_0 is typically not available in closed form. Therefore, computing or estimating the object of interest for assessing consistency, i.e., the score divergence $d_S(\hat{F}_m, F_0)$, is not feasible. In the subsequent data example, we thus compare the approximation methods via their out-of-sample predictive performance, and examine the variation of the mean scores across chains obtained by replicates with distinct random seeds. While studying the predictive performance does not allow to assess consistency of the approximation methods, it does allow us to assess the variability and applicability of the approximations in a practical setting.

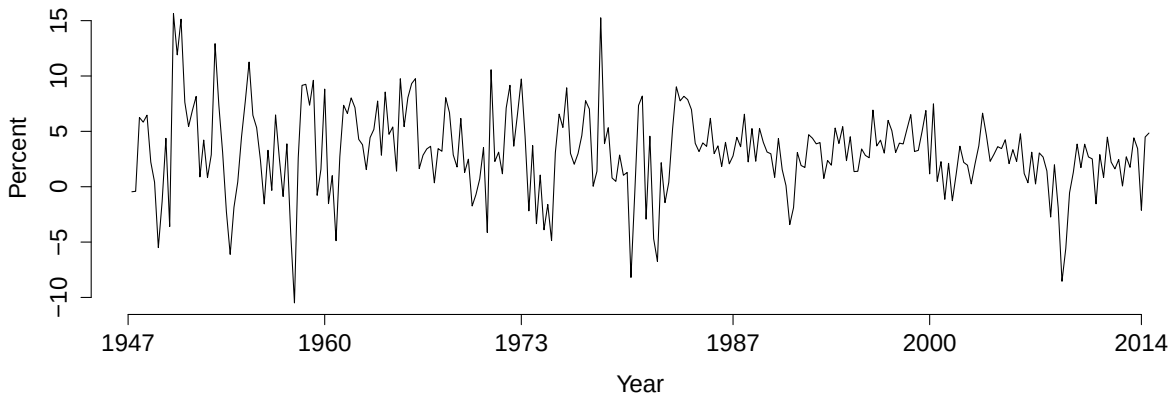

Figure 4: Quarterly growth rate of U.S. real GDP from the second quarter of 1947 to the third quarter of 2014, using the data vintage from the first quarter of 2015.

5.1 Data

We consider quarterly growth rates of U.S. real gross domestic product (GDP), as illustrated in Figure 4. The training sample used for model estimation is recursively expanded as forecasting moves forward in time. We use the real-time data set provided by the Federal Reserve Bank of Philadelphia[2], which provides historical snapshots of the data vintages available at any given date in the past, and consider forecasts for the period from the second quarter of 1996 to the third quarter of 2014, for a total of $T = 74$ forecast cases. For brevity, we present results for a prediction horizon of one quarter only. The Online Supplement contains results for longer horizons, which are qualitatively similar to the ones presented here.

5.2 Probabilistic forecasts

To construct density forecasts, we consider an autoregressive (AR) model with a single lag and state-dependent error term variance, in that

$$Y_t = \nu + \alpha Y_{t-1} + \varepsilon_t, \tag{18}$$

where $\varepsilon_t \sim \mathcal{N}(0, \eta^2_{s_t})$ and $s_t \in \{1, 2\}$ is a discrete state variable that switches according to a first-order Markov chain. The model, which is a variant of the Markov switching model proposed by Hamilton (1989), provides a simple description of time-varying heteroscedasticity. The latter is an important stylized feature of macroeconomic time series (see, e.g., Clark and Ravazzolo, 2015).

We conduct Bayesian inference via a Gibbs sampler, for which we give details in Appendix E. Let θ_i denote the complete set of latent states and model parameters at iteration i of the Gibbs sampler. Conditional on θ_i, the predictive distribution under the model in (18) is Gaussian with mean $\mu_i = \mu(\theta_i)$ and standard deviation $\sigma_i = \sigma(\theta_i)$, where we suppress time and forecast horizon for simplicity. As an illustration, Figure 5 presents histograms

[2] `http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/`

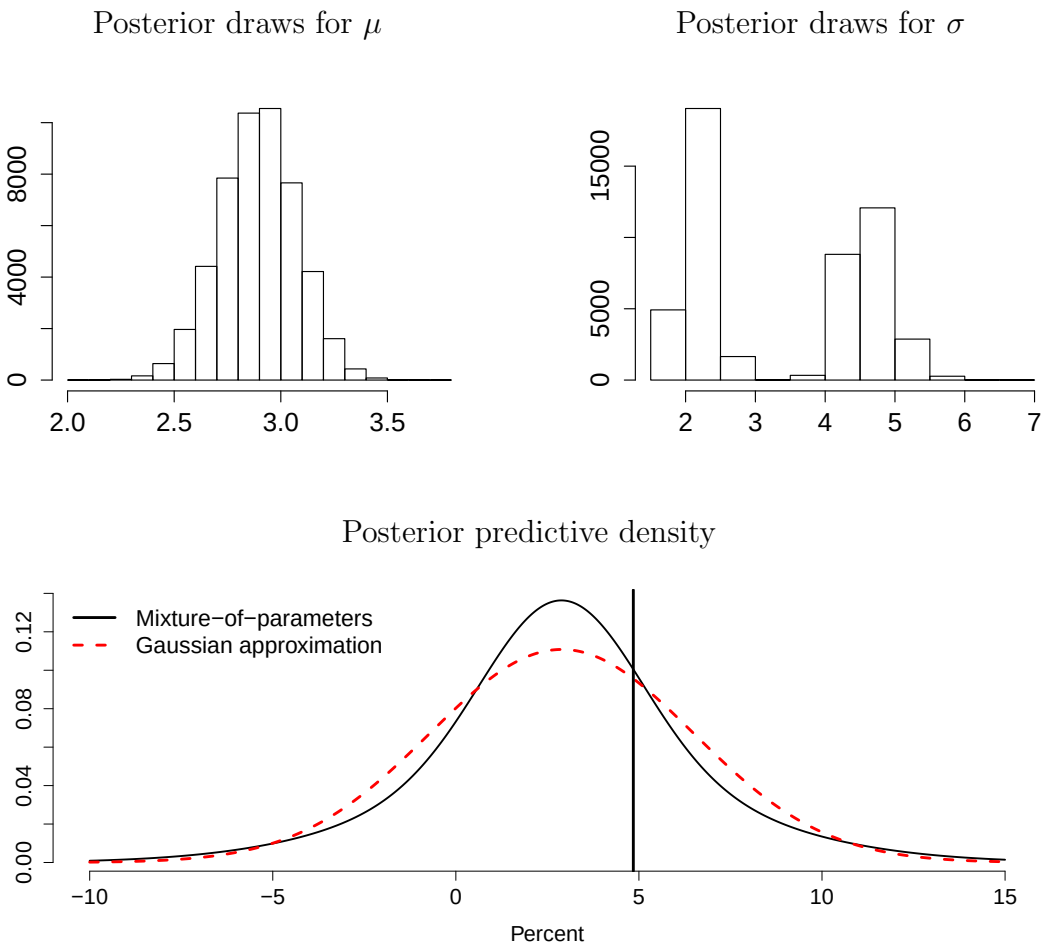

Figure 5: Forecasts of the growth rate of U.S. real GDP in the third quarter of 2014. Top: Histograms based on 40 000 posterior draws, without thinning. Bottom: Posterior predictive densities, with the vertical line indicating the realized rate. The kernel density estimator is visually indistinguishable from the mixture-of-parameters estimator, and is therefore omitted.

of the MCMC draws for μ and σ for forecasts made in the third quarter of 2014. The bimodal histogram for σ clearly reflects the two states; estimation uncertainty about the state-specific quantities creates additional variation. The predictive distributions obtained by the mixture-of-parameters and Gaussian approximations are also shown.[3]

At each forecast origin date $t = 1, \ldots, T = 74$, we produce 10 000 burn-in draws, and use 40 000 draws post burn-in. We construct 16 parallel chains in this way. The (time-averaged) mean score of a given approximation method, based on m MCMC draws within chain $c = 1, \ldots, 16$, is

$$\bar{\mathrm{S}}_{m,c} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{S}(\hat{F}_{m,c,t}, y_t),$$

[3]Alternatively, one could construct a mixture-of-parameters estimator by using a modified parameter vector $\tilde{\theta}_i$ that comprises *predicted probabilities* (rather than simulated realizations) of the future latent states. Given m MCMC iterations, this alternative estimator yields a mixture of $2m$ Gaussian component distributions. For reasons of computational and conceptual simplicity, we prefer the m-component estimator described in the text.

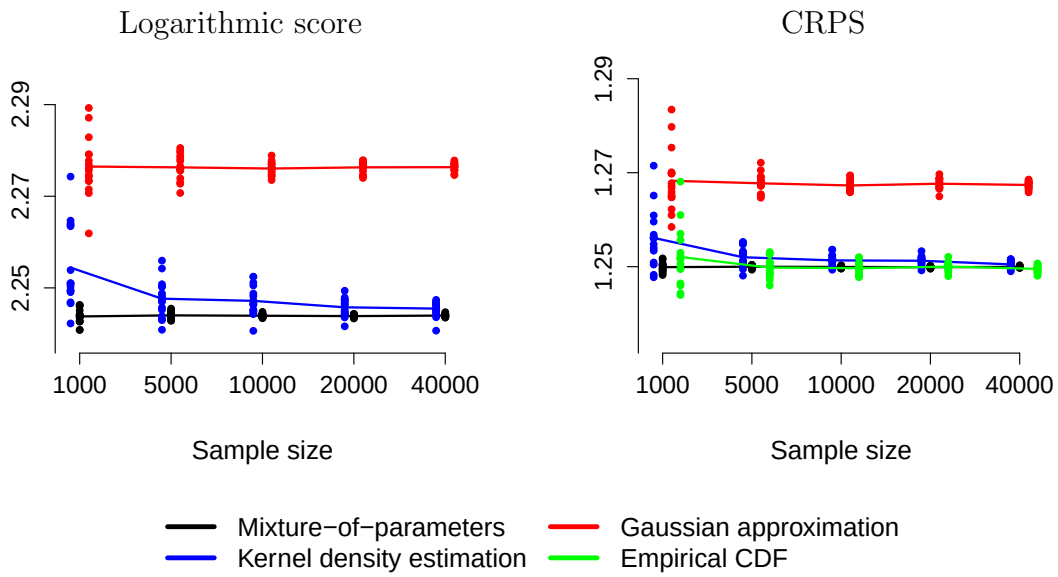

Figure 6: Mean score in the data example against sample size. The dots represent 16 parallel MCMC chains, and the lines connect averages across chains.

where $\hat{F}_{m,c,t}$ is the probabilistic forecast at time t. The variation of $\bar{\mathrm{S}}_{m,c}$ across chains c is due to differences in random seeds. From a pragmatic perspective, a good approximation method should be such that the values $(\bar{\mathrm{S}}_{m,c})_{c=1}^{16}$ are small and display little variation.

5.3 Results

In Figure 6, the mean score is plotted against the size of the MCMC sample. The mixture-of-parameters approximation outperforms its competitors: Its scores display the smallest variation across chains, for both the CRPS and the logarithmic score, and for all sample sizes. The scores of the mixture-of-parameter estimator also tend to be lower (i.e., better) than the scores for the other methods. The kernel density estimator performs poorly for small sample sizes, with the scores varying substantially across chains. Under the CRPS, the kernel density estimator is dominated by the empirical CDF technique, which can be interpreted as kernel density estimation with a bandwidth of zero.

6 Discussion

We have investigated how to make and evaluate probabilistic forecasts based on MCMC output. The formal notion of consistency allows us to assess the appropriateness of approximation methods within the framework of proper scoring rules. Despite their popularity in the literature, Gaussian approximations generally fail to be consistent. Conditions for consistency depend on the scoring rule of interest, and we have demonstrated that the mixture-of-parameters and empirical CDF-based approximations are consistent relative to the CRPS under minimal conditions. Proofs of consistency relative to the logarithmic score generally rely on stringent assumptions.

In view of these theoretical considerations as well as the practical perspective taken in our simulation and data examples, we recommend the use of the mixture-of-parameters

estimator, which provides an efficient approximation method and outperforms all alternatives. This can be explained by the fact that it efficiently exploits the parametric structure of the Bayesian model. Other approaches either impose restrictive assumptions, or fail to exploit what is known about the structure of the model, thus leading to unwarranted variability in the approximation. The empirical CDF-based approximation provides a good alternative if the conditional distributions fail to be available in closed form, or if for some reason the draws are to be made directly from the posterior predictive distribution, as opposed to using parameter draws.

Our recommendations have been implemented in the scoringRules package for R (R Core Team, 2016); see Jordan et al. (2016) for details. The functions and default choices aim to provide readily applicable and efficient approximations. The mixture-of-parameters estimator depends on the specific structure of the Bayesian model and can therefore not be covered in full generality. However, the implemented analytical solutions of the CRPS and logarithmic score allow for straightforward and efficient computation. The scoringRules package further contains functions and data for replicating the simulation and case study, with details provided at `https://github.com/FK83/scoringRules/blob/master/KLTG2016_replication.pdf`.

Ferro (2014) studies the notion of a fair scoring rule in the context of ensemble weather forecasts. A scoring rule is called *fair* if the expected score is optimal for samples with members that behave as though they and the verifying observation were sampled from the same distribution. While certainly relevant in the context of meteorological forecast ensembles, where the sample size m is typically between 10 and 50, these considerations seem less helpful in the context of MCMC output, where m is on the order of thousands and can be increased at low cost. Furthermore, the proposed small sample adjustments and the characterization of fair scores hold for independent samples only, an assumption that is thoroughly violated in the case of MCMC.

We are interested in evaluating probabilistic forecasts produced via MCMC, so that the predictive performance of a model during an out-of-sample, test or evaluation period can be used to estimate its forecast performance on future occasions. In contrast, information criteria suggest a different route towards estimating forecast performance (Spiegelhalter et al., 2002; Watanabe, 2010; Hooten and Hobbs, 2015). They consider a method's in-sample performance, and account for model complexity via penalty terms. Preferred ways of doing so have been the issue of methodological debate, and a consensus has not been reached; see, e.g., the comments in Gelman et al. (2014b) and Spiegelhalter et al. (2014). This present analysis does not concern in-sample comparisons, and does not address the question of whether these are more or less effective than out-of-sample comparisons. However, our results and observations indicate that out-of-sample comparisons of the type considered here yield robust results across a range of implementation choices.

Finally, we have focused on univariate forecast distributions. Some aspects of our analysis, such as the insights regarding the superiority of the mixture-of-parameters approximation, transfer readily to the evaluation of probabilistic forecasts for multivariate quantities. Nevertheless, the multivariate case features some novel challenges, including but not limited to the choice of the scoring rule, as discussed by Gneiting et al. (2008) and Scheuerer and Hamill (2015), and these we leave to future work.

Acknowledgements

The work of Tilmann Gneiting and Fabian Krüger was funded by the European Union Seventh Framework Programme under grant agreement 290976. Sebastian Lerch and Thordis L. Thorarinsdottir acknowledge support by the Volkswagen Foundation through the program "Mesoscale Weather Extremes — Theory, Spatial Modelling and Prediction (WEX-MOP)". Lerch further acknowledges support by Deutsche Forschungsgemeinschaft (DFG) through RTG 1953 "Statistical Modeling of Complex Systems and Processes" as well as project C7 ("Statistical postprocessing and stochastic physics for ensemble predictions") within SFB/TRR 165 "Waves to Weather". Gneiting, Krüger and Lerch thank the Klaus Tschira Foundation for infrastructural support at the Heidelberg Institute for Theoretical Studies (HITS). Helpful comments by Werner Ehm, Sylvia Frühwirth-Schnatter, Alexander Jordan, as well as seminar and conference participants at HITS, Karlsruhe Institute of Technology (KIT), University of Bern, the Extremes 2014 symposium (Hannover), Conference on Computational and Financial Econometrics (CFE, Pisa, 2014), German Probability and Statistics Days (GPSD, Bochum, 2016) and International Society for Bayesian Analysis meeting (ISBA, Sardinia, 2016) are gratefully acknowledged. Furthermore, we thank Gianni Amisano for sharing his program code for Bayesian Markov switching models.

References

Adolfson, M., Linde, J. and Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, 26, 289–328.

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

Baştürk, N., Cakmakli, C., Ceyhan, S. P. and Van Dijk, H. K. (2014). Posterior-predictive evidence on US inflation using extended new Keynesian Phillips curve models with non-filtered data. *Journal of Applied Econometrics*, 29, 1164–1182.

Bauwens, L., Koop, G., Korobilis, D. and Rombouts, J. V. K. (2015). The contribution of structural break models to forecasting macroeconomic series. *Journal of Applied Econometrics*, 30, 596–620.

Belmonte, M. A. G., Koop, G. and Korobilis, D. (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33, 80–94.

Berg, T. O. (2017). Forecast accuracy of a BVAR under alternative specifications of the zero lower bound. *Studies in Nonlinear Dynamics & Econometrics*, in press, `https://doi.org/10.1515/snde--2015--0084`.

Berg, T. O. and Henzel, S. R. (2015). Point and density forecasts for the Euro area using Bayesian VARs. *International Journal of Forecasting*, 31, 1067–1095.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2014). Assessing exceedance of ozone standards: A space-time downscaler for fourth highest ozone concentrations. *Environmetrics*, 25, 279–291.

Brandt, P. T., Freeman, J. R. and Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30, 944–962.

Broms, K. M., Hooten, M. B. and Fitzpatrick, R. M. (2016). Model selection and assessment for multi-species occupancy models. *Ecology*, 97, 1759–1770.

Carriero, A., Clark, T. E. and Marcellino, M. (2015a). Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30, 46–73.

Carriero, A., Clark, T. E. and Marcellino, M. (2015b). Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178, 837–862.

Carriero, A., Clark, T. E. and Marcellino, M. (2016). Common drifting volatility in large Bayesian VARs. *Journal of Business and Economic Statistics*, 34, 375–390.

Carriero, A., Mumtaz, H. and Theophilopoulou, A. (2015c). Macroeconomic information, structural change, and the prediction of fiscal aggregates. *International Journal of Forecasting*, 31, 325–348.

Clark, T. and McCracken, M. (2013). Advances in forecast evaluation. In *Handbook of Economic Forecasting* (G. Elliott and A. Timmermann, eds.), vol. 2. Elsevier, 1107–1201.

Clark, T. E. (2011). Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics*, 29, 327–341.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30, 551–575.

Craiu, R. V. and Rosenthal, J. S. (2014). Bayesian computation via Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 1, 179–201.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27, 65–81.

De la Cruz, R. and Branco, M. D. (2009). Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves. *Biometrical Journal*, 51, 588–609.

Dehling, H. and Philipp, W. (2002). Empirical process techniques for dependent data. In *Empirical Process Techniques for Dependent Data* (H. Dehling, T. Mikosch and M. Sørensen, eds.). Birkhäuser, Boston, 3–113.

Delatola, E.-I. and Griffin, J. E. (2011). Bayesian nonparametric modelling of the return distribution with stochastic volatility. *Bayesian Analysis*, 6, 901–926.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863–883.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263.

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

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140, 1917–1923.

Fox, E. B. and West, M. (2011). Autoregressive models for variance matrices: Stationary inverse Wishart processes. Preprint, available at `http://arxiv.org/abs/1107.5239`.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23, 579–594.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014a). *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC, Boca Raton.

Gelman, A., Hwang, J. and Vehtari, A. (2014b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–760.

Genon-Catalot, V., Jeantheau, T. and Larédo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli*, 6, 1051–1079.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons, Hoboken.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions

of asset returns. *International Journal of Forecasting*, 26, 216–230.

Geweke, J. and Amisano, G. (2011). Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26, 1–29.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7, 473–483.

Geyer, C. J. (1995). Conditioning in Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 4, 148–154.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74, 1545–1578.

Giannone, D., Lenza, M. and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97, 436–451.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59, 375–384.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29, 411–422.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds (with discussion and rejoinder). *Test*, 17, 211–264.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 14, 107–114.

Greenberg, E. (2013). *Introduction to Bayesian Econometrics*. 2nd ed. Cambridge University Press, New York.

Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942.

Groen, J. J. J., Paap, R. and Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business and Economic Statistics*, 31, 29–44.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007, 202–225.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer, Berlin.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Annals of Statistics*, 15, 1491–1519.

Hall, P., Lahiri, S. N. and Truong, Y. K. (1995). On bandwidth choice for density estimation with dependent data. *Annals of Statistics*, 23, 2241–2263.

Hall, P. and Presnell, B. (1999). Density estimation under constraints. *Journal of Computational and Graphical Statistics*, 8, 259–277.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357–384.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76, 675–684.

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures:*

Festschrift in Honour of Ludwig Fahrmeir (T. Kneib and G. Tutz, eds.). Physica-Verlag HD, Heidelberg, 91–110.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85, 3–28.

Jensen, M. J. and Maheu, J. M. (2010). Bayesian semiparametric stochastic volatility modeling. *Journal of Econometrics*, 157, 306–316.

Jensen, M. J. and Maheu, J. M. (2013). Bayesian semiparametric multivariate GARCH modeling. *Journal of Econometrics*, 176, 3–17.

Jensen, M. J. and Maheu, J. M. (2014). Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture. *Journal of Econometrics*, 178, 523–538.

Jin, X. and Maheu, J. M. (2013). Modeling realized covariances and returns. *Journal of Financial Econometrics*, 11, 335–369.

Jin, X. and Maheu, J. M. (2016). Bayesian semiparametric modeling of realized covariance matrices. *Journal of Econometrics*, 192, 19–39.

Jochmann, M., Koop, G. and Strachan, R. W. (2010). Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks. *International Journal of Forecasting*, 26, 326–347.

Jones, M. C. (1991). On correcting for variance inflation in kernel density estimation. *Computational Statistics & Data Analysis*, 11, 3–15.

Jordan, A. (2016). Facets of forecast evaluation. Ph.D. thesis, Karlsruhe Institute of Technology, available at `https://publikationen.bibliothek.kit.edu/1000063629`.

Jordan, A., Krüger, F. and Lerch, S. (2016). *The scoringRules package*. URL `https://cran.r-project.org/web/packages/scoringRules/`.

Kallache, M., Maksimovich, E., Michelangeli, P.-A. and Naveau, P. (2010). Multimodel combination by a Bayesian hierarchical model: Assessment of ice accumulation over the oceanic Arctic region. *Journal of Climate*, 23, 5421–5436.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kastner, G. (2016). Dealing with stochastic volatility in time series using the R package stochvol. *Journal of Statistical Software*, 69, 1–30.

Kim, H. J., MacEachern, S. N. and Jung, Y. (2016). Bandwidth selection for kernel density estimation with a Markov chain Monte Carlo sample. Preprint, available at `http://arxiv.org/abs/1607.08274`.

Krüger, F., Clark, T. E. and Ravazzolo, F. (2015). Using entropic tilting to combine BVAR forecasts with external nowcasts. *Journal of Business and Economic Statistics*, in press, `doi:10.1080/07350015.2015.1087856`.

Krüger, F. and Nolte, I. (2016). Disagreement versus uncertainty: Evidence from distribution forecasts. *Journal of Banking & Finance*, 72, 172–186.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons.

Leininger, T. J., Gelfand, A. E., Allen, J. M. and Silander Jr, J. A. (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 314–334.

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

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3, 112–115.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*,

60, 213–223.

Liu, C. and Maheu, J. M. (2009). Forecasting realized volatility: a Bayesian model-averaging approach. *Journal of Applied Econometrics*, 24, 709–733.

Lopes, H. F., Salazar, E. and Gamerman, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis*, 3, 759–792.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48, 188–190.

Maheu, J. M. and Gordon, S. (2008). Learning, forecasting and structural breaks. *Journal of Applied Econometrics*, 23, 553–583.

Maheu, J. M. and McCurdy, T. H. (2009). How useful are historical data for forecasting the long-run equity return distribution? *Journal of Business and Economic Statistics*, 27, 95–112.

Maheu, J. M., McCurdy, T. H. and Song, Y. (2012). Components of bull and bear markets: Bull corrections and bear rallies. *Journal of Business and Economic Statistics*, 30, 391–403.

Maheu, J. M. and Song, Y. (2014). A new structural break model, with an application to Canadian inflation forecasting. *International Journal of Forecasting*, 30, 144–160.

Maneesoonthorn, W., Martin, G. M., Forbes, C. S. and Grose, S. D. (2012). Probabilistic forecasts of volatility and its risk premia. *Journal of Econometrics*, 171, 217–236.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

Panagiotelis, A. and Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. *International Journal of Forecasting*, 24, 710–727.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data — Borrowing strength from related populations. *Annals of Applied Statistics*, 6, 304–333.

Risser, M. D. and Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26, 284–297.

Rodrigues, L. R. L., García-Serrano, J. and Doblas-Reyes, F. (2014). Seasonal forecast quality of the West African monsoon rainfall regimes by multiple forecast systems. *Journal of Geophysical Research: Atmospheres*, 119, 7908–7930.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832–837.

Roussas, G. G. (1988). Nonparametric estimation in mixing sequences of random variables. *Journal of Statistical Planning and Inference*, 18, 135–149.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.

Sahu, S. K., Bakar, K. S. and Awang, N. (2015). Bayesian forecasting using spatio-temporal models with applications to ozone levels in the eastern United States. In *Geometry Driven Statistics* (I. L. Dryden and J. T. Kent, eds.), chap. 13. John Wiley & Sons, 260–281.

Salazar, E., Sansó, B., Finley, A. O., Hammerling, D., Steinsland, I., Wang, X. and Delamater, P. (2011). Comparing and blending regional climate model predictions for the American Southwest. *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 586–605.

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143, 1321–1334.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 53, 683–690.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Annals of Applied Statistics*, 6, 1452–1477.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation

based modelling of large space-time data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77, 3–33.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smith, M. S. and Vahey, S. (2016). Asymmetric density forecasting of U.S. macroeconomic variables. *Journal of Business and Economic Statistics*, 34, 416–434.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion and rejoinder). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76, 485–493.

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144, 2375–2393.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540.

Thorarinsdottir, T. L., Gneiting, T. and Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1, 522–534.

Tran, L. T. (1989). The L_1 convergence of kernel density estimates under dependence. *Canadian Journal of Statistics*, 17, 197–208.

Tran, M.-N., Pitt, M. K. and Kohn, R. (2016). Adaptive Metropolis–Hastings sampling using reversible dependent mixture proposals. *Statistics and Computing*, 26, 361–381.

Trombe, P.-J., Pinson, P. and Madsen, H. (2012). A general probabilistic forecasting framework for offshore wind power fluctuations. *Energies*, 5, 621–657.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Warne, A., Coenen, G. and Christoffel, K. (2016). Marginalized predictive likelihood comparisons of linear Gaussian State-Space models with applications to DSGE, DSGE-VAR and VAR models. *Journal of Applied Econometrics*, 32, 103–119.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.

Yu, B. (1993). Density estimation in the L^∞ norm for dependent data with applications to the Gibbs sampler. *Annals of Statistics*, 21, 711–735.

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G. and Micheas, A. C. (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110, 6–15.

Appendix

A Literature survey

To survey how probabilistic forecasts based on MCMC output are evaluated in the literature, we made an attempt to conduct a systematic survey. To obtain a broad set of

candidate articles, we performed Web of Science[4] and Google Scholar[5] searches in May 2016 for the terms "forecast", "probabilistic forecast", "proper scoring rule", "CRPS" or "logarithmic score" along with either "Bayesian" or "MCMC". This approach left us with about 200 papers. Furthermore, we added papers that we knew from other sources (such as the websites of individual researchers).

Among these candidate articles, we considered published or pre-published articles in scientific journals or proceedings only. In particular, working papers and preprints were excluded from further consideration. We only retained studies where forecasts based on Bayesian MCMC methods are produced, and evaluated via proper scoring rules, and we restricted our attention to studies of real-valued linear variables, ignoring articles that deal with binary or categorical observations only. Furthermore, as we are interested in full probabilistic forecasts based on MCMC output, we disregarded articles where only functionals of the forecast distributions, such as means or medians, are evaluated, and we retained those studies only where the computation of the scores is documented in sufficient detail. Finally, Table 1 lists only papers that employ the CRPS or the logarithmic score. Very few studies have used scoring rules other than these: Riebler et al. (2012) use the Dawid–Sebastiani score, and Smith and Vahey (2016) and Tran et al. (2016) use weighted versions of the logarithmic score and the CRPS, as described by Diks et al. (2011) and Gneiting and Ranjan (2011), respectively.

Retaining only articles that meet the above selection criteria leaves us with the studies listed in Table 1, with multiple listings indicating that a study employs several choices.

B Computing the CRPS for mixtures of Gaussians

Here we discuss the computation of the CRPS in (7) when the predictive distribution is an equally weighted mixture of normal distributions, say $F = \hat{F}_m^{\mathrm{MP}}$, where $F_c(\cdot|\theta_i)$ is Gaussian with mean μ_i and variance σ_i^2. Grimit et al. (2006) note that in this case (8) can be written as

$$\mathrm{CRPS}\left(\hat{F}_m^{\mathrm{MP}}, y\right) = \frac{1}{m}\sum_{i=1}^{m} A(y-\mu_m, \sigma_m^2) - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} A(\mu_i-\mu_j, \sigma_i^2+\sigma_j^2), \qquad (19)$$

where $A(\mu, \sigma^2) = 2\sigma\phi(\frac{\mu}{\sigma}) + \mu(2\Phi(\frac{\mu}{\sigma}) - 1)$, with ϕ and Φ denoting the standard normal density and CDF, respectively. The scoringRules software package (Jordan et al., 2016) contains R/C++ code for the evaluation of (19), which requires $\mathcal{O}(m^2)$ operations.

A potentially much faster, but not exact, alternative is to evaluate the integral in (7) numerically.[6] Here we provide some evidence on the viability of this strategy, which we implement via the R function integrate, with arguments rel.tol and abs.tol of integrate set to 10^{-6}. As a first experiment, we use numerical integration to re-compute the CRPS scores of the mixture-of-parameters estimator in our data example for the first quarter of 2011. Figure 7 summarizes the results for 16 parallel chains. The left panel shows that the approximate scores are visually identical to the exact ones across all sample

[4] http://webofscience.com

[5] https://scholar.google.com

[6] Numerical integration could also be based on another representation of the CRPS that has recently been derived by Taillardat et al. (2016, p. 2390, bottom right).

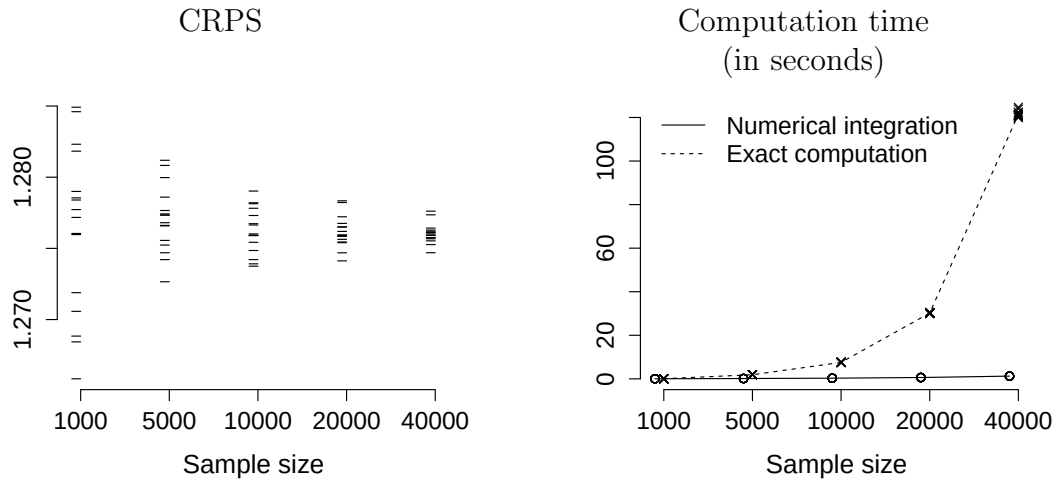

Figure 7: CRPS scores for the first quarter of 2011 in the data example, for 16 parallel chains and various MCMC sample sizes. Left: The segments connect the CRPS value obtained using numerical integration (left node) to the score obtained using the exact formula (right node). Right: Computation times in seconds, for numerical integration (dots; solid line) and exact formula (crosses; dashed line).

sizes and chains. Indeed, the maximal absolute error incurred by numerical integration is 8.0×10^{-8}. The approximation errors are dwarfed by the natural variation of the scores across MCMC chains. The right panel compares the computation time for exact evaluation vs. numerical integration. The latter is much faster, especially for large samples. For a sample of size 40,000 numerical integration requires less than 1.5 seconds, whereas exact evaluation requires about two minutes on an Intel i7 processor.

To obtain broad-based evidence, we next compare exact evaluation vs. numerical integration for all 74 forecast dates, from the second quarter of 1996 to the third quarter of 2014, employing 16 parallel chains for each date. We focus on the two largest MCMC sample sizes, 20 000 and 40 000, and find that across all 2 368 instances (74 dates times 2 sample sizes times 16 chains), the absolute difference of the two CRPS values never exceeds 6.3×10^{-7}. Therefore, we feel that numerical integration allows for the efficient evaluation of the CRPS for mixtures of normal distributions. The differences to the exact values are practically irrelevant and well in line with the error bounds in R's integrate function.

C Consistency of mixture-of-parameters approximations

Proof of Theorem 1

In the case of the CRPS, we prove the stronger result that $\int_{\mathbb{R}} |\hat{F}_m^{\mathrm{MP}}(z) - F_0(z)| \, \mathrm{d}z \to 0$ almost surely as $m \to \infty$. Putting $H(z) = 1 - F_0(z) + F_0(-z)$ and $\hat{H}_m(z) = 1 - \hat{F}_m^{\mathrm{MP}}(z) +$

$\hat{F}_m^{\mathrm{MP}}(-z)$ for $z \in \mathbb{R}$, we find that, for every fixed $N > 0$,

$$\limsup_{m\to\infty} \int_{\mathbb{R}} |\hat{F}_m^{\mathrm{MP}}(z) - F_0(z)| \, \mathrm{d}z \;\leq\; \limsup_{m\to\infty} \int_{-N}^{N} |\hat{F}_m^{\mathrm{MP}}(z) - F_0(z)| \, \mathrm{d}z \tag{20}$$

$$+ \int_N^\infty H(z) \, \mathrm{d}z \;+\; \limsup_{m\to\infty} \int_N^\infty \hat{H}_m(z) \, \mathrm{d}z. \tag{21}$$

The Ergodic Theorem implies that the right-hand side of (20) tends to zero, and that

$$\int_N^\infty \hat{H}_m(z) \, \mathrm{d}z = \int_N^\infty \frac{1}{m} \sum_{i=1}^{m} \left(1 - F_c(z|\theta_i) + F_c(-z|\theta_i)\right) \mathrm{d}z \longrightarrow \int_N^\infty H(z) \, \mathrm{d}z$$

almost surely as $m \to \infty$. In view of (20) and (21) we conclude that

$$\limsup_{m\to\infty} \int_{\mathbb{R}} |\hat{F}_m^{\mathrm{MP}}(z) - F_0(z)| \, \mathrm{d}z \leq 2 \int_N^\infty H(z) \, \mathrm{d}z \tag{22}$$

almost surely as $m \to \infty$. As the right-hand side of (22) decreases to zero as N grows without bounds, the proof of the claim is complete.

In the case of the DSS, let $K(z) = 1 - F_0(z) - F_0(-z)$ and $\hat{K}_m(z) = 1 - \hat{F}_m^{\mathrm{MP}}(z) - \hat{F}_m^{\mathrm{MP}}(-z)$ for $z \in \mathbb{R}$. Due to the finiteness of the first moments of F_0 and $\hat{F}_m^{\mathrm{MP}}$, $\int_{\mathbb{R}} z \, \mathrm{d}F_0(z) = \int_0^\infty K(z) \, \mathrm{d}z$ and $\int_{\mathbb{R}} z \, \mathrm{d}\hat{F}_m^{\mathrm{MP}}(z) = \int_0^\infty \hat{K}_m(z) \, \mathrm{d}z$. For the second moments, we find similarly that $\int_{\mathbb{R}} z^2 \, \mathrm{d}F_0(z) = 2\int_0^\infty zH(z) \, \mathrm{d}z$ and $\int_{\mathbb{R}} z^2 \, \mathrm{d}\hat{F}_m^{\mathrm{MP}}(z) = 2\int_0^\infty z\hat{H}_m(z) \, \mathrm{d}z$. Proceeding as before, the Ergodic Theorem implies almost sure convergence of the first and second moments, and thereby consistency relative to the DSS.

Proof of Theorem 2

By Lemma 2.1 in Chapter 4 of Kullback (1959),

$$\sup_{z\in\mathbb{R}} \left| 1 - \frac{\hat{f}_m^{\mathrm{MP}}(z)}{f_0(z)} \right| \longrightarrow 0$$

almost surely as $m \to \infty$ implies the desired convergence of the Kullback-Leibler divergence. Let P_m denote the empirical CDF of the parameter draws $(\theta_i)_{i=1}^m$. Under assumption (B) almost sure strong uniform consistency,

$$\sup_{z\in\Omega} \left| \hat{f}_m^{\mathrm{MP}}(z) - f_0(z) \right| = \sup_{z\in\Omega} \left| \int_\Theta f_c(z|\theta) \left[\mathrm{d}P_m(\theta) - \mathrm{d}P_{\mathrm{post}}(\theta) \right] \right| \longrightarrow 0$$

almost surely as $m \to \infty$, yields Kullback's condition. Finally, we establish almost sure strong uniform convergence under assumptions (A) and (B) by applying Theorem 19.4 and Example 19.8 of van der Vaart (2000).

D Consistency of empirical CDF-based approximations

Proof of Theorem 3

In the case of the CRPS, we proceed in analogy to the proof of Theorem 1 and demonstrate the stronger result that $\int_{\mathbb{R}} |\hat{F}_m^{\mathrm{ECDF}}(z) - F_0(z)| \, \mathrm{d}z \to 0$ almost surely as $m \to \infty$. Putting

Table 5: Prior parameters in the Markov switching model.

Symbol in Amisano and Giacomini (2007)	$\underline{\boldsymbol{\mu}}_{\boldsymbol{\delta}}$	$\underline{\mathbf{H}}_{\delta}^{-\mathbf{1}}$	$\underline{s}$	$\underline{\nu}$	$\mathbf{R}$
Parameter choice	$0_{[2,1]}$	$25 \times I_2$	0.3	3	$\begin{bmatrix} 8 & 2 \\ 2 & 8 \end{bmatrix}$
Relation to our Eq. (18)	Prior mean for $(\nu, \alpha)'$	Prior variance for $(\nu, \alpha)'$	Prior parameters for $\eta_{s_t}^2$		Dirichlet prior state transitions

$H(z) = 1 - F_0(z) + F_0(-z)$ and $\hat{H}_m(z) = 1 - \hat{F}_m^{\mathrm{ECDF}}(z) + \hat{F}_m^{\mathrm{ECDF}}(-z)$ for $z \in \mathbb{R}$, we see that, for every fixed $N > 0$,

$$\limsup_{m\to\infty} \int_{\mathbb{R}} |\hat{F}_m^{\mathrm{ECDF}}(z) - F_0(z)| \, \mathrm{d}z \;\le\; \limsup_{m\to\infty} \int_{-N}^{N} |\hat{F}_m^{\mathrm{ECDF}}(z) - F_0(z)| \, \mathrm{d}z \tag{23}$$

$$+ \int_{N}^{\infty} H(z) \, \mathrm{d}z \;+\; \limsup_{m\to\infty} \int_{N}^{\infty} \hat{H}_m(z) \, \mathrm{d}z. \tag{24}$$

The Generalized Glivenko-Cantelli Theorem (Dehling and Philipp, 2002, Theorem 1.1) implies that the right-hand side of (23) tends to zero almost surely as $m \to \infty$. If Z_0 has distribution F_0, then $\int_N^\infty H(z) \, \mathrm{d}z = \mathbb{E}(|Z_0| - N)_+$, where $(W)_+ = \max(W, 0)$ denotes the positive part of W. Furthermore, by the Ergodic Theorem

$$\int_{N}^{\infty} \hat{H}_m(z) \, \mathrm{d}z = \frac{1}{m} \sum_{i=1}^{m} (|X_i| - N)_+ \longrightarrow \mathbb{E}(|Z_0| - N)_+$$

almost surely as $m \to \infty$, which along with (23) and (24) implies that

$$\limsup_{m\to\infty} \int_{\mathbb{R}} |\hat{F}_m^{\mathrm{ECDF}}(z) - F_0(z)| \, \mathrm{d}z \le 2 \, \mathbb{E}(|Z_0| - N)_+ \tag{25}$$

almost surely as $m \to \infty$. As the right-hand side of (25) gets arbitrarily close to zero as N grows without bounds, the proof of the claim is complete.

In the case of the DSS, it suffices to note that the moments of the empirical CDF are the sample moments of $(X_i)_{i=1}^m$, and then to apply the Ergodic Theorem.

E Implementation details for the data example

Here we provide additional information on the Markov switching model for the quarterly U.S. GDP growth rate, Y_t. As described in equation (18) in Section 5, the model is given by $Y_t = \nu + \alpha Y_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, \eta_{s_t}^2)$, and $s_t \in \{1, 2\}$ is a discrete state variable that switches according to a Markov chain.

Our implementation follows Amisano and Giacomini (2007, Section 6.3), in that our prior distributions have the same functional forms but possibly different parameter choices, as summarized in Table 5. However, note that we use prior parameters for the residual variances in both latent states, whereas Amisano and Giacomini (2007) assume the residual variance to be constant across states.

Let $\beta = (\nu, \alpha)'$ denote the parameters for the conditional mean equation (18), $\overline{s}_t = (s_1, \ldots, s_t)'$ the sequence of latent states up to time t, $h = (\eta_1^{-2}, \eta_2^{-2})'$ the inverses of the state-dependent residual variances, and $\mathbf{P}$ the 2×2 transition matrix for the latent states. Our Gibbs sampler can then be sketched as follows:

- Draw $\beta \,|\, h, \overline{s}_t$ from a Gaussian posterior. The mean and variance derive from a generalized least squares problem, with observation t receiving weight $\eta_{s_t}^{-2}$.
- Draw $h \,|\, \beta, \overline{s}_t$ from a Gamma posterior. The Gamma distribution parameters for $\eta_s^{-2}, s \in \{1, 2\}$, are calculated from the observations t for which $s_t = s$. If necessary, permute the draws such that $\eta_1^2 > \eta_2^2$.
- Draw $\overline{s}_t \,|\, \beta, h, \mathbf{P}$ using the algorithm described by Greenberg (2013, pp. 194–195).
- Draw $\mathbf{P} \,|\, \overline{s}_t$ from a Dirichlet posterior.

Gianni Amisano kindly provides implementation details and Matlab code via his website (`https://sites.google.com/site/gianniamisanowebsite/home/teaching/istanbul-2014`, last accessed: August 15, 2016). An R implementation of his code is available within the R package scoringRules (Jordan et al., 2016), see `https://github.com/FK83/scoringRules/blob/master/KLTG2016_replication.pdf` for details.

Online supplement: Additional figures

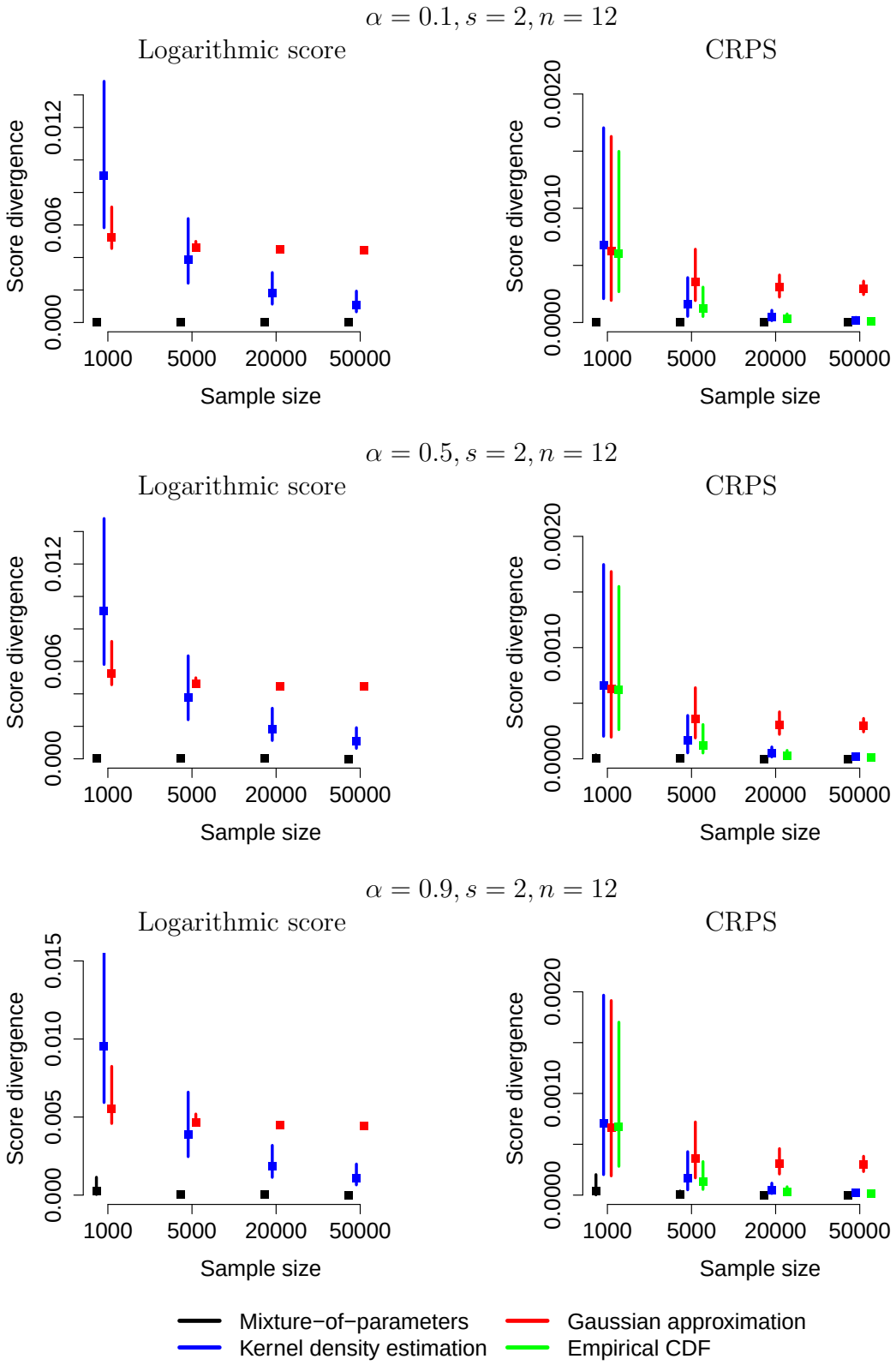

Figure 8: Same as Figure 1 in the paper, but for various parameter constellations of the data-generating process (cf. Table 4 in the paper).

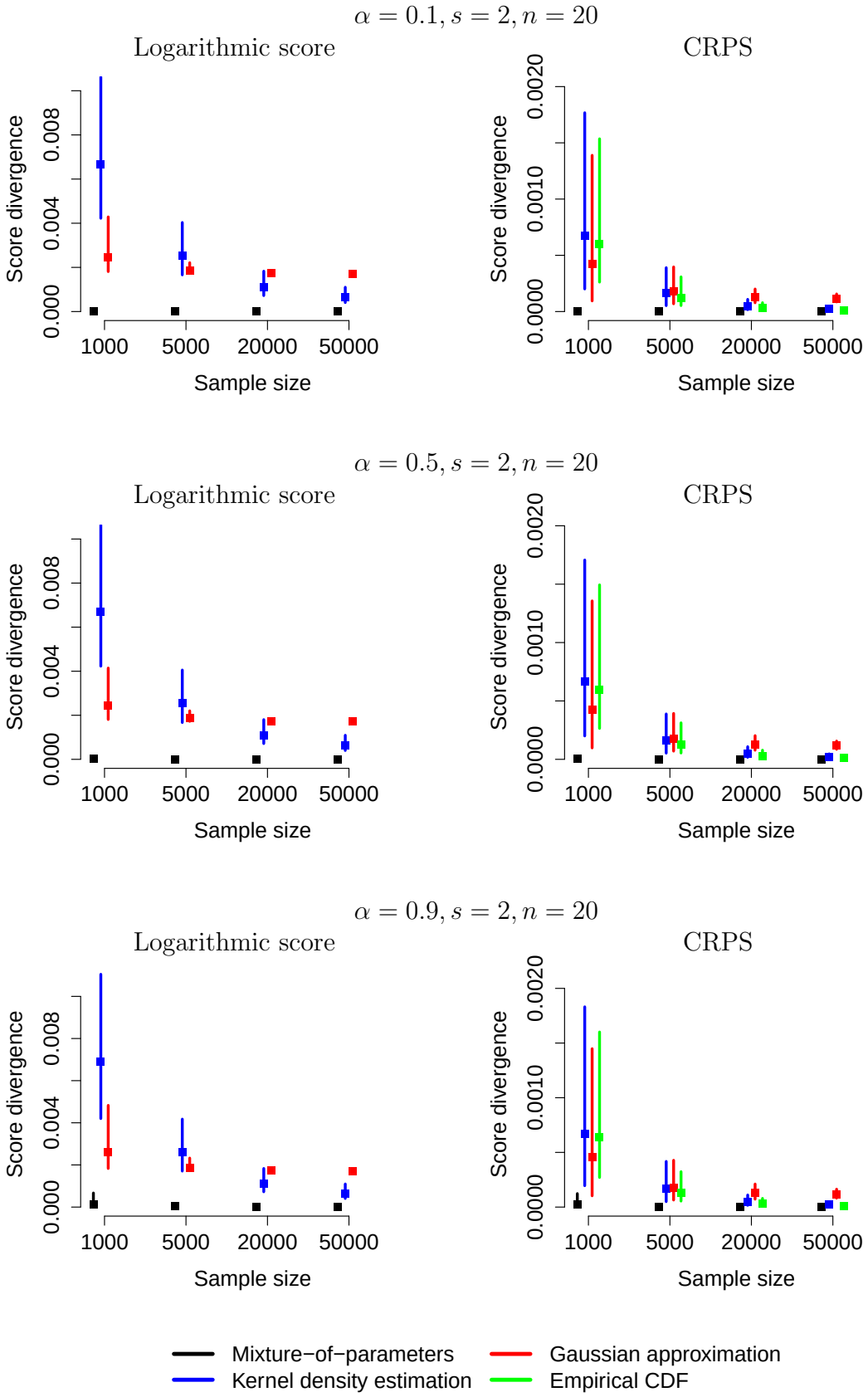

Figure 9: Same as Figure 1 in the paper, but for various other parameter constellations of the data-generating process (cf. Table 4 in the paper).

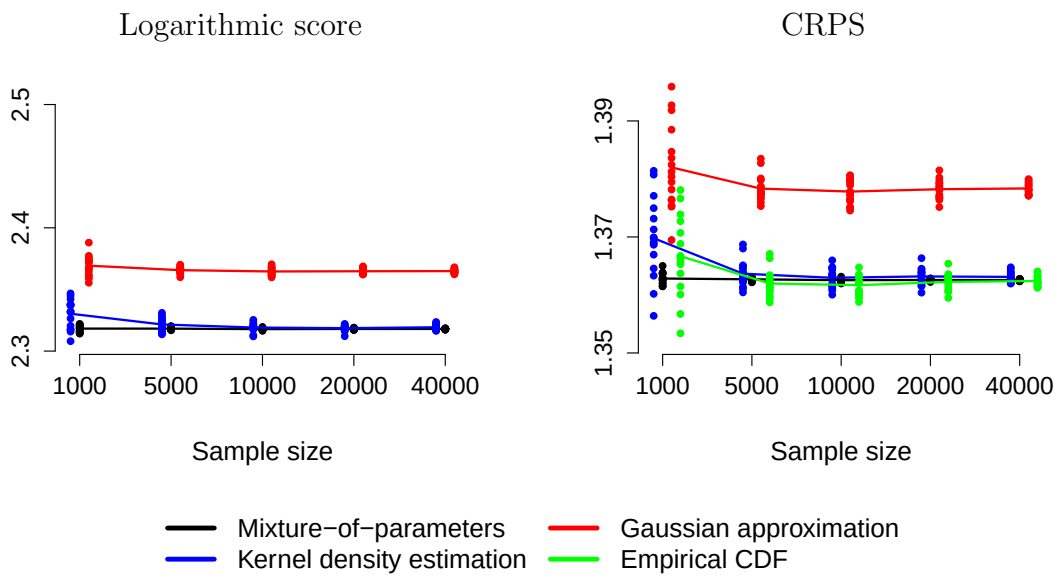

Figure 10: Same as Figure 6 in the paper, but at a prediction horizon of two quarters.

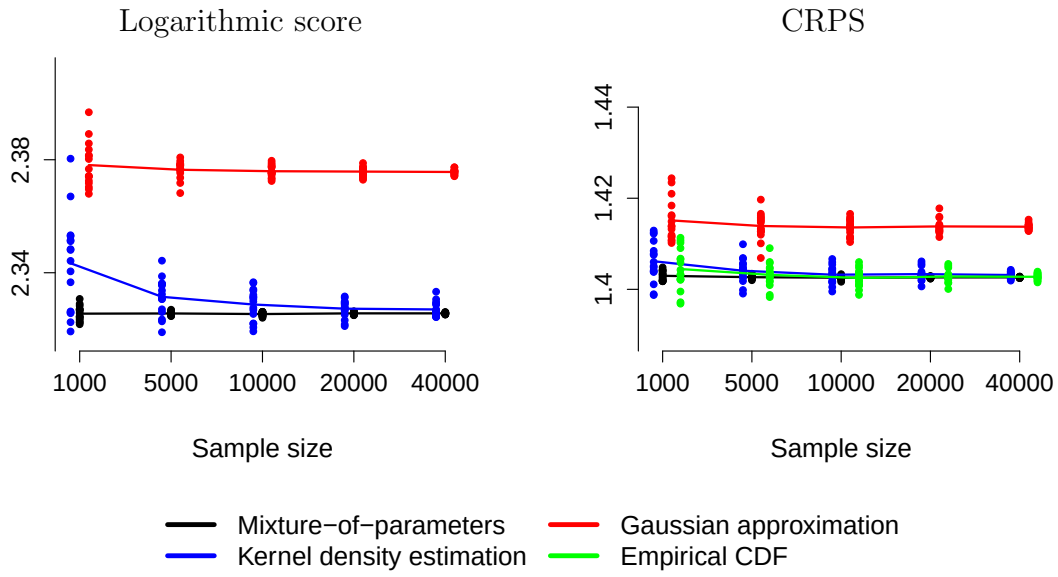

Figure 11: Same as Figure 6 in the paper, but at a prediction horizon of three quarters.