Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output

Fabian Krüger[1], Sebastian Lerch[2,3], Thordis L. Thorarinsdottir[4] and Tilmann Gneiting[2,3]

[1]Heidelberg University, Heidelberg, Germany; [2]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany; [3]Karlsruhe Institute of Technology, Karlsruhe, Germany; [4]Norwegian Computing Center, Oslo, Norway

1. Introduction

A rapidly growing transdisciplinary literature uses Bayesian inference to produce posterior predictive distributions. The posterior predictive CDF is of the generic form

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

where P_{post} is the posterior distribution of the parameter, θ, and $F_c(\cdot|\theta)$ is the conditional predictive CDF given $\theta \in \Theta$.
Frequently, the posterior predictive CDF must be approximated in some way, typically using some form of Markov chain Monte Carlo (MCMC); see, e.g., Gelfand and Smith (1990).
A generic MCMC algorithm designed to sample from F_0 can be sketched as follows.

- Fix $\theta_0 \in \Theta$ at some arbitrary value.
- For $i = 1, 2, \dots$ 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 $X_i \sim F_c(\cdot|\theta_i)$.

This generic MCMC algorithm allows for two 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 $(X_i)_{i=1}^m$.

We provide a systematic assessment of how to make and evaluate probabilistic forecasts based on such simulation output.

2. Proper scoring rules

Scoring rules are functions

$$S : \mathcal{F} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\},$$

where $\mathcal{F}$ denotes a class of probability distributions on $\mathbb{R}$. A scoring rule is called *proper* if

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

for all $F, G \in \mathcal{F}$ (Gneiting and Raftery, 2007). The *score divergence* associated with the scoring rule S is given by

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

Examples include

- the *logarithmic score*,

$$\text{LogS}(f, y) = -\log(f(y)), \tag{2}$$

  with $d_{\text{LogS}}(F, G) = \int g(z)\,\log\left(\frac{g(z)}{f(z)}\right)\,\mathrm{d}z$,
- the *continuous ranked probability score*

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \le z\})^2\,\mathrm{d}z \tag{3}$$

  with $d_{\text{CRPS}}(F, G) = \int (F(z) - G(z))^2\,\mathrm{d}z$.

3. Approximation methods

To compute LogS or CRPS for F_0, simulated samples $(\theta_i)_{i=1}^m$ and $(X_i)_{i=1}^m$ must be used to estimate F_0. The following approximation methods are frequently used in the literature.

Approximations based on $(\theta_i)_{i=1}^m$

- *Mixture-of-parameters estimator*: Approximate F_0 by

$$\hat{F}_m^{\text{MP}}(x) = \frac{1}{m}\sum_{i=1}^m F_c(x|\theta_i). \tag{4}$$

Approximations based on $(X_i)_{i=1}^m$

- *Gaussian approximation*

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

  where $\hat{\mu}_m$ and $\hat{\sigma}_m$ are the empirical mean and standard deviation of $(X_i)_{i=1}^m$.
- *Empirical CDF*: Estimate CDF F_0 by

$$\hat{F}_m^{\text{ECDF}}(x) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{x \ge X_i\}. \tag{6}$$

- *Kernel density estimation*: Estimate PDF f_0 by

$$\hat{f}_m^{\text{KD}}(x) = \frac{1}{mh_m}\sum_{i=1}^m K\left(\frac{x - X_i}{h_m}\right) \tag{7}$$

Occurrences in 53 recently published articles from economics, environmental sciences and other disciplines:

	LogS	CRPS
Mixture-of-parameters estimator	25	3
Kernel density estimation	6	1
Gaussian approximation	7	2
Empirical CDF	n/a	16

4. Theoretical consistency results

How to assess the adequacy of approximation methods from a theoretical perspective?

4.1 Consistency and score divergences

An approximation method is *consistent relative to scoring rule* S *at 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 $m \to \infty$.
Note that

- properties of $d_{\mathrm{S}}(\hat{F}_m, F_0)$ and required convergence of $\hat{F}_m$ to F_0 strongly depend on S,
- consistency is independent of forecast quality.

4.2 Consistency results

We investigate sufficient conditions for consistency of the aforementioned approximation methods. Assumptions:

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

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

Mixture-of-parameters approximation

Under assumption (A), the MP approximation is consistent relative to the CRPS.
Under assumptions (A) and (B), the MP approximation is consistent relative to the logarithmic score.

Gaussian approximation

Can only be consistent if F_0 is Gaussian – unlikely to hold in many applications.

Empirical CDF-based approximation

Under assumption (A), the empirical CDF technique is consistent relative to the CRPS.

Kernel density estimation

Requires stringent assumptions on mixing coefficients and bandwidth as tail properties of kernel K and f_0 need to be carefully matched (e.g., Hall, 1987).

5. Simulation study

Investigate approximation methods in a setup that emulates realistic MCMC behavior with dependent samples. Here, F_0 is known by construction, and we can compare the different approximations to the true forecast distribution.

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

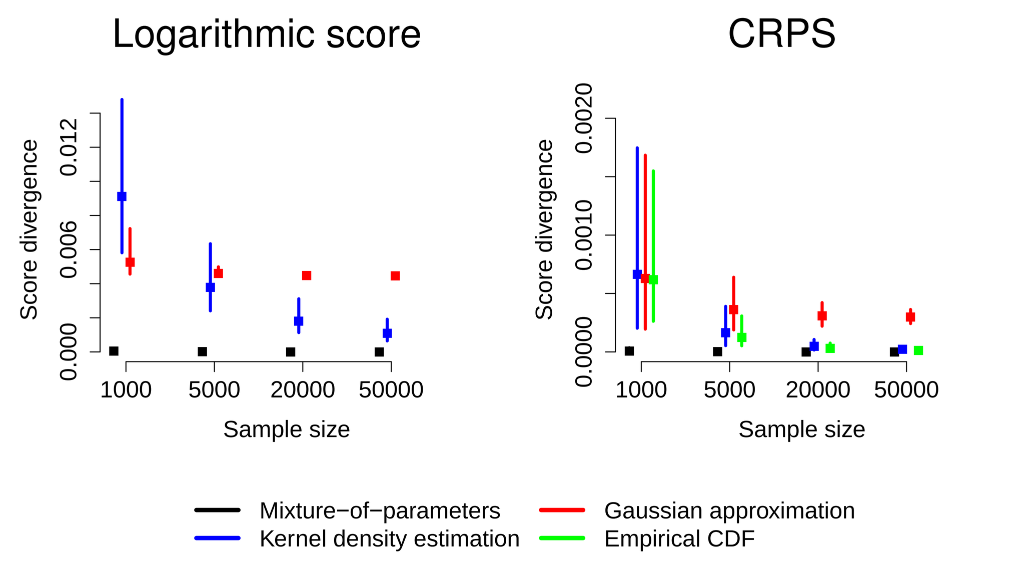
Data-generating process

Generate sequences $(\theta_i)_{i=1}^m$ and $(X_i)_{i=1}^m$ such that

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

is a compound Gaussian distribution.
To mimic a realistic MCMC scenario with dependent draws, we use the Fox and West (2011) model for θ^2 that implies autoregressive-type dependence, and an unconditional Student t distribution F_0.

Results



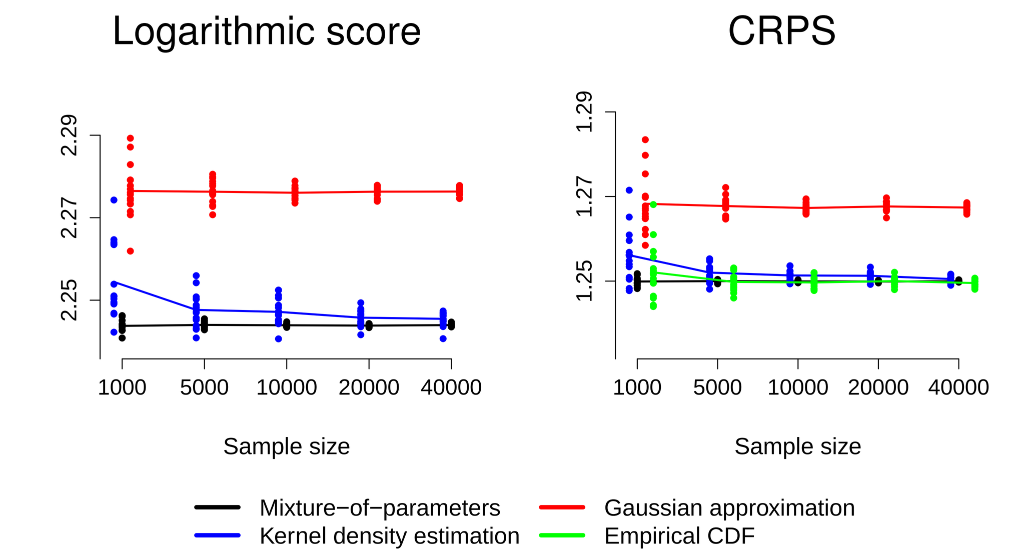
The MP estimator dominates the other methods by a wide margin with divergences very close to zero, and little variation across replicates.

6. Case study

Markov switching AR model for one quarter ahead forecasts of quarterly growth rates of U.S. GDP, 1996-2014,

$$Y_t = \nu + \alpha Y_{t-1} + \varepsilon_t, \tag{8}$$

where $\varepsilon_t \sim \mathcal{N}(0, \eta_{s_t}^2)$ and $s_t \in \{1, 2\}$ is a discrete state variable. Conditional on θ_i, the predictive distribution in (8) is Gaussian, but F_0 is not.
As F_0 is unknown, we are unable to compute $d_{\mathrm{S}}(\hat{F}_m, F_0)$. Instead, we compare predictive performance of approximation methods across multiple chains.



MP approximated scores display the smallest variation across chains for all sample sizes. KDE performs poorly for small sample sizes, and is dominated by the empirical CDF-based approximation in case of the CRPS.

7. Discussion

Theoretical and practical implications:

- We derive conditions for consistency of various approximation methods.
- CRPS requires less stringent regularity assumptions compared to LogS.
- MPE works best, KDE is problematic for LogS, Gaussian approximations are generally problematic.

All details are available in Krüger et al. (2016). Considerations presented here have been implemented in the `R` package `scoringRules` (Jordan et al., 2017) that provides functions to efficiently compute scoring rules for many parametric distributions, and forecasts given as simulated samples.

References

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`.

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.

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

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

Jordan, A., Krüger, F. and Lerch, S. (2017). *scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts*. R package version 0.9.3, URL `https://CRAN.R-project.org/package=scoringRules`.

Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. Working paper. Preprint available at `https://arxiv.org/abs/1608.06802`.

Heidelberg Institute for Theoretical Studies
HITS
Karlsruhe Institute of Technology
WAVES TO WEATHER