WORKING PAPER

Bayesian nonparametric calibration and combination of predictive distributions

NORGES BANK
RESEARCH

03 | 2015

AUTHORS:

FEDERICO BASSETTI
ROBERTO CASARIN
FRANCESCO RAVAZZOLO

NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:

servicesenter@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail

servicesenter@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8143 (online)
ISBN 978-82-7553-847-3 (online)

Bayesian Nonparametric Calibration and Combination of Predictive Distributions*

Federico Bassetti[§] Roberto Casarin[†] Francesco Ravazzolo[‡]

[§]University of Pavia
[†]University of Venice
[‡]Norges Bank and BI Norwegian Business School

February 25, 2015

Abstract

We introduce a Bayesian approach to predictive density calibration and combination that accounts for parameter uncertainty and model set incompleteness through the use of random calibration functionals and random combination weights. Building on the work of Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013), we use infinite beta mixtures for the calibration. The proposed Bayesian nonparametric approach takes advantage of the flexibility of Dirichlet process mixtures to achieve any continuous deformation of linearly combined predictive distributions. The inference procedure is based on Gibbs sampling

[*] This Working Paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. We are much indebted to Tilmann Gneiting for helpful discussions and for his contribution to an earlier version of this work. We thank Concepcion Ausin, Luc Bawuens, Sylvia Früwirth-Schnatter, James Mitchell, Jacek Osiewalski, Dimitris Korobilis, Gary Koop, Enrique ter Horst, Shaun Vahey, Ken Wallis, Michael Wipper, and Herman van Dijk for their constructive comments, and the conference and seminar participants at the: CORE Louvain University, Cracow Polish Science Academy seminar series, the 2014 "Italian Statistical Society" meeting at University of Cagliari, the 2014 "Econometric Modelling and Forecasting in Central Banks" workshop at University of Glasgow, Norges Bank, the 2013 "Economic Modelling and Forecasting Group" workshop at University of Warwick, the "8th International Conference on Computational and Financial Econometrics", Pisa, Universidad Carlos III de Madrid, University Ca' Foscari of Venice, Wien University. Casarin‘s research is supported by funding from the European Union, Seventh Framework Programme FP7/2007-2013 under grant agreement SYRTO-SSH-2012-320270, by the Institut Europlace of Finance, "Systemic Risk grant", the Global Risk Institute in Financial Services, the Louis Bachelier Institute, "Systemic Risk Research Initiative", and by the Italian Ministry of Education, University and Research (MIUR) PRIN 2010-11 grant MISURA. Bassetti's research is supported by the INdAm-GNAMPA Project 2014.

and allows accounting for uncertainty in the number of mixture components, mixture weights, and calibration parameters. The weak posterior consistency of the Bayesian nonparametric calibration is provided under suitable conditions for unknown true density. We study the methodology in simulation examples with fat tails and multimodal densities and apply it to density forecasts of daily S&P returns and daily maximum wind speed at the Frankfurt airport.

AMS 2000 subject classifications: Primary 62; secondary 91B06.

JEL codes: C13, C14, C51, C53.

Keywords: Forecast calibration, Forecast combination, Density forecast, Beta mixtures, Bayesian nonparametrics, Slice sampling.

1 Introduction

Combining forecasts from different statistical models or other sources of information is a crucial problem in many important applications. A wealth of papers have addressed this issue with Bates and Granger (1969) being one of the first attempts in this field. The initial focus of the literature was on defining and estimating combination weights for point forecasts. For instance, Granger and Ramanathan (1984) propose to combine point forecasts with unrestricted least squares regression coefficients as weights. The ubiquitous Bayesian model averaging technique relies on weighted averages of posterior distributions from different models and implies linearly combined posterior means (Hoeting et al., 1999). Recently, probabilistic forecasts in the form of predictive probability distributions have become prevalent in various fields, including macro economics with routine publications of fancharts from central banks, finance with asset allocation strategies based on higher-order moments, and meteorology with operational ensemble forecasts of future weather (Tay and Wallis, 2000; Gneiting and Katzfuss, 2014).

Therefore, research interest has shifted to the construction of combinations of predictive distributions, which poses new challenges (Gneiting and Ranjan, 2013). A prominent, critically important issue is that predictive distributions ought to be calibrated (Dawid, 1984; Kling and Bessler, 1989; Diebold et al., 1998; Gneiting et al., 2007; Mitchell and Wallis, 2011). Moreover, the traditional linear pool (Stone, 1961; Hall and Mitchell, 2007) has been generalized to nonlinear aggregation schemes (Fawcett et al., 2013; Gneiting and Ranjan, 2013), and time-varying approaches can account for time instabilities and estimation uncertainty in the combination weights (Billio et al., 2013).

In this paper, we propose a flexible Bayesian nonparametric approach to calibration and combination that relies on beta mixtures, and nests the beta transformed linear pool introduced by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013). We develop tools for Bayesian inference for both cases of known and unkown number of mixture components. In the case the number of component is not known we assume an infinite mixture representation and a Dirichlet process prior (Ferguson, 1973; Lo, 1984; Sethuraman, 1994). This type of prior and its multivariate extensions (e.g., see Müller et al. (2004), Griffin and Steel (2006), Hatjispyros et al. (2011)), is now widely used due to the availability of efficient algorithms for posterior computations (Escobar and West, 1995; MacEachern and Müller, 1998; Papaspiliopoulos and Roberts, 2008; Taddy, 2010), including but not limited to applications in time series settings (Hirano, 2002; Chib and Hamilton, 2002; Rodriguez and ter Horst, 2008; Taddy and Kottas, 2009; Jensen and Maheu, 2010; Griffin, 2011; Griffin and Steel, 2011; Burda et al., 2014; Bassetti et al., 2014; Wiesenfarth et al., 2014; Jochmann, 2015). A recent account of Bayesian non-parametric inference can be found in Hjort et al. (2010). In this paper we develop a slice sampling approach that builds on the work of Walker (2007) and Kalli et al. (2011).

Also, we contribute to the recent literature on posterior consistency of Bayesian nonparametric inference in econometrics, for example, see the recent studies of Norets and Pelenis (2012), Pati et al. (2013), Pelenis (2014), Norets and Pelenis (2015). In this paper we focus on the posterior consistency of the nonparametric estimates of the calibration function and of the linear combination of densities. We build on Wu and Ghosal (2009a,b) and provide weak consistency under general conditions on the combined densities and under both model set completeness and incompleteness assumptions.

The remainder of the paper is organized as follows. Section 2 introduces our beta mixture calibration and combination model and places it in the context of the general density combination approach introduced by Fawcett et al. (2013). This is followed by Section 3, where we propose Bayesian inference based on slice and Gibbs sampling methods. Section 4 provides posterior consistency of the Bayesian nonparametric calibration and combination in the weak sense under suitable conditions for unknown true density and under the assumption of incomplete model set. In Section 5 we illustrate the effectiveness of our approach on simulation examples. Section 6 provides case studies including some well-studied datasets in weather forecast and finance and see major improvements in the predictive performance for daily stock returns and daily maximum wind speed. The paper closes with a discussion in Section 7.

2 Beta mixture calibration and combination

Let $F_1, \ldots, F_M$ be a set of predictive cumulative distribution functions (CDFs) for a real-valued variable of interest, y, which might be based on distinct statistical models or experts. Following Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013), we consider combination formulas that map the M-tuple $(F_1, \ldots, F_M)$ into a single, aggregated predictive CDF, F. Let

$$\Delta_M = \left\{ \boldsymbol{\omega} = (\omega_1, \ldots, \omega_M) \in [0,1]^M : \sum_{m=1}^{M} \omega_m = 1 \right\}$$

denote the unit simplex in $\mathbb{R}^M$. The beta transformed linear pool introduced by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) considers combination formulas of the form

$$F(y|\boldsymbol{\theta}) = B_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y) \right) \tag{1}$$

for $y \in \mathbb{R}$, where $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\omega})$, $B_{\alpha,\beta}$ denotes the CDF of the beta distribution with parameters $\alpha > 0$ and $\beta > 0$ and density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on the unit interval. We interpret $B_{\alpha,\beta}$ as a parametric calibration function, which acts on a linear combination of $F_1, \ldots, F_M$ with mixture weights $\boldsymbol{\omega} \in \Delta_M$. In the particular case in which $\alpha = 1$ and $\beta = 1$, the calibration function is the identity function, and the beta transformed linear pool reduces to the traditional linear pool. If $F_1, \ldots, F_M$ admit Lebesgue densities $f_1, \ldots, f_M$, respectively, the combination formula (1) can be written equivalently in terms of the aggregated probability density function (PDF), namely

$$f(y|\boldsymbol{\theta}) = \left(\sum_{m=1}^{M} \omega_m f_m(y) \right) b_{\alpha,\beta}\left(\sum_{m=1}^{M} \omega_m F_m(y) \right) \tag{2}$$

for $y \in \mathbb{R}$, where $b_{\alpha,\beta}$ is the PDF of the beta distribution. In the case $M = 1$ of a single predictive distribution, the transformation serves to achieve calibration; when $M = 2$, we seek to combine and calibrate simultaneously. The linear combination weights assign relative importance to the individual predictive distributions, and the beta transformed linear pool admits exchangeable flexible dispersivity in a certain well defined sense (Gneiting and Ranjan, 2013). However, the approach allows for a rather limited, parametric class of calibration functions only.

In this paper we extend the approach and propose the use of mixtures of beta calibration and combination models. We generalize (1) and (2) to

$$F(y|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \, B_{\alpha_k, \beta_k} \left(\sum_{m=1}^{M} \omega_{km} F_m(y) \right) \tag{3}$$

and

$$f(y|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \left(\sum_{m=1}^{M} \omega_{km} f_m(y) \right) b_{\alpha_k, \beta_k} \left(\sum_{m=1}^{M} \omega_{km} F_m(y) \right) \tag{4}$$

for $y \in \mathbb{R}$, where $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega})$, the vector $\boldsymbol{w} = (w_1, \ldots, w_K) \in \Delta_K$ comprises the beta mixture weights, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ are beta calibration parameters, and $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M)$, with $\boldsymbol{\omega}_1 = (\omega_{11}, \ldots, \omega_{1M}), \ldots, \boldsymbol{\omega}_K = (\omega_{K1}, \ldots, \omega_{KM}) \in \Delta_M$ the component specific sets of linear combination weights.

It is well known that any continuous function g on the unit interval can be approximated by a beta mixture. Specifically, if we let $w_{k,K} = \int_{(k-1)/K}^{k/K} g(x) \, \mathrm{d}x$ for for $K = 1, 2, \ldots$ and $k = 1, \ldots, K$, then

$$\lim_{K \to \infty} \left(\sup_{y \in [0,1]} \left| \sum_{k=1}^{K} w_{k,K} \, b_{k,\, K-k+1}(y) - g(y) \right| \right) = 0.$$

This result illustrates the flexibility of the beta mixture approach and raises the possibility of parsimonious representations, where we assume that $\omega_{1m} = \cdots = \omega_{Km} = \omega_m$ for $m = 1, \ldots, M$ and $\alpha_k = k$ and $\beta_k = K - k + 1$ for $k = 1, \ldots, K$. When $K < \infty$ we refer to the general beta mixture model in (3) and (4) as the BM_K model, which is much more flexible, and nests the beta transformed linear pool proposed by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) that arises in the special case in which $K = 1$. Bayesian inference can provide guidance in choosing appropriate compromises between parsimony and flexibility, especially when K is unknown. In particular, our Bayesian approach allows us to treat the parameter K as unbounded and random. We refer to this latter setting as the infinite beta mixture or BM_∞ calibration, for which we give details in the following section.

The beta mixture calibration and combination model can also be interpreted in terms of generalized linear pool, introduced by Fawcett et al. (2013). Specifically, we can write the aggregated PDF (4) as

$$f(y|\boldsymbol{\theta}) = \sum_{m=1}^{M} \tilde{\omega}_m(y)\, f_m(y)$$

for $y \in \mathbb{R}$, where the generalized weight functions are given by

$$\tilde{\omega}_m(y) = \sum_{k=1}^{K} \omega_{km} w_k \, b_{\alpha_k,\,\beta_k}\left(\sum_{m=1}^{M} \omega_{km} F_m(y) \right)$$

for $m = 1, \ldots, M$. We should notice that this simple result provides an alternative interpretation of the generalized combination model in Fawcett et al. (2013) as a calibration and combination model. One of the major differences with respect to Fawcett et al. (2013) is that they use weight functions that are piecewise constant, whereas the weight functions implied by the beta mixture model are continuous.

For inference on our model we use a flexible Bayesian approach, which we describe in the following section.

3 Bayesian inference

In Bayesian settings, it is convenient to express the standard beta distribution with parameters $\alpha > 0$ and $\beta > 0$ and density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ in terms of its mean $\mu = \alpha/(\alpha+\beta)$ and the parameter $\nu = \alpha + \beta > 0$ (Epstein, 1966; Robert and Rousseau, 2002; Billio and Casarin, 2011; Casarin et al., 2012). We refer to the reparameterized PDF as

$$b^*_{\mu,\nu}(x) = \frac{\Gamma(\nu)}{\Gamma(\mu\nu)\Gamma((1-\mu)\nu)}\, x^{\mu\nu-1}(1-x)^{(1-\mu)\nu-1}\, \mathbb{1}_{[0,1]}(x),$$

where Γ denotes the gamma function, and we use the symbol $B^*_{\mu,\nu}$ to denote the corresponding CDF.

We discuss inference in the time series setting at the unit prediction horizon, where the training data comprise the predictive CDFs $F_{1t}, \ldots, F_{Mt}$, which are conditional on information available at time $t-1$, along with the respective realization, y_t, at time $t = 1, \ldots, T$, respectively. We then wish to estimate a calibration and combination formula of the form (3) that maps the tuple $F_{1t}, \ldots, F_{Mt}$ into an aggregated CDF, F_t. In practice, we use the estimated calibration and combination formula to aggregate the predictive CDFs $F_{1,T+1}, \ldots, F_{M,T+1}$, which are based on information available at time T, into a single predictive CDF, F_{T+1}, for the subsequent value, y_{T+1}, of the variable of interest. Extensions to multi-step ahead forecasts is possible, and we leave this for further research.

To ease the notational burden in the time series setting, let $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kM}) \in \Delta_M$, and write

$$H_t(y_t|\boldsymbol{\omega}_k) = \sum_{m=1}^{M} \omega_{km}\, F_{mt}(y_t) \tag{5}$$

and

$$h_t(y_t|\boldsymbol{\omega}_k) = \sum_{m=1}^{M} \omega_{km}\, f_{mt}(y_t) \tag{6}$$

for $t = 1, \ldots, T$ and $k = 1, 2, \ldots, K$, respectively.

3.1 Bayesian finite beta mixture model

We work with a reparameterized version of the finite beta mixture calibration and combination model (i.e., $K < \infty$), in which the aggregated CDF and PDF can be represented as

$$F_t(y_t|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, B^*_{\mu_k,\nu_k}(H_t(y_t|\boldsymbol{\omega}_k)) \tag{7}$$

and

$$f_t(y_t|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, h(y_t|\boldsymbol{\omega}_k) b^*_{\mu_k,\nu_k}(H_t(y_t|\boldsymbol{\omega}_k)) \tag{8}$$

for $t = 1, \ldots, T$. The parameter vector for the BM$_K$ model can then be written as $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\omega})$, where $\boldsymbol{w} = (w_1, \ldots, w_K) \in \Delta_K$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in (0,1)^K$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K) \in (0,\infty)^K$ and $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K) \in \Delta_M^K$, with K being a fixed positive integer. The parameter space is defined as $\Theta = \Delta_K \times (0,1)^K \times (0,\infty)^K \times \Delta_M^K$.

Our Bayesian approach assumes that

$$\boldsymbol{w} \sim \mathcal{D}ir(\xi_{w1}, \ldots, \xi_{wM}) \tag{9}$$

and

$$\begin{aligned} \mu_k &\sim \mathcal{B}e(\xi_{\mu 1}, \xi_{\mu 2}), & (10)\\ \nu_k &\sim \mathcal{G}a(\xi_{\nu 1}, \xi_{\nu 2}), & (11)\\ \boldsymbol{\omega}_k &\sim \mathcal{D}ir(\xi_{\omega 1}, \ldots, \xi_{\omega M}) & (12) \end{aligned}$$

for $k = 1, \ldots, K$, where $\mathcal{B}e(\alpha, \beta)$ is a Beta distribution with density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ for $x \in \Delta_1$, $\mathcal{G}a(\gamma, \delta)$ is a Gamma distribution with density proportional to $x^{\gamma} \exp\{-\delta x\}$ for $x > 0$, and $\mathcal{D}ir(\xi_1, \ldots, \xi_M)$ is a Dirichlet distribution with density proportional to $\prod_{m=1}^{M} w_m^{\xi_m - 1}$ for $(w_1, \ldots, w_M) \in \Delta_M$, with all these distributions being independent. Guided by symmetry arguments in the Beta and Dirchlet case, and using a standard, uninformative prior in the Gamma case (Spiegelhalter et al., 2004)¡, we parameterize parsimoniously and set $\xi_{w1} = \cdots = \xi_{wM}$, $\xi_{\mu} = \xi_{\mu 1} = \xi_{\mu 2}$, $\xi_{\nu 1} = \xi_{\nu 2}$, and $\xi_{\omega 1} = \cdots = \xi_{\omega M}$. In what follows, we refer to the common hyperparameter values as ξ_w, ξ_μ, ξ_ν, and ξ_ω, respectively

Adopting a data augmentation framework (Frühwirth-Schnatter, 2006), we introduce the allocation variables $d_{kt} \in \{0, 1\}$, where $k = 1, \ldots, K$ and $t = 1, \ldots, T$. The likelihood of the BM_K calibration model is the marginal of the complete data likelihood

$$L(Y, D \,|\, \boldsymbol{\theta}) = \prod_{t=1}^{T} \prod_{k=1}^{K} \left(w_k \, h_t(y_t | \boldsymbol{\omega}_k) \, b^*_{\mu_k, \nu_k} \left(H_t(y_t | \boldsymbol{\omega}_k) \right) \right)^{d_{kt}},$$

where we let $Y = (y_1, \ldots, y_T)$ and $D = (d_{11}, \ldots, d_{K1}, \ldots, d_{1T}, \ldots, d_{KT})$. The implied joint posterior of D and $\boldsymbol{\theta}$ given the observations Y satisfies

$$\pi(D, \boldsymbol{\theta} \,|\, Y) \propto g(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\omega}) \prod_{k=1}^{K} w_k^{\xi_w + T_k - 1} \prod_{t \in \mathcal{D}_k} h_t(y_t | \boldsymbol{\omega}_k) \, b^*_{\mu_k, \nu_k}(H_t(y_t | \boldsymbol{\omega}_k)),$$

where $g(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\omega})$ is the prior density, $\mathcal{D}_k = \{t = 1, \ldots, T \,|\, d_{kt} = 1\}$, and T_k is the number of elements in $\mathcal{D}_k$. To sample from the joint posterior, we use a Gibbs sampler that draws iteratively from $\pi(D \,|\, \boldsymbol{\theta}, Y)$, $\pi(\boldsymbol{\mu}, \boldsymbol{\nu} \,|\, \boldsymbol{w}, \boldsymbol{\omega}, D, Y)$, $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, D, Y)$, and $\pi(\boldsymbol{w} \,|\, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\nu}, D, Y)$, respectively, for which we give details in Appendix A.1.

The output of the algorithm is a sample $\boldsymbol{\theta}^{(i)} = (\boldsymbol{w}^{(i)}, \boldsymbol{\mu}^{(i)}, \boldsymbol{\nu}^{(i)}, \boldsymbol{\omega}^{(i)})$ for $i = 1, \ldots, I$, where I is the number of iterations in the Gibbs sampler. The sample is used to approximate with $\hat{F}_{T+1}(y_{T+1})$ the desired one-step-ahead cumulative posterior predictive distribution $F_{T+1}(y_{T+1}) = \int_{\Theta} F_{T+1}(y_{T+1} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | Y) d\boldsymbol{\theta}$, where $\pi(\boldsymbol{\theta} | Y)$ is the marginal distribution of $\pi(D, \boldsymbol{\theta} | Y)$. In the special case when $K = 1$ we get

$$\hat{F}_{T+1}(y_{T+1}) = \frac{1}{I} \sum_{i=1}^{I} B^*_{\mu^{(i)}, \nu^{(i)}} \left(\sum_{m=1}^{M} \omega_m^{(i)} \, F_{m,T+1}(y_{T+1}) \right), \tag{13}$$

which can be thought of as a Bayesian implementation of the beta transformed linear pool (1) of Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013). An advantage of the proposed approach based on Gibbs approximation is that parameter uncertainty can be take into consideration in the prediction. A plug-in approximation of the predictive, which does not account for the parameter uncertainty, can be used, namely $\hat{F}_{T+1}(y_{T+1}) = F_{T+1}(y_{T+1}|\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the parameter posterior mean which can be approximated by the empirical average of $\boldsymbol{\theta}^{(i)}$ $i = 1, \ldots, I$. Another advantage of our approach is that credible intervals for the calibrated predictive CDF can be easily approximated by using the output of the Gibbs sampler.

3.2 Bayesian infinite beta mixture model

In the finite-mixture beta calibration and combination model the number of the beta densities is given, and model selection procedures can be used to choose the number of mixture components. As evidenced in previous studies (see Billio et al. (2013) and Fawcett et al. (2013)), in a time series context the model pooling scheme can be subject to time instability, thus as a new group of observations arrives the pooling scheme can change dramatically. Geweke (2010) discusses how standard weights converge to select one model (or a subset of models), therefore not properly coping with such instability. For these reasons, one would like to start with an infinite prior number of calibration functions and local pooling schemes, only a finite number of which are selected on a given finite sample. The consequence is that the number K of beta mixture components can vary and increase with the sample size. One of the side benefits of the model with infinite calibration components is that it provides an answer to the problem of selecting the number of components in the finite mixture approach.

We propose here a Bayesian non-parametric models which allows for estimating the number of components and also for including the model uncertainty in the posterior predictive. We refer to this model as the infinite-mixture calibration model BM_∞. Let us assume

$$f_t(y_t|\boldsymbol{\theta}) = b^*_{\mu,\nu}\left(H_t(y_t|\boldsymbol{\omega})\right) h_t(y_t|\boldsymbol{\omega}),$$

where $\boldsymbol{\theta} = (\mu, \nu, \boldsymbol{\omega})$, with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$. Our prior for the BM_∞ parameters $\boldsymbol{\theta}$ is nonparametric, i.e. $\boldsymbol{\theta} \sim G(\boldsymbol{\theta})$ where G is a random probability measure

$$G \sim DP(\psi, G_0)$$

and $DP(\psi, G_0)$ denotes a Dirichlet process (DP) (Ferguson (1973)) with concentration parameter ψ and base measure G_0. Following the standard result of Sethuraman (1994), the Dirichlet process prior can be represented as

$$G(d\boldsymbol{\theta}) = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\theta}_k}(d\boldsymbol{\theta})$$

with random weights w_k generated by the stick-breaking construction

$$w_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$$

where the stick-breaking components v_l are i.i.d. random variables from $\mathcal{B}e(1, \varphi)$. The atoms θ_k are i.i.d. random variables from the base measure G_0. In our model the base measure is given by the product of the following distribution

$$\mathcal{B}e(\xi_\mu, \xi_\mu)\mathcal{G}a(\xi_\nu/2, \xi_\nu/2)\mathcal{D}ir(\xi_\omega, \ldots, \xi_\omega).$$

The Dirichlet process prior assumption and the stick-breaking representation of the DP allow us to write the combination and calibration model in terms of infinite mixtures of random beta distributions with the following random pdf

$$\begin{aligned} f_t(y_t|G) &= \int f_t(y_t|\boldsymbol{\theta})G(d\boldsymbol{\theta}) \\ &= \sum_{k=1}^{\infty} w_k b^*_{\mu_k,\nu_k}\left(H_t(y_t|\boldsymbol{\omega}_k)\right) h_t(y_t|\boldsymbol{\omega}_k). \end{aligned}$$

The number of components sampled in the first T observations is random and its prior distribution is (Antoniak (1974))

$$P(K = k|\psi, T) = \frac{T!\Gamma(\psi)}{\Gamma(\psi + T)}|s_{Tk}|\psi^k$$

for $k = 1, 2, \ldots$, where s_{Tk} is the signed Stirling number (Abramowitz and Stegun, 1972, p. 824). The dispersion hyper-parameter $\psi > 0$ is driving the prior expected number of parameters. Large values of ψ increase the probability of introducing new components in the mixture. As the prior dispersion depends crucially on this parameter, the results of the posterior inference on the infinite mixture model are usually presented for different values of ψ. It also possible to extend the nonparametric models by assuming a further stage of the prior hierarchical structure and assuming a prior for ψ. A common choice for the prior is a gamma distribution, $\mathcal{G}a(c, d)$ (see Escobar and West (1995)). The second important feature is that our inference approach provides, as a natural product, the posterior distribution of the number of components given a sample of data and allows for the inclusion of the number of components uncertainty in the predictive density.

Inference on infinite mixture models resulting from a Dirichlet prior assumption requires the use of simulation methods. Gibbs samplers have been proposed in Escobar (1994) and Ishwaran and James (2001), which make use of the Polya-urn representation of the Dirichlet process. Ishwaran and Zarepour (2000) proposed a sampler based on a truncation of the infinite mixture representation. Papaspiliopoulos and Roberts (2008) proposed an exact simulation algorithm based on retrospective sampling. In this paper we apply the slice sampling algorithm proposed in Walker (2007) and Kalli et al. (2011). The algorithm uses a set of auxiliary variables to deal with the infiniteness problem of the mixture model. More specifically, let us introduce a sequence of slice sampling variables u_t, $t = 1, 2, \ldots, T$, then $f_t(y_t|G)$ is the marginal of

$$f_t(y_t, u_t|G) = \sum_{k=1}^{\infty} \mathbb{1}_{\{u_t<w_k\}} b^*_{\mu_k,\nu_k}\left(H_t(y_t|\boldsymbol{\omega}_k)\right) h_t(y_t|\boldsymbol{\omega}_k)$$

Note that given a set of observations, y_t and slice variables, u_t, $t = 1, \ldots, T$, the complete data likelihood can be written as

$$L(Y, U|G) = \prod_{t=1}^{T} \sum_{k\in A_t} b^*_{\mu_k,\nu_k}\left(H_t(y_t|\boldsymbol{\omega}_k)\right) h_t(y_t|\boldsymbol{\omega}_k),$$

where $Y = (y_1, \ldots, y_T)$, $U = (u_1, \ldots, u_T)$, $A_t = \{k|u_t < w_k\}$. Note that $N_t = Card(A_t)$, that is the number of components of the infinite sum, is finite when conditioning on the slice variables. Thus, the introduction of the auxiliary variables allows us to have a finite mixture representation of the infinite mixture model. Following a standard approach to inference for mixture models (e.g., see Frühwirth-Schnatter (2006)) we now introduce a sequence of allocation variables, d_t, $t = 1, \ldots, T$, with $d_t \in A_t$. Each of these variables indicates which component of the finite mixture provides the observation y_t. The complete data likelihood is

$$L(Y, U, D|G) = \prod_{t=1}^{T} \mathbb{1}_{\{u_t<w_{d_t}\}} b^*_{\mu_{d_t},\nu_{d_t}}\left(H_t(y_t|\boldsymbol{\omega}_{d_t})\right) h_t(y_t|\boldsymbol{\omega}_{d_t})$$

where $D = (d_1, \ldots, d_T)$.

Let us denote by $V = (v_1, v_2, \ldots)$ and $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots)$, with $\boldsymbol{\theta}_k = (\mu_k, \nu_k, \boldsymbol{\omega}_k)$, $\boldsymbol{\omega}_k = (\omega_{1k}, \ldots, \omega_{Mk})$, the infinite dimensional vectors of the stick-breaking components and atoms respectively. In what follows we assume the dispersion parameter ψ is unknown with prior distribution $\pi(\psi)$.

From the completed likelihood function and our assumptions on the prior distributions, the joint posterior distribution of U, D, V, Θ and ψ given Y is

$$\pi(U, D, V, \Theta, \psi|Y) \propto \prod_{t=1}^{T} \mathbb{1}_{\{u_t < w_{d_t}\}} b^*_{\mu_{d_t},\nu_{d_t}} \left(H_t(y_t|\boldsymbol{\omega}_{d_t})\right) h_t(y_t|\boldsymbol{\omega}_{d_t})$$

$$\times \prod_{k\geq 1}(1-v_k)^{\psi-1}\mu_k^{\xi_\mu-1}(1-\mu_k)^{\xi_\mu-1}\nu_k^{\xi_\nu/2}\exp\{-\xi_\nu\nu_k/2\}\prod_{i=1}^{M}\omega_{ik}^{\nu/2-1}\pi(\psi).$$

Joint sampling from the posterior is not possible and this calls for the application of a Gibbs sampling procedure. Adapting the sampler described in Walker (2007) and Kalli et al. (2011) to our setting, we develop an efficient collapsed Gibbs sampling procedure which generates sequentially the parameters and the latent variables from the full conditional distributions $\pi(\Theta|U, D, V, Y, \psi)$, $\pi(V, U|\Theta, D, Y, \psi)$, $\pi(D|\Theta, V, U, Y, \psi)$ and $\pi(\psi|Y)$. The details of the steps of the Gibbs sampler are given in Appendix A.2.

The output of the algorithm are samples $\boldsymbol{w}^{(i)}$ and $\boldsymbol{\theta}^{(i)} = (\boldsymbol{\mu}^{(i)}, \boldsymbol{\nu}^{(i)}, \boldsymbol{\omega}^{(i)})$ for $i = 1, \ldots, I$ where I is the number of MCMC iterations, and can be used to sample from the one-step-ahead cumulative predictive distribution. For further details see Appendix A.2.

4 Posterior consistency

In this section we discuss the weak posterior consistency of the infinite mixture model BM_∞. Weak consistency guarantees that asymptotically the posterior accumulates in weak neighbourhoods of the "true" density f_0. Roughly speaking, the posterior learns from the data and puts more and more mass near f_0.

In the following, we focus on the i.i.d. case and provide general results which cover the models considered in the simulation examples and the application to weather forecast. As regards the non i.i.d. case, posterior consistency proof is case-specific depending heavily on the model used. For instance, see Tang and Ghosal (2007) for posterior consistency of Bayesian nonparametric estimates with transition kernel of an ergodic Markov process and Choudhuri et al. (2004) for the estimation of the spectral density of stationary and short-memory Gaussian time series. Posterior consistency results for calibration in the non i.i.d. case are left for future research.

Let $\mathcal{F}$ be the set of all possible densities on the sample space $\mathcal{Y} \subset \mathbb{R}$ and Π^* be a prior on $\mathcal{F}$. The posterior is said to be *weakly consistent* at f_0 if $\Pi^*(U|y_1, \dots, y_n)$ converges a.s. to 1 for every weak neighbourhood U of f_0, whenever $y_1, y_2, \dots$ are i.i.d. observations with common density f_0.

The Schwartz theorem states that the consistency at a "true density" f_0 holds if the prior assigns positive probabilities to Kullback-Leibler neighborhoods of f_0. Hence one only needs to check if the Kullback-Leibler property is satisfied by the prior setting and the true density f_0, see Theorem 4.4.2 in Ghosh and Ramamoorthi (2003).

More formally, a Kullback-Leibler neighbourhood of a density $f \in \mathcal{F}$ of size ε is defined as

$$\mathcal{K}_\varepsilon(f_0) = \left\{ g \in \mathcal{F} | \int f \log\left(\frac{f}{g}\right) \leq \varepsilon \right\},$$

and the Kullback-Leibler property holds at $f_0 \in \mathcal{F}$, for short $f_0 \in KL(\Pi^*)$, if $\Pi^*(\mathcal{K}_\varepsilon(f_0)) > 0$ for all $\varepsilon > 0$. We will denote with $supp(\mu)$ the weak support of a probability measure μ and with $KL(f, g)$ the Kullback-Leibler divergence between the two densities f and g, i.e. $KL(f, g) := \int f \log\left(\frac{f}{g}\right)$.

In this section we will exploit the type I mixture prior representation of Π^*. Let us recall that a prior on $\mathcal{F}$ is said to be a type I mixture prior if it is induced via the map

$$G \mapsto f_G(y) = \int_\Theta K(y; \boldsymbol{\theta}) G(d\boldsymbol{\theta}), \tag{14}$$

where Θ is the mixing parameter space, $K(y; \boldsymbol{\theta})$ a density kernel on $\mathcal{Y} \times \Theta$ and G has distribution Π on the space $\mathcal{M}(\Theta)$ of probability measures on Θ (see Wu and Ghosal (2009a)).

In our joint calibration and combination model, the kernel is

$$K(y; \boldsymbol{\theta}) = b^*_{\mu,\nu}(H(y|\boldsymbol{\omega}))h(y|\boldsymbol{\omega}) \tag{15}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_c)$, where $\boldsymbol{\theta}_p = \boldsymbol{\omega}$ indicates the pooling parameters, and $\boldsymbol{\theta}_c = (\mu, \nu)$ the calibration parameters. Since we deal only with the case of i.i.d. observations, we drop from the kernel K the observation index, that is the conditioning on other variables. The random mixing distribution Π is given by a Dirichlet process prior, so that

$$\boldsymbol{\theta}|G \sim G \tag{16}$$

where $G \sim DP(\psi, G_0)$. Again, for the sake of simplicity we assume that the concentration parameter ψ is given.

4.1 Joint calibration and combination consistency

Let us first consider the case in which both the pooling parameters and the calibration parameters are unknown. In this case $\Theta = \Delta_M \times [0,1] \times \mathbb{R}^+$ and G is a DP process on $\mathcal{M}(\Delta_M \times [0,1] \times \mathbb{R}^+)$ with base measure G_0 on $\Delta_M \times [0,1] \times \mathbb{R}^+$ and concentration parameter $\psi > 0$.

Here Π^* turns out to be the prior on $\mathcal{F}$ induced by

$$G \mapsto \int b^*_{\mu,\nu}(H(y|\boldsymbol{\omega}))h(y|\boldsymbol{\omega})G(d\boldsymbol{\omega}d\mu d\nu)$$

when $G \sim DP(\psi, G_0)$.

Before stating the first result, let us recall that $h(y|\boldsymbol{\omega}) = \sum_{m=1}^M \boldsymbol{\omega}_m f_m(y)$.

Theorem 4.1. *Assume that the functions $f_m(\cdot)$ are continuous on $\mathcal{Y}$. Let u_0 be a continuous density on $(0,1)$ such that*

$$\begin{aligned} &\int_0^1 [|\log(x)| + |\log(1-x)|]u_0(x)dx < +\infty \\ \textit{and} \quad &\int_0^1 \log(u_0(x))u_0(x)dx < +\infty. \end{aligned} \tag{17}$$

Let $f_0(y) = u_0(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0})$ with $\boldsymbol{\omega_0}$ in the interior of Δ_M and assume that, for every compact set $C \subset \mathcal{Y}$,

$$\inf_{y \in C} h(y|\boldsymbol{\omega_0}) > 0. \tag{18}$$

Then $f_0 \in KL(\Pi^*)$ whenever G_0 has full support.

The proof of the previous theorem is postponed to Appendix B. A useful restatement of the previous result is contained in the following Corollary.

Corollary 4.1. *Assume that there is $\boldsymbol{\omega_0}$ in the interior of Δ_M such that $h(\cdot|\boldsymbol{\omega_0})$ is continuous and (18) holds for every compact set $C \subset \mathcal{Y}$ and let f_0 be a continuous density on $\mathcal{Y}$ such that*

$$\begin{aligned} &\int [|\log(H(y|\boldsymbol{\omega_0}))| + |\log(1 - H(y|\boldsymbol{\omega_0}))|]f_0(y)dy < +\infty \\ \textit{and} \quad &KL(f_0, h(\cdot|\boldsymbol{\omega_0})) < +\infty. \end{aligned} \tag{19}$$

If G_0 has full support, then $f_0 \in KL(\Pi^*)$.

Proof. Write H_0 and h_0 for $H(\cdot|\boldsymbol{\omega_0})$ and its density. By assumption (18) one gets that H_0 is continuous and strictly increasing. Hence, if one defines

$$u_0(x) := \frac{f(H_0^{-1}(x))}{h_0(H_0^{-1}(x))},$$

it follows that $f_0(y) = u_0(H_0(y))h_0(y)$. Note that u_0 turns out to be a continuous function on $(0,1)$. It remains to check that assumption (19) yields (17). Now, a change of variable gives

$$\int |\log(H(y|\boldsymbol{\omega_0}))|f_0(y)dy = \int |\log(H(y|\boldsymbol{\omega_0}))|u_0(H_0(y))h_0(y)dy = \int |\log(x)|u_0(x)dx.$$

Similarly for $\int |\log(1 - H(y|\boldsymbol{\omega_0}))|]f_0(y)dy$. Finally

$$KL(f_0, h(\cdot|\boldsymbol{\omega_0})) = \int \log(u_0(H_0(y))u_0(H_0(y))h_0(y)dy = \int u_0(x)\log(u_0(x))dx.$$

□

The assumptions of Corollary 4.1 can be easily checked for many applied contexts. Here we show that the assumptions are satisfied for the Gaussian mixture and Student-t mixture examples considered later on in this paper for the simulation study.

Example 4.1. *Consider the case in which*

$$h(y|\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \varphi(y|\mu_m, \sigma_m^2), \quad f_0(y) = \sum_{i=1}^{K} p_i \varphi(y|\mu_i^*, \sigma_i^{*2})$$

where $\varphi(\cdot|\mu, \sigma^2)$ is the pdf of a normal distribution of mean μ and variance σ^2. Denote by $\Phi(\cdot|\mu, \sigma^2)$ the cumulative distribution function of $\varphi(\cdot|\mu, \sigma^2)$. Let us prove that that $f_0 \in KL(\Pi^*)$.

In order to apply Corollary 4.1 one needs to check that (19) is satisfied for some $\boldsymbol{\omega_0}$ in the interior of Δ_M. E.g., consider the equal weights linear pooling, $\boldsymbol{\omega_0} = (1/M, \dots, 1/M)$. To this end observe that:

(i) given a mixture of M normal distributions with means and variances (μ_m, σ_m^2), $m = 1, \dots, M$, if $0 < \sigma_- < \min_m \sigma_m \leq \max_m \sigma_m < \sigma_+$, then there are two constants C^- and C^+ such that, for every y,

$$C^- \varphi(y|0, \sigma_-^2) \leq \sum_{m=1}^{M} \omega_m \varphi(y|\mu_m, \sigma_m^2) \leq C^+ \varphi(y|0, \sigma_+^2);$$

(ii) as $y \to +\infty$, one has $(1 - \Phi(y|0,1))/\varphi(y|0,1) \sim 1/y)$ and hence

$$|\log(1 - \Phi(y|0,\sigma^2))| \sim y^2/\sigma^2.$$

Using (i) and (ii) one can check that

$$|\log(1 - H(y|\boldsymbol{\omega_0}))| \leq C \max\{|\log(1 - \Phi(y|0,\sigma-^2)|, |\log(1 - \Phi(y|0,\sigma+^2)|\} \leq C' y^2$$

for suitable constants C, C'. Analogous considerations hold for $|\log(H(y|\boldsymbol{\omega_0}))|$. Hence the first condition in (19) *is satisfied. Using (i) and the fact that $KL(\varphi(\cdot|\mu_1,\sigma_1^2), \varphi(\cdot|\mu_2,\sigma_2^2)) < +\infty$, it is easy to obtain also that $KL(h(\cdot|\boldsymbol{\omega_0}), f_0) < +\infty$.*

Example 4.2. *Consider the case in which*

$$h(y|\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \varphi(y|\mu_m, \sigma_m^2), \quad f_0(y) = \sum_{i=1}^{K} p_i \mathcal{T}_{\mu_i^*,\sigma_i^*,\nu}(y),$$

where $\mathcal{T}_{\mu,\sigma,\nu}$ is a t-distribution with location, scale and degrees of freedom paramters μ, σ and ν respectively. Since $f_0(y) \sim C y^{-\nu-1}$ as $|y| \to +\infty$, arguing as in the previous example it is easy to see that (19) *is satisfied whenever $\nu > 2$. In this case $f_0 \in KL(\Pi^*)$.*

4.2 Calibration consistency

If the pooling parameters $\boldsymbol{\omega}_0$ are known, the inference is limited to the calibration parameters $\boldsymbol{\theta}_c = (\mu, \nu)$, hence $\Theta = [0,1] \times \mathbb{R}^+$ and G is a DP process on $\mathcal{M}([0,1] \times \mathbb{R}^+)$ with base measure G_0 and concentration parameter ψ.

In this special case Π^* turns out to be the prior induced by

$$G \mapsto \int b^*_{\boldsymbol{\theta}_c}(H(y|\boldsymbol{\omega_0}))G(d\boldsymbol{\theta}_c)h(y|\boldsymbol{\omega_0})$$

when $G \sim DP(\psi, G_0)$.

The analogous of Corollary 4.1 is given below. Note that here $\boldsymbol{\omega_0}$ is not necessarily assumed to be in the interior of Δ_M, which means that the set of models in the combination scheme can be complete.

Theorem 4.2. *Let $\boldsymbol{\omega_0}$ be a given point in Δ_M such that $h(\cdot|\boldsymbol{\omega_0})$ is continuous and* (18) *holds for every compact set $C \subset \mathcal{Y}$ and let f_0 be a continuous density on $\mathcal{Y}$ such that* (19) *holds. If G_0 has full support, then $f_0 \in KL(\Pi^*)$.*

In some situations, it is useful to consider a base measure G_0 without full support. In this spirit, following the techniques of Tang et al. (2007), we can prove the next result.

Theorem 4.3. *Let $\boldsymbol{\omega_0}$ be a given point in Δ_M and let*

$$f_0(x) = u_0(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0})$$

with $u_0(x) = w_0 b^*_{\mu_0,\nu_0}(x) + (1-w_0)\int_{(0,1)\times\mathbb{R}^+} b^*_{\mu,\nu}(x)P_0(d\mu d\nu)$, P_0 being a probability measure on $(0,1)\times\mathbb{R}^+$. If (μ_0,ν_0) belong to supp(G_0), supp$(P_0) \subset$ supp(G_0), and for some $\zeta > 0$ and $0 < \eta < \min(\mu_0, 1-\mu_0, \nu_0, w_0)$ one has

$$\int_0^1 \frac{u_0(x)^{\zeta+1}}{x^{\zeta A}(1-x)^{\zeta B}}dx < +\infty, \tag{20}$$

for $A = (\mu_0+\eta)(\nu_0+\eta) - 1$ and $B = (1-\mu_0+\eta)(\nu_0+\eta) - 1$, then $f_0 \in KL(\Pi^*)$.

5 Simulation examples

We assume that a combined predictive distribution can be obtained from the two normal predictive distributions with different location and equal scale parameters, $\mathcal{N}(-1,1)$ and $\mathcal{N}(2,1)$, where $\mathcal{N}(\mu,\sigma^2)$ denotes the normal distribution with location μ and scale σ.

In the simulation experiments, the hyperparameter setting for the BC and BMC model is $\xi_{j\mu} = 2$, $\xi_{j\nu} = 0.1$ and $\xi_{j\omega} = 1$, and $\xi_{jw} = 1$, $j = 1,2$. The priors are informative, but with a large prior variance, thus one can expect posterior inference should not be affected by the hyperparameter settings. Our experiments show that the results, in terms of calibration, do not change when considering less informative prior settings, and secondly that the use of improper prior distributions in mixtures model, even if possible, still remains an open issue. See e.g. Wasserman (2000) for a discussion on the use of improper prior in mixture modelling.

- Linear pool (LP)

  $$f(y|\boldsymbol{\theta}) = \omega\varphi(y|-1,1) + (1-\omega)\varphi(y|2,1),$$

  where $\boldsymbol{\theta} = \omega$. The model weights in the linear pooling are estimated using the recursive log score, see e.g. Jore et al. (2010). Equals BM_1 with $\alpha = \beta = 1$ fixed.

- Beta-transformed linear pool (BM$_1$)

$$f(y|\boldsymbol{\theta}) = f_{\alpha,\beta}\left(H(y|\omega)\right) h(y|\omega),$$

where $\boldsymbol{\theta} = (\alpha, \beta, \omega)$, $h(y|\omega) = \omega\varphi(y|-1,1) + (1-\omega)\varphi(y|2,1)$ and $H(y|\omega) = \omega\nu(y|-1,1) + (1-\omega)\nu(y|2,1)$.

- Two-component finite beta mixture model (BM$_2$)

$$f(y|\boldsymbol{\theta}) = wf_{\alpha_1,\beta_1}\left(H(y|\omega_1)\right) h(y|\omega_1) + (1-w)f_{\alpha_2,\beta_2}\left(H(y|\omega_2)\right) h(y|\omega_2),$$

where $\boldsymbol{\theta} = (w, \alpha_1, \alpha_2, \beta_1, \beta_2, \omega_1, \omega_2)$, and $h(y|\omega)$ and $H(y|\omega)$ have been defined as in the BC model.

- Infinite beta mixture model (BM$_\infty$)

Estimation: Based on a set of 1,000 MCMC iterations after a burn-in period of 2,000 iterations.

For expository purposes we arbitrarily set, in Table 1, $\alpha_1 = \alpha$, $\beta_1 = \beta$ and $w = 1$ for the BC models and $\omega_1 = \omega$ for the models with common linear combination.

5.0.1 Multimodality

Let us denote with $\varphi(x|\mu, \sigma^2)$ and $\Phi(x|\mu, \sigma^2)$ the pdf and cdf respectively of a $\mathcal{N}(\mu, \sigma^2)$. We assume that the data are generated by the following mixture of the three normal distributions

$$y_t \overset{i.i.d.}{\sim} p_1\mathcal{N}(-2, 0.25) + p_2\mathcal{N}(0, 0.25) + p_3\mathcal{N}(2, 0.25), \quad t = 1, \ldots, 1000,$$

where $p = (p_1, p_2, p_3) \in \Delta_3$.

The posterior means of the parameter of the calibration and combination models are reported in Table 1. Figure 1 shows the empirical cdfs of different sequences of probability integral transform (PIT). In all the experiments, the PIT of the non-calibrated model (red lines) is far from the standard uniform (black lines). In these datasets, the BC clearly lacks calibration. The BC cdf (green line) is closer to uniformity than the NC model, but it has difficulties in deforming the combination density some parts of the support.

More specifically, the two-component beta calibrations are able to achieve a more flexible deformation of the cdf linear combination providing a calibrated cdf (blue and magenta lines) which is close to the uniform cdf. Figure 2 shows the results of the calibration and combination procedure

Table 1: Parameter settings (posterior means) for the calibration models BM$_1$ and BM$_2$, for different datasets, of i.i.d. 1000 observations each, simulated from the mixture model $p_1\mathcal{N}(-2, 0.25) + p_2\mathcal{N}(0, 0.25) + p_3\mathcal{N}(2, 0.25)$, for different values of $\boldsymbol{p} = (p_1, p_2, p_3)$. Note that for expository purposes we arbitrarily set $\alpha_1 = \alpha$, $\beta_1 = \beta$ and $w = 1$ for the BC models and $\omega_1 = \omega$ for the common linear pooling models.

$\boldsymbol{p}$	$(1/5, 1/5, 3/5)$		$(1/7, 1/7, 5/7)$	
$\boldsymbol{\theta}$	BM$_1$	BM$_2$	BM$_1$	BM$_2$
α	0.97	0.94	1.04	0.87
β	1.50	27.48	1.47	2.08
ω	0.20	0.04	0.17	0.29
w		0.36		0.44
α^*		22.19		17.71
β^*		4.87		5.09
ω^*		0.67		0.54

$\boldsymbol{p}$	$(1/5, 1/5, 3/5)$		$(1/7, 1/7, 5/7)$	
$\boldsymbol{\theta}$	BM$_1$	BM$_2$	BM$_1$	BM$_2$
w	1.00	0.48	1.00	0.29
α	0.74	2.47	0.74	6.61
β	1.72	2.11	2.03	2.44
ω	0.52	0.54	0.54	0.72
α^*		2.30		1.96
β^*		34.21		51.00
ω^*		0.39		0.19

decomposed along the different components of the mixture. As an example consider the first dataset, generated with $\boldsymbol{p} = (1/5, 1/5, 3/5)$. The solid and dashed blue lines in the top-left plot of Figure 2 show the contribution of the first and second component respectively of the BMC1 mixture model to the calibration of the density. The first component mainly calibrates the pdf on the positive part of the support and the second component calibrates the pdf on the negative part of the support. The results in Table 1 show that both components assign the same weights ($\omega = 0.449$) to the first model in the pool, i.e. $\mathcal{N}(-1, 1)$. This weight is higher than in the BC model, which has a less flexible calibration function and thus assigns a lower weight $\omega = 0.202$ to the first model in the pool. The solid and dashed magenta lines in the top-left plot of Figure 2 show a behaviour similar to the BMC1 components.

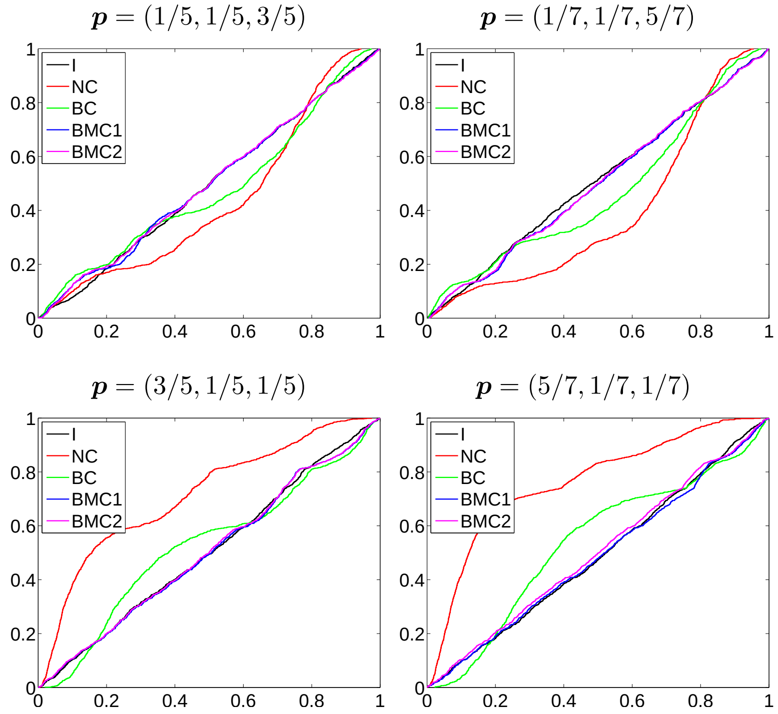

Figure 1: PITs cdf for different calibration models and different datasets.

Table 1 indicates that the first BMC2 component assigns weight $\omega_1 = 0.043$ to the first model in the pool. This means that the calibration on the negative part of the support set is done mainly using the predictive distribution of the second model, $\mathcal{N}(2, 1)$. The calibration of the positive part of the set is obtained thanks to the second BMC2 component which assigns weight $\omega_2 = 0.667$ to the first model in the pool.

To investigate the sensitivity of the posterior quantities to the choice of the hyperparameters, we combine and calibrate the cdfs, on the same dataset, using two different values of the dispersion parameters, $\psi = 1$ and $\psi = 5$,

The top charts of Figure 3 report the PITs of the average infinite beta mixture calibration (BMC) model and their 99% credibility intervals obtained from 1,000 MCMC samples after convergence. Usually a burn-in sample of 1,000 is used. The PITs of the calibrated model (black lines) belong to the credibility interval of the BMC, thus the resulting predictive cdf is well calibrated. We should notice that the credibility intervals are

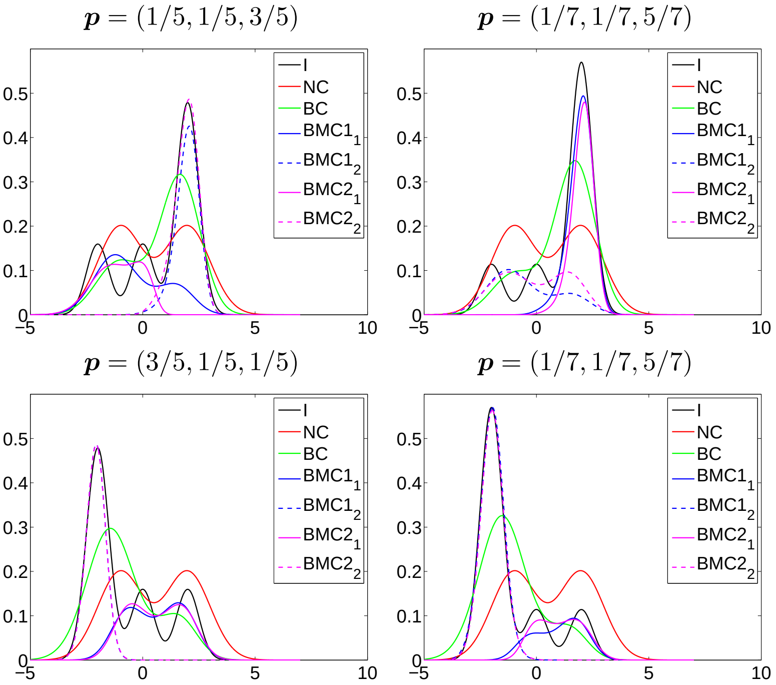

Figure 2: Contribution of the calibration components for different models BC, BMC1 (first and second mixture component, $BMC1_1$ and $BMC1_2$) and BMC2 (first and second mixture component, $BMC2_1$ and $BMC2_2$), and different datasets.

usually larger than the one obtained using a beta mixture with a fixed number of components. In fact the calibrated density accounts for both calibration parameter uncertainty and also for the uncertainty about the number of mixture components. A comparison between the left- and right-top chart also show that an increase in the value ofs the dispersion parameter usually increases the uncertainty.

The credibility intervals (gray lines) obtained with the infinite beta mixture calibration model, see Figure 3, always contain the PITs (first row) and the predictive density function (second row) of the correct model. The infinite BMC seems particularly accurate in the tails (right column). We also note that the uncertainty of the number components in the infinite beta mixture implies a wider high probability density region (HPD), see gray lines in 3, than that given by the finite beta mixture calibration, see third panel in 4. The prior and posterior distributions of the number of mixture components in BMC are given in the bottom graph in Figure 3.

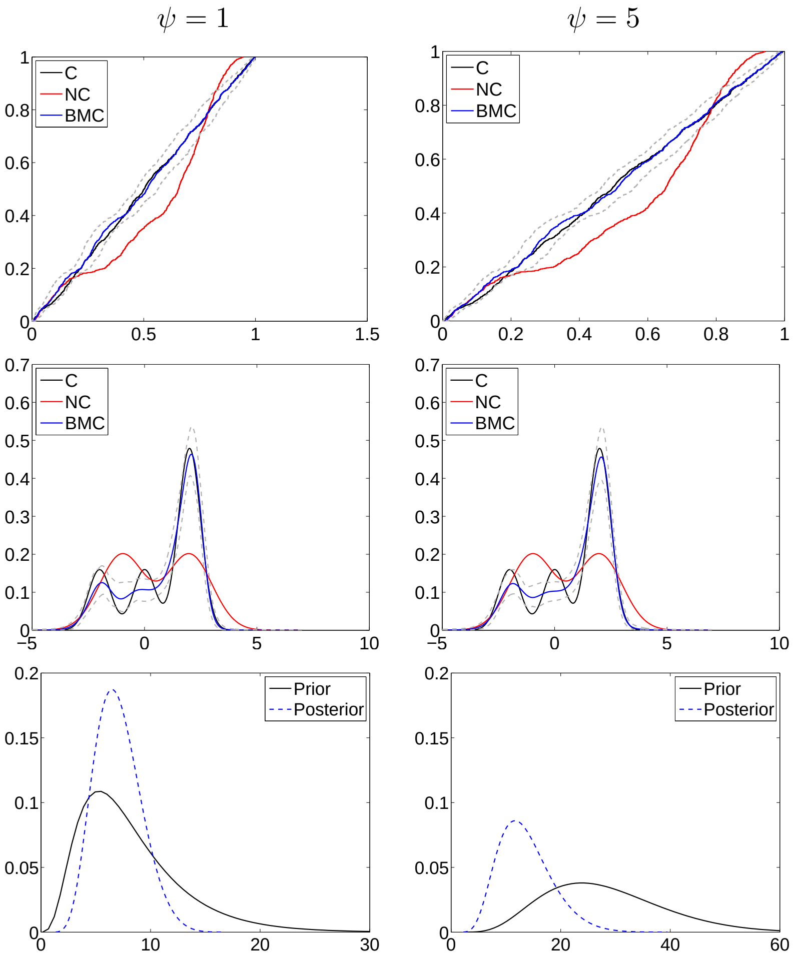

Figure 3: Infinite beta mixture calibrated (BMC), calibrated (C) and non calibrated (NC) combinations for a dataset of 1,000 samples from $p_1\mathcal{N}(-2, 0.25) + p_2\mathcal{N}(0, 0.25) + p_3\mathcal{N}(2, 0.25)$ with $\boldsymbol{p} = (1/5, 1/5, 3/5)$. PITs cdf (top graph) and calibrated pdf (middle graph) of the combination models C (black), NC (red) and BMC (blue) and BMC 99% HPD (gray). Prior (black) and posterior (blue) number of components of the random BMC model (bottom graph).

The posterior density is more concentrated than the prior, suggesting that data are informative on the number of calibration components.

5.0.2 Heavy tails

In second example, we assume that the data are generated by the following mixture of t-distributions, i.e.

$$y_t \overset{i.i.d.}{\sim} \frac{1}{2}\mathcal{T}(-1,1,6) + \frac{1}{2}\mathcal{T}(2,1,6), \quad t = 1,\dots,3000,$$

where $\mathcal{T}(\mu,\sigma,\nu)$ denotes a t-distribution with location, scale and degrees of freedom parameters μ, σ and ν respectively. As before, we assume that the predictive distribution is obtained from the combination of the two normal distributions given in the example of the previous section, which are $\mathcal{N}(-1,1)$ and $\mathcal{N}(2,1)$. The NC, BC, BMC1 and BMC2 models are defined as in the first example. Fig 4 focuses on the right tail of the predictive pdf and shows results for the calibrated and beta calibrated PITs cdf and their 99% HPD. There is strong evidence of the difficulties of the BC model in calibrating the tails. The BC underestimates the tail probability and over-estimates the central part of the distribution. The BMC1 and BMC2 models are able instead to provide well-calibrated PITs on the tails of the distribution.

In the second set of experiments we assume ψ is unknown. The resulting hierarchical model is a continuous mixture of infinite Dirichlet mixture, which usually leads to more dispersed predictive distributions. The results of the infinite mixture calibration are given in Figure 5.

Both cdf (top panel) and pdf (second panel) indicate that the Bayesian BC has problems producing well-calibrated predictions. The Bayesian nonparametric calibration BMC, on the contrary, produces well-calibrated densities, in particular on the tails; see also the 99% credibility intervals. We note that the posterior distribution of the number of clusters is more concentrated than the prior, thus there is learning from the data on the number of mixture components. Finally, our experiments changing the dispersion parameter indicate no substantial changes in the posterior density over different hyperparameter values.

6 Empirical applications

Then, we investigate relative predictability accuracy for the out-of-sample period. Precisely, as in Geweke and Amisano (2010), Geweke and Amisano

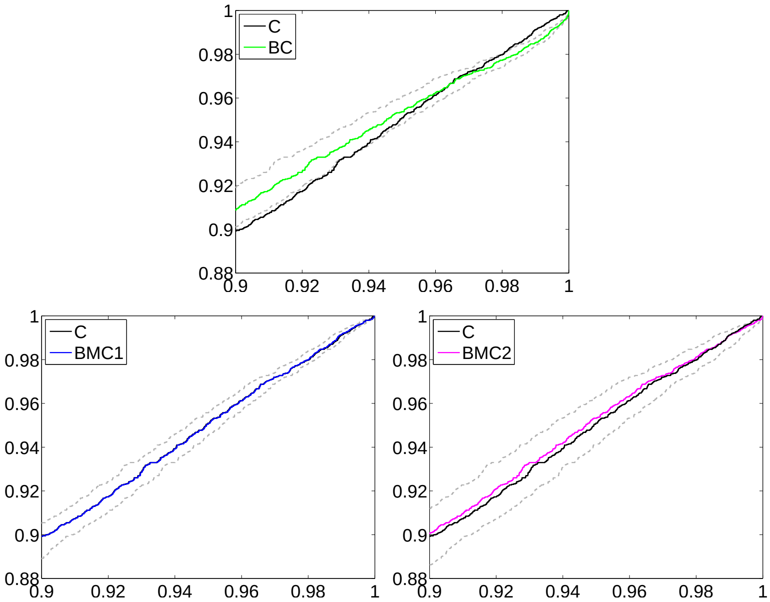

Figure 4: Results of the Bayesian calibration model BC (top), BMC1 (middle) and BMC2 (bottom) for the right tail of the predictive distribution. In each plot the PITs cdf of the calibrated (solid black line) and beta calibrated model (solid coloured line) and their 99% HPD region (gray dashed lines)

(2011) and Fawcett et al. (2013), we evaluate the predictive densities using the Kullback Leibler Information Criterion (KLIC) based measure, utilizing the expected difference in the Logarithmic Scores of the candidate forecast densities. The KLIC computes the distance between the true density of a random variable and some candidate density. Even though the true density is not known, for the comparison of two competing models, it is sufficient to consider the average Logarithmic Score (AvLS). The continuous ranked probability score (CRPS), defined at time t for model k as:

$$\mathrm{CRPS}_{t,k} = \int \left(F_{t,k}(z_t) - \mathbb{1}_{[y_t,+\infty)}(z) \right)^2 \mathrm{d}z$$

where $F_{t,k}(y_t)$ and $f_{t,k}(y_t)$ are the cdf and pdf, respectively, for model k.

6.1 Stock returns

The first application considers S&P500 daily percent log returns data from 3 January 1972 to 31 December 2008, an updated version of the database used

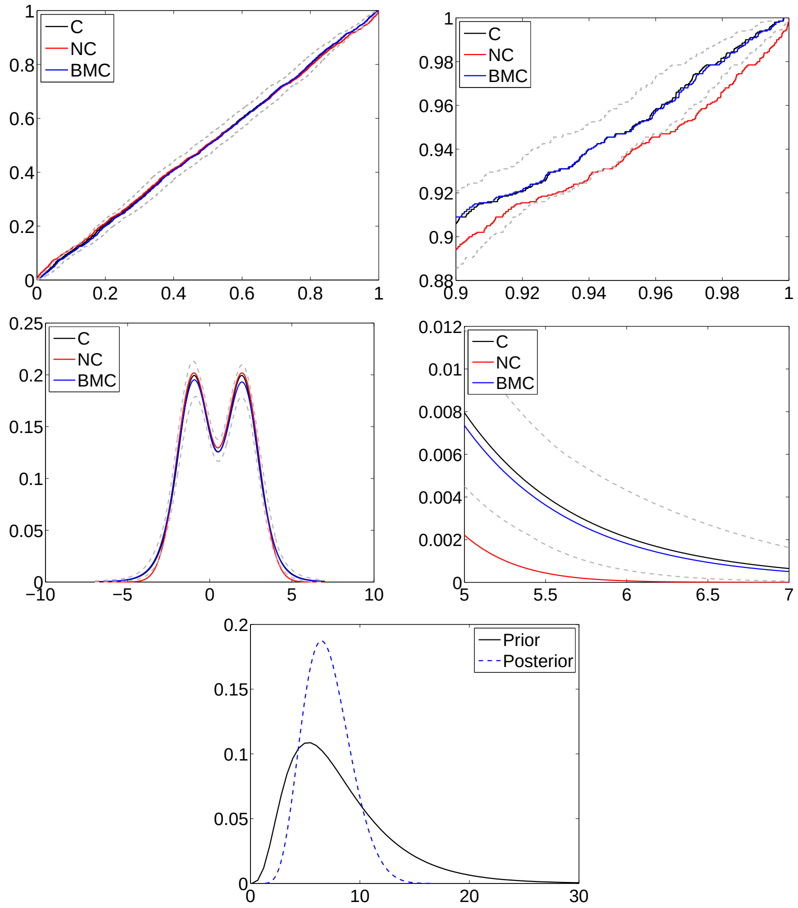

Figure 5: Infinite beta mixture calibrated (BMC), calibrated (C) and non calibrated (NC) combinations for a dataset of 2,000 samples from $1/2\mathcal{T}(-1,1,6) + 1/2\mathcal{T}(2,1,6)$. PITs cdf (top graphs) and calibrated pdf (middle graphs) of the combination models C (black), NC (red) and BMC (blue) and BMC 99% HPD (gray). Prior (black) and posterior (blue) number of components of the random BMC model (bottom).

in studies such as Geweke and Amisano (2010), Geweke and Amisano (2011) and Fawcett et al. (2013).[1] We estimate a Normal GARCH(1,1) model and a t-GARCH(1,1) model via maximum likelihood (ML) using rolling samples of 1250 trading days (about five years) and produce one day ahead density forecasts. The first one day ahead forecast refers to December 15, 1975. The predictive densities are formed by substituting the ML estimates for the unknown parameters. We combine the two predictive densities using a linear pooling with recursive log score weights, see description in Section 5.[2] Also in this section, we refer to it as the non-calibrated model. Furthermore we consider our mixture of beta probability density functions (BMC) to achieve better calibration properties. We split the sample in two periods. The data from December 15, 1975 to December 31, 2006 are used for an in-sample calibration of our method to investigate its properties over a long period. The data from January 3, 2007 to December 31, 2008 for a total of 504 observations, are used for our out-of-sample analysis. Therefore, we extend evidence in Geweke and Amisano (2010) and Geweke and Amisano (2011) by focusing on the period related to Great Financial Crisis, with the first semester of 2007 considered a tranquil period and the remaining part of the sample corresponding to the most turbulent times. In this experiment, we fit the calibration over a moving window of 250 days and produce one-day ahead forecasts.[3]

First, we compare the two individual models and the two combinations in terms of calibration, measured as PIT, over the full sample period (in-sample and out-of-sample periods).[4] Figure 6 reports calibration results for the in-sample analysis. The BMC line is the closer to the 45 degree line, which represents the PIT plot for the unknown true/ideal model. This 45 degree line always belongs to the confidence interval of the BMC. NC is not calibrated for all quantiles. In particular, on the upper and lower tails, the NC differs substantially from the BMC. As in the simulation exercises the posterior density for the numbers of beta components in BMC is more concentrated than the prior.

We now turn our attention to the out-of-sample analysis and the log score

[1] We thank James Mitchell for providing data.

[2] More flexible weighting schemes, such as time-varying weights, can also be computed.

[3] We also investigated out-of-sample performance over the period from 15 December 1976 to 16 December 2002, the same sample applied in Geweke and Amisano (2010) and Geweke and Amisano (2011). Superior performance of our BMC are confirmed in this longer sample.

[4] Figures focusing only on the out-of-sample period provide similar evidence and available upon request from the authors.

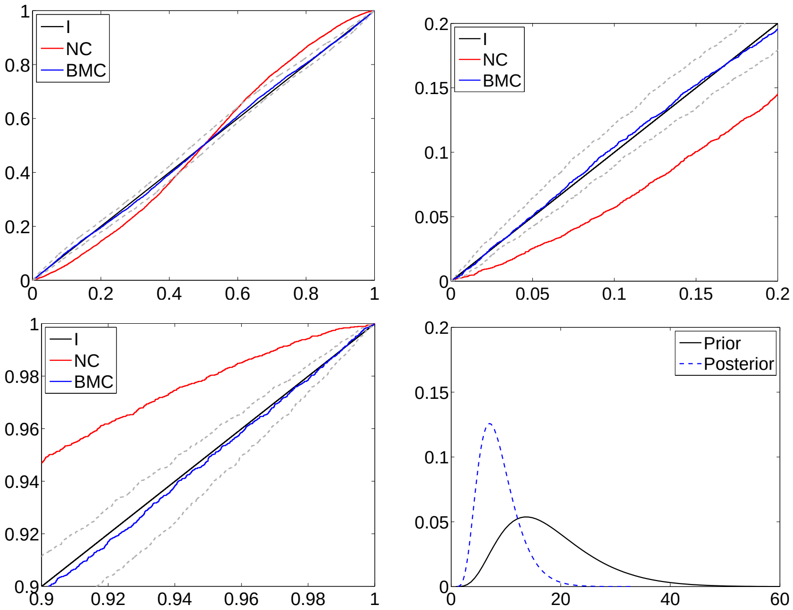

Figure 6: Infinite beta mixture calibrated (BMC), calibrated (C) and non calibrated (NC) combinations for the S&P500 daily percent log returns data. PITs cdf (lines from 1 to 3) of the idea model I C (black), combination models NC (red) and BMC (blue) and BMC 99% HPD (gray). Prior (black) and posterior (blue) number of components of the random BMC model (bottom).

results. Table 2 reports average log scores for the 4 forecasting methods. BMC provides the highest score. Figure 7 shows that after the initial weeks of January 2007 where models perform similarly, BMC outperforms the other three approaches. The t-model provides higher scores than the normal one and the non-calibrated combination. The accuracy of the normal Garch model is very low during our OOS period, in particular on the extreme events, which results in deteriorating NC performance after August 2007, the beginning of the turbulent times. Just selecting the t-GARCH version or, even better, applying local weights as in our BMC improves accuracy. Figure 8 shows the BMC-based predictive density.

Table 2: Average log scores for the Normal GARCH model (Normal), t-GARCH model (Student-t), linear pooling (NC) and beta mixture calibration (BMC) over the sample period from January 1, 2007 to December 31, 2008.

	Normal	Student-t	NC	BMC
AvLS	-2.311	-1.650	-1.827	-1.450

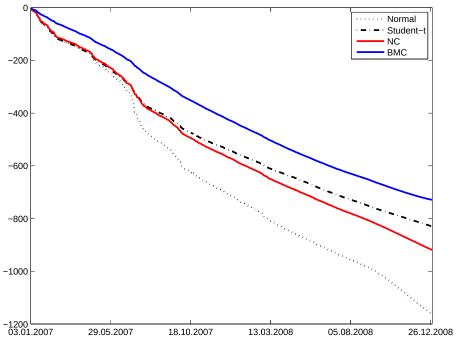

Figure 7: Cumulative log scores for the Normal GARCH model (Normal), t-GARCH model (Student-t), linear pooling (NC) and beta mixture calibration (BMC) over the sample period from January 1, 2007 to December 31, 2008.

6.2 Wind speed

The second empirical example considers the dataset used in Lerch and Thorarinsdottir (2013).[5] It consists of 50 ensemble member predictions (Molteni et al., 1996) of wind speed at ten meters above the ground, obtained from the global ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF). We restrict our attention to the ensemble predictions for the maximum wind speed at the station at Frankfurt airport. The station ensemble forecasts are obtained by bilinear interpolation of the gridded model output.

We consider the ECMWF ensemble run initialized at 00 hours UTC with a horizontal resolution of about 33 km, a temporal resolution of 3-6 hours

[5] We thank Sebastian Lerch and Thordis Thorarinsdottir for providing data and forecasts.

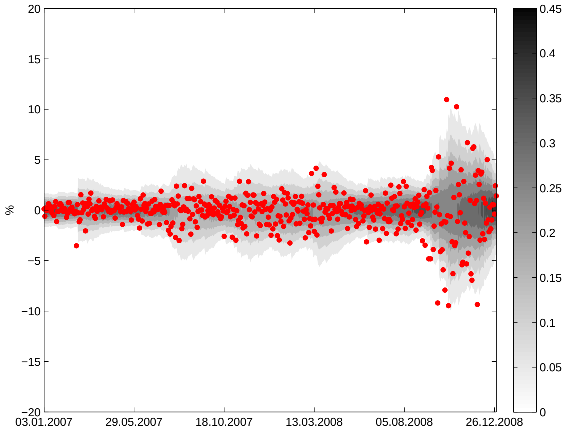

Figure 8: Fanchart of the BMC model and observations (red points) over the sample period from January 1, 2007 to December 31, 2008.

and lead times of 3, 6 and 24 hours. To obtain predictions of daily maximum wind speed, we take the daily maximum of each ensemble member at the Frankfurt location. One day ahead forecasts are given by the maximum over lead times. The observations are hourly observations of 10-minute average wind speed which is measured over the 10 minutes before the hour. To obtain daily maximum wind speed, from 1 May 2010 to 30 of April 2011, we take the maximum over the 24 hours corresponding to the time frame of the ensemble forecast.

The results presented below are based on a verification period from 9 August 2010 to 30 April 2011, consisting of 263 individual forecast cases. Additionally, we use data from 1 February 2010 to 30 April 2011 to obtain training periods of equal length for all days in the verification period and for model selection purposes and forecasts from May 1, 2010 to 8 August 2010 (100 observations) as initial training period for the combination methods.

Following Lerch and Thorarinsdottir (2013), we consider two competing models: the truncated normal distribution (TN) and the generalized extreme value distribution (GEV). The TN model is estimated by minimizing the CRPS. The GEV model is estimated by maximum likelihood estimation.

First, we report in-sample results over the sample from May 1, 2010 to 30 April 2011. Then, we implement an out-of-sample exercise for the period from August 9, 2010 to 30 April 2011. We report both log score and CRPS results.

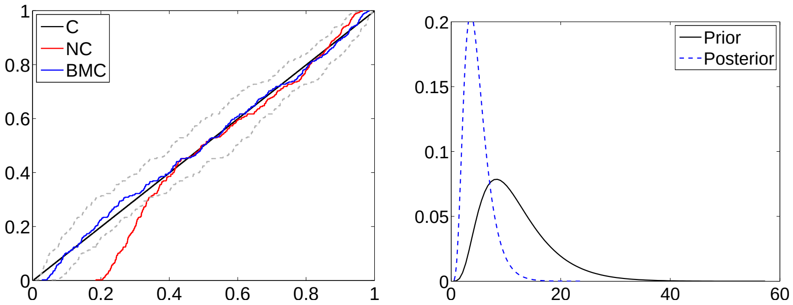

Figure 9: Infinite beta mixture calibrated (BMC), calibrated (C) and non calibrated (NC) combinations for the maximum wind speed at the station at Frankfurt airport. PITs (left top) of the combination models C (black), NC (red) and BMC (blue) and BMC 99% HPD (gray). Prior (black) and posterior (blue) number of components of the random BMC model (right top).

Table 3: Average log scores (AvLS) and average CRPS (AvCRPS) for the truncated normal (TN), the generalized extreme value distribution (GEV), linear pooling (NC) and beta mixture calibration (BMC) over the sample period from August 9, 2010 to 30 April 2011.

	TN	GEV	NC	BMC
AvLS	-2.812	-2.904	-2.433	-1.997
AvCRPS	1.346	1.802	1.314	0.982

Figure 9 reports in-sample calibration results. The BMC line is close to the ideal model and always includes the 45 degree line in the confidence interval. The NC performs poorly for small quantiles. The posterior density for the numbers of beta components in BMC is more concentrated than the prior, confirming also in this exercise that data are informative on the number of mixture components. When focusing on the OOS exercise, the BMC predictive distribution predicts accurately and provides the highest average LS and the lowest average CRPS in Table 3. Gains are substantial, as Figure 10 shows. The distribution is often multimodal, see Figure 11, with the highest mode at low values of wind speed, and a second mode concentrated around values of wind speed greater than 5 meters per second.

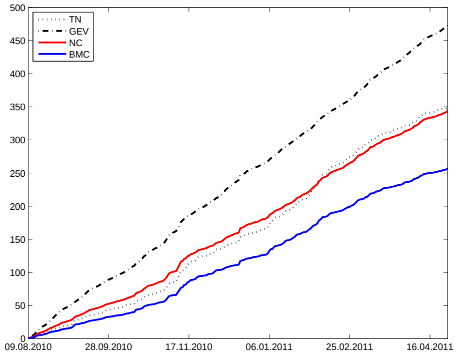

Figure 10: Cumulative CRPS for the truncated normal (TN), the the generalized extreme value distribution (GEV), linear pooling (NC) and beta mixture calibration (BMC) over the sample period from August 9, 2010 to 30 April 2011.

The truncated normal has too many values in the lower and upper tails; the GEV is too skewed to the upper tail, thus predicting on average too high values. The NC is also upper biased by the GEV. The BMC shifts the probability mass of the predictive distribution from the upper tail to the central part and the left tail, thus producing better calibrated forecasts.

7 Discussion

We propose a Bayesian approach to predictive density calibration accounting for parameter uncertainty. We build on the predictive density calibration and combination framework of Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) and propose the use of infinite mixtures of beta densities for the calibration. We rely upon the flexibility of the infinite beta mixtures to achieve a continuous deformation of the linear combination of predictive distributions. Each component of the beta mixture calibrates different parts of the predictive cdf and uses component-specific combination weights. Thanks to these features, our calibration model can also be viewed as a mixture of local combination models. Furthermore, our Bayesian framework allows for estimating the number of mixture components, the beta parameters and the predictive weights, and including various sources of uncertainty in the predictive density. We discuss properties of our methodology in simulation exercises, showing how the infinite beta

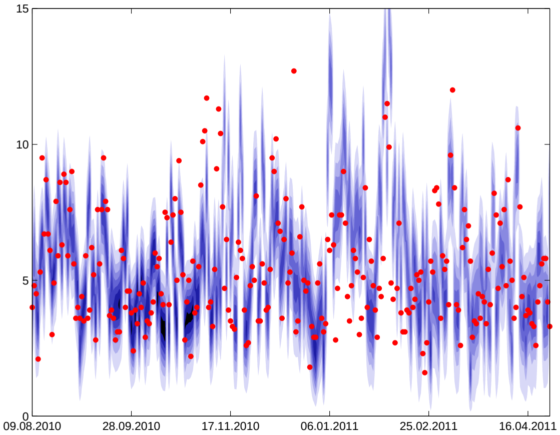

Figure 11: Fanchart of the BMC model and observations (red points) over the sample period from August 9, 2010 to 30 April 2011.

components are adequate in applications with multimodal densities and heavy tails. In empirical applications to stock returns and wind speed data, our infinite beta mixture approach provides well-calibrated and accurate density forecasts.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174.

Bassetti, F., Casarin, R., and Leisen, F. (2014). Beta-product dependent Pitman-Yor processes for Bayesian inference. *Journal of Econometrics*, 180:49–72.

Bates, J. M. and Granger, C. W. J. (1969). Combination of forecasts. *Operational Research Quarterly*, 20:451–468.

Billio, M. and Casarin, R. (2011). Beta autoregressive transition Markov-switching models for business cycle analysis. *Studies in Nonlinear Dynamics and Econometrics*, 15:1–32.

Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177:213–232.

Bouguila, N., Ziou, D., and Monga, E. (2006). Practical Bayesian estimation of a finite beta mixture through Gibbs sampling and its applications. *Statistics and Computing*, 16:215–225.

Burda, M., Harding, M., and Hausman, J. (2014). A Bayesian semiparametric competing risk model with unobserved heterogeneity. *Journal of Applied Econometrics*.

Casarin, R., Dalla Valle, L., and Leisen, F. (2012). Bayesian model selection for beta autoregressive processes. *Bayesian Analysis*, 7:1–26.

Chib, S. and Hamilton, B. H. (2002). Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110:67–89.

Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147:278–290.

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.

Epstein, E. S. (1966). Quality control for probability forecasts. *Monthly Weather Review*, 94:487–494.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89:268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Fawcett, N., Kapetanios, G., Mitchell, J., and Price, S. (2013). Generalised density forecast combinations. Technical report, Warwick Business School.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer-Verlag, Berlin.

Geweke, J. (2010). *Complete and Incomplete Econometric Models.* Princeton University Press, Princeton.

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). Optimal prediction pools. *Journal of Econometrics*, 164:130–141.

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics.* Springer Series in Statistics. Springer-Verlag, New York.

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

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

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3:197–204.

Griffin, J. E. (2011). Inference in infinite superpositions of non-Gaussian Ornstein-Uhlenbeck processes using Bayesian nonparametric methods. *Journal of Financial Econometrics*, 1:1–31.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101:179–194.

Griffin, J. E. and Steel, M. F. J. (2011). Stick-breaking autoregressive processes. *Journal of Econometrics*, 162:383–396.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23:1–13.

Hatjispyros, S. J., Nicoleris, T. N., and Walker, S. G. (2011). Dependent mixtures of Dirichlet processes. *Computational Statistics & Data Analysis*, 55:2011–2025.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica*, 70:781–799.

Hjort, N. L., Homes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87:371–390.

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

Jochmann, M. (2015). Modeling U.S. inflation dynamics: A Bayesian nonparametric approach. *Econometric Reviews*, 34(5):537–558.

Jore, A. S., Mitchell, J., and Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25:621–634.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21:93–105.

Kling, J. L. and Bessler, D. A. (1989). Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices, and output. *Journal of Business*, 62:477–499.

Lerch, S. and Thorarinsdottir, T. L. (2013). Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus Series A*, 65:21206.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12:351–357.

MacEachern, S. N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238.

Mitchell, J. and Wallis, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26:1023–1040.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society B*, 66:735–749.

Norets, A. and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168(332-346).

Norets, A. and Pelenis, J. (2015). Posterior consistency in conditional density estimation by covariate dependent mixtures. *EconometricTheory, forthcoming.*

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective Markov chain Monte Carlo for Dirichlet process hierarchical models. *Biometrika*, 95:169–186.

Pati, D., Dunson, D., and Tokdar, S. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, 178:624–638.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society Series B*, 72:71–91.

Robert, C. P. and Rousseau, J. (2002). A mixture approach to Bayesian goodness of fit. Technical Report 02009, CEREMADE, Université Paris-Dauphine.

Rodriguez, A. and ter Horst, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis*, 3:339–366.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley, Chichester.

Stone, M. (1961). The linear pool. *Annals of Mathematical Statistics*, 2:1339–1342.

Taddy, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking the intensity of violent crime. *Journal of the American Statistical Association*, 105:1403–1417.

Taddy, M. A. and Kottas, A. (2009). Markov switching Dirichlet process mixture regression. *Bayesian Analysis*, 4:793–816.

Tang, Y. and Ghosal, S. (2007). Posterior consistency of Dirichlet mixtures for estimating a transition density. *J. Statist. Plann. Inference*, 137(6):1711–1726.

Tang, Y., Ghosal, S., and Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, 63(4):1126–1134, 1312.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, 19:235–254.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics – Simulation and Computation*, 36:45–54.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society Series B*, 62:159–180.

Wiesenfarth, M., C.M., H., Kneib, T., and Cadarso-Suarez, C. (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. *Journal of Business and Economic Statistics*, 32(3):468–482.

Wu, Y. and Ghosal, S. (2009a). Correction to: "Kullback Leibler property of kernel mixture priors in Bayesian density estimation" [mr2399197]. *Electron. J. Stat.*, 3:316–317.

Wu, Y. and Ghosal, S. (2009b). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331.

A Computational details

A.1 Gibbs sampler for the finite beta mixture model

1. Full conditional distribution of D. Samples from the full conditional of D given $(\boldsymbol{\theta}, Y)$ are obtained by drawing sequentially over t, vectors $d_t = (d_{1t}, \ldots, d_{Kt})$ from multinomial distributions with probabilities

$$\pi(d_{kt} = 1|\boldsymbol{\theta}, Y) \propto w_k b^*_{\mu_k,\nu_k}(H(y_t|\boldsymbol{\omega}_k))\, h(y_t|\boldsymbol{\omega}_k)$$

for $k = 1, \ldots, K$.

2. Full conditional distribution of $(\boldsymbol{\mu}, \boldsymbol{\nu})$. Samples from the full conditional of $(\boldsymbol{\mu}, \boldsymbol{\nu})$ given $(\boldsymbol{w}, \boldsymbol{\omega}, D, Y)$ are obtained in a sequence of Metropolis-Hastings (MH) steps on a transformed space. Following Bouguila et al. (2006), we let: $\mu_k = 1/(1+\exp\{-\gamma_k\})$ and $\nu_k = \exp\{\lambda_k\}$, $k = 1, \ldots, K$ and draw iteratively from

$$\begin{aligned}&\pi(\gamma_k, \lambda_k|\boldsymbol{w}, \boldsymbol{\omega}, D, Y) \propto \\ &\prod_{t\in\mathcal{D}_k} b^*_{\mu_k,\nu_k}(H(y_t|\boldsymbol{\omega}_k))\, \mu_k^{\xi_{1\mu}-1}(1-\mu_k)^{\xi_{2\mu}-1}\nu_k^{\xi_{1\nu}-1}\exp\{-\xi_{2\nu}\nu_k\}J(\mu_k, \nu_k),\end{aligned}$$

where $J(\mu_k, \nu_k) = \exp\{-\gamma_k - \lambda_k\}(1 + \exp\{-\gamma_k\})^{-2}(1 + \exp\{-\lambda_k\})^{-2}$ is the Jacobian of the transform. In the MH we use a Gaussian random walk proposal distribution with covariance matrix $\Sigma = 0.05 I_2$, which yields acceptance rates of about 0.4.

3. Full conditional distribution of $\boldsymbol{\omega}$. Samples from the full conditional of $\boldsymbol{\omega}$ given $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, D, Y)$ are obtained by drawing iteratively $\boldsymbol{\omega}_k$, $k = 1, \ldots, K$. At each step we apply a MH with the prior distribution as proposal. The acceptance probability of each MH step is:

$$\min\left\{\prod_{t\in\mathcal{D}_k}\frac{b^*_{\mu_k,\nu_k}(H(y_t|\boldsymbol{\omega}^*))h(y_t|\boldsymbol{\omega}^*)}{b^*_{\mu_k,\nu_k}(H(y_t|\boldsymbol{\omega}_k))h(y_t|\boldsymbol{\omega}_k)}, 1\right\},$$

where $\boldsymbol{\omega}^* \sim \mathcal{D}ir(\xi_\omega, \ldots, \xi_\omega)$.

4. Full conditional distribution of $\boldsymbol{w}$. Samples from the full conditional of $\boldsymbol{w}$ given $(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\nu}, D, Y)$ are obtained by exploiting the conjugacy of the prior distribution, in that

$$\pi(w_1, \ldots, w_k | \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\nu}, D, Y) \propto \mathcal{D}ir(\xi_w + T_1, \ldots, \xi_w + T_k).$$

When $K = 1$, we replace the single-move Gibbs sampler with a global MH sampler with target distribution obtained by applying to the joint posterior

$$\pi(\mu, \nu, \omega | Y) \propto \prod_{t=1}^{T} b^*_{\mu,\nu} \left(H(y_t|\omega)\right) h(y_t|\omega) \mu^{\xi_{1\mu}-1}(1-\mu)^{\xi_{2\mu}-1}$$
$$\times\, \nu^{\xi_{1\nu}-1} \exp\{-\xi_{2\nu}\nu\} \omega^{\xi_{1\omega}-1}(1-\omega)^{\xi_{2\omega}-1}$$

where $Y = (y_1, \ldots, y_t)$, the change of variable $\mu = 1/(1 + \exp\{-\theta_1\})$, $\nu = \exp\{\theta_2\}$ and $\omega = 1/(1 + \exp\{-\theta_3\}$. We consider a random walk proposal on the transformed parameter space accounting for the Jacobian of the transformation, that is, $J = \exp\{\theta_2 - \theta_1 - \theta_3\}(1 + \exp\{-\theta_1\})^{-2}(1 + \exp\{-\theta_3\})^{-2}$. Setting the covariance matrix to $\Sigma = \text{diag}\{0.1, 0.05, 0.1\}$, we achieve acceptance rates of about 0.4.

A.2 Gibbs sampler for the infinite beta mixture model

Let $\mathcal{D}_k = \{t = 1, \ldots, T | d_t = k\}$ denote the set of indexes of the observations allocated to the k-th component of the mixture and with $\mathcal{D} = \{k | \mathcal{D}_k \neq \emptyset\}$ the set of indexes of the non-empty mixture components. Then the cardinality of $\mathcal{D}$, $Card(\mathcal{D})$, gives the number of mixture components and $D^* = \sup \mathcal{D}$ can be interpreted as the number of stick-breaking components used in the mixture. As noted by Kalli et al. (2011), the sampling of an infinite numbers of Θ and V is not necessary, since only the elements in the full conditional pdfs of D are needed. The maximum number of atoms and stick-breaking components to sample is $N^* = \max\{t = 1, \ldots, T | N^*_t\}$, where N^*_t is the smallest integer such that $\sum_{j=1}^{N^*_t} w_j > 1 - u_t$. Thus sampling from the joint $\pi(V, U | \Theta, D, Y, \psi)$ is obtained by splitting $V = (V^*, V^{**})$, where $V^* = (v_1, \ldots, v_{D^*})$ and $V^{**} = (v_{D^*+1}, \ldots, v_{N^*})$, and by further collapsing the Gibbs, that is by sampling from $\pi(V^* | \Theta, D, Y, \psi)$ and $\pi(U | V^*, \Theta, D, Y, \psi)$ and then from $\pi(V^{**} | V^*, U, \Theta, D, Y, \psi)$.

1. Full conditional distribution of V^*. Sampling from the full conditional of V^* given (D, Θ, Y, ψ) is obtained by drawing v_k, with $k \leq D^*$, from the full conditionals

$$\pi(v_k | D, Y) \propto (1 - v_k)^{\psi + b_k - 1} v_k^{a_k},$$

that is, the PDF of a $\mathcal{B}e(a_k+1, b_k+\psi)$ with $a_k = \sum_{t=1}^{T} \mathbb{1}_{\{d_t=k\}}$ and $b_k = \sum_{t=1}^{T} \mathbb{1}_{\{d_t>k\}}$.

2. Full conditional distribution of U. Samples from the full conditional of U given (V, D, Θ, Y, ψ) is obtained by simulating from the uniform

$$\pi(u_t|V, D, Y) \propto \frac{1}{w_{d_t}} \mathbb{1}_{\{u_t<w_{d_t}\}}$$

for $t = 1, \ldots, T$.

3. Full conditional distribution of V^{**}. Sampling from the full conditional of V^{**} given $(V^*, U, D, \Theta, Y, \psi)$ is obtained by sampling from

$$\pi(v_k|U, D, Y) \propto (1 - v_k)^{\psi-1},$$

that is, the PDF of a $\mathcal{B}e(1, \psi)$, with $k = D^* + 1, \ldots, N^*$.

4. Full conditional distribution of Θ. Sample the elements k, $k = 1, \ldots, N^*$, of Θ given (U, D, V, Y, ψ), from the full conditional

$$\begin{aligned}\pi(\boldsymbol{\theta}_k|U, D, V, Y) &\propto \prod_{t\in\mathcal{D}_k} b^*_{\mu_k,\nu_k}(H(y_t|\boldsymbol{\omega}_k))\, h(y_t|\boldsymbol{\omega}_k)\\ &\times \mu_k^{\xi_\mu-1}(1-\mu_k)^{\xi_\mu-1}\nu_k^{\xi_\nu/2}\exp\{-\xi_\nu\nu_k/2\}\prod_{i=1}^{M}\omega_{ik}^{\xi_\omega-1}\mathbb{1}_{\{\boldsymbol{\omega}_k\in\Delta_M\}}\end{aligned}$$

for $k \in \mathcal{D}$, and from the prior G_0 for $k \notin \mathcal{D}$. We sample from the full conditional by iterating over the following steps:

(a) $\pi(\mu_k, \nu_k|\boldsymbol{\omega}_k, U, D, V, Y, \psi)$

(b) $\pi(\boldsymbol{\omega}_k|\mu_k, \nu_k, U, D, V, Y, \psi)$.

We apply here the same sampling strategy described for the parameter of the finite beta mixture model.

5. Full conditional distribution of D. Samples from the full conditional of D given (V, U, Θ, Y, ψ) are obtained by sampling from

$$\pi(d_t|V, U, Y) \propto \mathbb{1}_{\{u_t<w_{d_t}\}} f_{\mu_{d_t},\nu_{d_t}}(H(y_t|\boldsymbol{\omega}_{d_t}))\, h(y_t|\boldsymbol{\omega}_{d_t}),$$

with $d_t \in \{1, \ldots, N_t^*\}$, where N_t^* is defined above.

6. Full conditional distribution of ψ. If the dispersion parameter ψ is assumed to be random with $\mathcal{G}a(c,d)$ prior, then an extra step is needed in the Gibbs sampler. More specifically, the full conditional distribution of ψ given U, D, V and Θ has density

$$\pi(\psi|K,T) \propto B(\psi,T)\psi^{K+c-1}\exp\{-d\psi\}\mathbb{1}_{\psi\in(0,+\infty)},$$

which depends only on the number of observations T and the number of mixture components N^*, which has been defined above.

The Gibbs sampler can used to generate draws from the predictive distribution $\hat{F}_{T+1}(y_{T+1})$. At each iteration a uniform random variable $u^{(i)}$ is sampled from the unit interval and $\boldsymbol{\theta}_j^{(i)}$ is used such that $w_{j-1}^{(i)} < u^{(i)} < w_j^{(i)}$. If $j > N^{*\,(i)}$, then more weights are required than currently exist, and they can be sampled from $\mathcal{B}e(1,\psi)$ and the additional $\boldsymbol{\theta}_j^{(i)}$ from G_0. Having taken $\boldsymbol{\theta}_j^{(i)}$, $y_{T+1}^{(i)}$ can be sampled from $B^*_{\mu_j^{(i)},\nu_j^{(i)}}\Big(H(y_{T+1}|\boldsymbol{\omega}_j^{(}i))\Big)$.

B Proofs of the results of Section 4

The proof of Theorem 4.1 is based on an application of Theorem 1 and Lemma 3 of Wu and Ghosal (2009a,b). For the shake of clarity we report the statements of these results in Theorems B.1 and B.2 below.

To prove that $f_0 \in KL(\Pi^*)$, Wu and Ghosal (2009a) suggest to split the problem in two steps, as shown in the next very simple theorem.

Theorem B.1 (Thm. 1 of Wu and Ghosal (2009a))**.** *If for any ϵ there is a probability measure G_ϵ and a measurable set $\mathcal{W} \subset \mathcal{M}(\Theta)$, with $G_\epsilon \in \mathcal{W}$ and $\Pi(\mathcal{W}) > 0$, such that*

(H1) $\int \log(f_0/f_{G_\epsilon})f_0 < \epsilon$,

(H2) $\int \log(f_{G_\epsilon}/f_G)f_0 < \epsilon$ *for every G in $\mathcal{W}$;*

then $f_0 \in KL(\Pi^*)$.

Proof. Since

$$\int \log\big(\frac{f_0(y)}{f_G(y)}\big)f_0(y) = \int \log\big(\frac{f_0(y)}{f_{G_\epsilon}(y)}\big)f_0(y)dy + \int \log\big(\frac{f_{G_\epsilon}(y)}{f_G(y)}\big)f_0(y)dy,$$

one has

$$\Pi^*\Big\{\int \log(f_0(y)/f_G(y))f_0(y)dy < 2\epsilon\Big\}$$
$$\geq \Pi^*\Big\{\int \log(f_{G_\epsilon}(y)/f_G(y))f_0(y)dy < \epsilon\Big\} \geq \Pi(\mathcal{W}) > 0.$$

□

Given a probability G_ϵ satisfying (H1), in general some work is need to verify assumption (H2). A useful sufficient condition is contained in Lemma 3 of of Wu and Ghosal (2009a).

Theorem B.2 (Lemma 3 of Wu and Ghosal (2009a))**.** *Let* Θ *be a separable metric space. If for any* $\epsilon > 0$ *there is a probability measure* $G_\epsilon \in supp(\Pi)$ *such that* (H1) *holds and there is a closed set* D_ϵ *such that*

(H3) D_ϵ *contains* $\sup(G_\epsilon)$ *in its interior and*

$$\int \log\Big(\frac{f_{G_\epsilon}(y)}{\inf_{\theta\in D_\epsilon} K(y;\theta)}\Big)f_0(y)dy < +\infty,$$

(H4) $\inf_{y\in C}\inf_{\theta\in D_\epsilon} K(y;\theta) > 0$ *for every compact set* $C \subset \mathcal{Y}$*,*

(H5) $\{\theta \mapsto K(y;\theta) : y \in C\}$ *is uniformly equicontinuous on* D_ϵ*,*

then (H2) *holds for a suitable* $\mathcal{W} \subset \mathcal{M}(\Theta)$ *with* $\Pi(\mathcal{W}) > 0$*, and hence* $f_0 \in KL(\Pi^*)$*.*

Assumption (H1) –(H2), respectively– corresponds to (A1) –(A3), respectively– in Theorem 1 of Wu and Ghosal (2009a). Note the Theorem Wu and Ghosal (2009a) is stated for Type II prior and it has an additional assumption (A2), which is not needed for Type I prior. Note also that assumptions (H3)-(H4) correspond to assumptions (A7)-(A8) of Lemma 3 of Wu and Ghosal (2009a), while (H5) is slightly different from the original assumption (A9), see Wu and Ghosal (2009b).

Proof of Theorem 4.1. Here we need to think Δ_M as the set $\{(\omega_1,\dots,\omega_{M-1}) \in [0,1]^{M-1} : \sum_{i=1}^{M-1}\omega_i \leq 1\}$ endowed with the topology induced by the euclidean norm. Clearly, ω_M will denote $1-\sum_{i=1}^{M-1}\omega_i$.

Verification of H1 of Theorem B.1. Since u_0 is continuous on $(0,1)$ and $\int_0^1 \log(u_0(x))u_0(x)dx < +\infty$ by Theorem 1 in Robert and Rousseau (2002) there is

$$u_\epsilon(x) = u_{\tilde{P}_\epsilon}(x) = \sum_{i=1}^{K_\epsilon} w_{i,\epsilon} b^*_{\mu_{i,\epsilon},\nu_{i,\epsilon}}(x),$$

where $\tilde{P}_\epsilon(d\mu d\nu) := \sum_{i=1}^{K_\epsilon} w_{i,\epsilon}\delta_{\mu_{i,\epsilon},\nu_{i,\epsilon}}(d\mu d\nu)$, such that $KL(u_0, u_\epsilon) \leq \epsilon$. If $G_\epsilon(d\boldsymbol{\omega} d\mu d\nu) := \delta_{\boldsymbol{\omega_0}}(d\boldsymbol{\omega}) \times \tilde{P}_\epsilon(d\mu d\nu)$, then

$$f_{G_\epsilon}(y) = \int b^*_{\mu,\nu}(H(y|\boldsymbol{\omega}))h(y|\boldsymbol{\omega})G_\epsilon(d\boldsymbol{\omega} d\mu d\nu) = u_\epsilon(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0}).$$

By a simple change of variables,

$$\begin{aligned} KL(f_0, f_{G_\epsilon}) &= \int u_0(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0}) \log\Big(\frac{u_0(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0})}{u_\epsilon(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0})}\Big)dy \\ &= \int_0^1 u_0(z)\log\Big(\frac{u_0(z)}{u_\epsilon(z)}\Big)dz. \end{aligned}$$

That is

$$KL(f_0, f_{G_\epsilon}) = KL(u_0, u_\epsilon) \leq \epsilon.$$

Note that $supp(G_\epsilon) = \{\boldsymbol{\omega_0}\} \times \cup_{i=1}^{K_\epsilon}\{(\mu_{i,\epsilon}, \nu_{i,\epsilon})\}$ and, since G_0 has full support, $G_\epsilon \subset supp(Dir(\psi, G_0))$.

Verification of H3 of Theorem B.2. One can find a compact set D^*_ϵ in $(0,1) \times (0, +\infty)$ such that D^*_ϵ contains $\cup_{i=1}^{K_\epsilon}\{(\mu_{i,\epsilon}, \nu_{i,\epsilon})\}$ in its interior. Moreover, recalling that $h(y|\boldsymbol{\omega}) = \sum_{i=1}^M \boldsymbol{\omega}_i f_i(y)$ and that $\boldsymbol{\omega_0}$ is in the interior of Δ_M, one can find a (sufficiently small) compact set $\Delta^*_\epsilon \subset \Delta_M$ containing $\boldsymbol{\omega_0}$ in its interior such that if $\boldsymbol{\omega} \in \Delta^*_\epsilon$ then $C_{1,\epsilon}h(y|\boldsymbol{\omega_0}) \leq h(y|\boldsymbol{\omega}) \leq C_{2,\epsilon}h(y|\boldsymbol{\omega_0})$ for every y. It follows that $D_\epsilon = \Delta^*_\epsilon \times D^*_\epsilon$ is a compact set containing $supp(G_\epsilon)$ in its interior. Noticing that if $\boldsymbol{\omega} \in \Delta^*_\epsilon$ then $C_{2,\epsilon}H(y|\boldsymbol{\omega_0}) \geq H(y|\boldsymbol{\omega}) \geq C_{1,\epsilon}H(y|\boldsymbol{\omega_0})$ and $C_{2,\epsilon}(1 - H(y|\boldsymbol{\omega_0})) \geq (1 - H(y|\boldsymbol{\omega})) \geq C_{1,\epsilon}(1 - H(y|\boldsymbol{\omega_0}))$, one can write

$$\begin{aligned} I_\epsilon(y) &:= \inf_{(\boldsymbol{\omega},\mu,\nu)\in D_\epsilon} K(y; \boldsymbol{\omega}, \mu, \nu) \\ &= \inf_{(\boldsymbol{\omega},\mu,\nu)\in D_\epsilon} h(y|\boldsymbol{\omega})\frac{H(y|\boldsymbol{\omega})^{\mu\nu-1}(1 - H(y|\boldsymbol{\omega}))^{(1-\mu)\nu-1}}{B(\mu\nu, (1-\mu)\nu)} \\ &\geq C_{3,\epsilon}h(y|\boldsymbol{\omega_0})H(y|\boldsymbol{\omega_0})^{A_\epsilon-1}(1 - H(y|\boldsymbol{\omega_0}))^{B_\epsilon-1} =: I^*_\epsilon(y) \end{aligned}$$

where $C_{3,\epsilon} = C_{1,\epsilon}C_{2,\epsilon}^{-2}\inf\{C_{1,\epsilon}^{\mu\nu+(1-\mu)\nu}/B(\mu\nu, (1-\mu)\nu) : (\mu,\nu) \in D^*_\epsilon\} > 0$, $A_\epsilon = \sup\{\mu\nu : (\mu,\nu) \in D^*_\epsilon\} > 0$ and $B_\epsilon = \sup\{(1-\mu)\nu : (\mu,\nu) \in D^*_\epsilon\} > 0$. Hence, one the one hand $f_{G_\epsilon}(y) \geq I_\epsilon(y)$ and hence $\log(f_{G_\epsilon}(y)/I_\epsilon(y)) \geq 0$, on the other hand

$$\begin{aligned} \int \log\Big(\frac{f_{G_\epsilon}(y)}{I_\epsilon(y)}\Big)f_0(y)dy &\leq \int \log\Big(\frac{f_{G_\epsilon}(y)}{I^*_\epsilon(y)}\Big)f_0(y)dy \\ &\leq \int_0^1 \log\Big(\frac{u_\epsilon(x)}{x^{A_\epsilon-1}(1-x)^{B_\epsilon-1}}\Big)u_0(x)dx + |\log(C_{3,\epsilon})|. \end{aligned}$$

Since $C_{4,\epsilon}x^{A'_\epsilon-1}(1-x)^{B'_\epsilon-1} \leq u_\epsilon(x) \leq C_{5,\epsilon}x^{A''_\epsilon-1}(1-x)^{B''_\epsilon-1}$ for suitable constants, it follows that

$$\int \Big|\log\Big(\frac{u_\epsilon(x)}{x^{A_\epsilon-1}(1-x)^{B_\epsilon-1}}\Big)\Big|u_0(x)dx \leq C_{6,\epsilon}\int[|\log(x)|+|\log(1-x)|]u_0(x)dx < +\infty$$

by assumption (17). Hence

$$0 < \int \log\Big(\frac{f_{G_\epsilon}(y)}{\inf_{(\boldsymbol{\omega},\mu,\nu)\in D_\epsilon} K(y;\boldsymbol{\omega},\mu,\nu)}\Big)f_0(y)dy < +\infty.$$

Verification of H4 of Theorem B.2. It follows immediately that, for every compact set C,

$$\inf_{y\in C}\inf_{(\boldsymbol{\omega},\mu,\nu)\in D_\epsilon} K(y;\boldsymbol{\omega},\mu,\nu) \geq \inf_{y\in C} I^*_\epsilon(y)$$

and the right hand side is strictly positive by (18).

Verification of H5 of Theorem B.2. The function $(\boldsymbol{\omega},\mu,\nu,y)\mapsto K(y;\boldsymbol{\omega},\mu,\nu)$ is continuous and hence uniformly continuous on the compact set $C\times D_\epsilon$. It follows that the family $\{(\boldsymbol{\omega},\mu,\nu)\mapsto K(y;\boldsymbol{\omega},\mu,\nu) : y\in C\}$ is uniformly equicontinuous on D_ϵ. □

Proof of Theorem 4.2. The proof is a simple modification of the proof of Theorem 4.1. Note that here the assumption that $\boldsymbol{\omega_0}$ belongs to the interior of Δ_M is not needed. □

Proof of Theorem 4.3. Given any measure Q on $(0,1)\times\mathbb{R}^+$ recall that $f_Q(x) = u_Q(H(y|\boldsymbol{\omega_0}))h(y|\boldsymbol{\omega_0})$ where $u_Q(x) = \int b^*_{\mu,\nu}(x)Q(d\mu d\nu)$. Again, by a simple change of variables, $KL(f_G,f_0) = KL(u_G,u_0)$. Hence to prove that $f_0\in KL(\Pi^*)$ it suffices to prove that for every $\epsilon>0$, $P\{KL(u_G,u_0)\leq\epsilon\}>0$.

Now recall that if $G\sim DP(\psi,G_0)$ then G admits the representation $G = w_1\delta_{\theta_1}+(1-w_1)G_1$ where $w_1,\theta_1=(\mu_1,\nu_1)$ and G_1 are stochastically independent, $G_1\sim DP(\psi,G_0)$, $w_1\sim Beta(1,\phi_1)$ and $\theta_1\sim G_0$.

Given $\eta,\eta'>0$ define $\mathcal{U}_\eta := \{(w,\mu,\nu)\in(0,1)^2\times\mathbb{R}^+ : |w-w_0|\leq\eta, |\mu-\mu_0|\leq\eta, |\nu-\nu_0|\leq\eta\}$ and $\mathcal{U}^*_{\eta,\eta'} := \{G = w_1\delta_{\theta_1}+(1-w_1)G_1 : (w_1,\theta_1)\in\mathcal{U}_\eta, |u_{G_1}-u_{P_0}|_1\leq\eta'\}$, where we denote by $|u_1-u_2|_1 = \int|u_1-u_2|dx$ be the L_1 distance between two densities u_1 and u_2.

Note that if $G\in\mathcal{U}^*_{\eta,\eta'}$ then

$$u_G(x)\geq w_1b^*_{\mu_1,\nu_1}(x)\geq c_\eta x^{A_\eta}(1-x)^{B_\eta}$$

where

$$c_\eta := \frac{w_0 - \eta}{B((\mu_0 - \eta)(\nu_0 - \eta), (1 - \mu_0 - \eta)(\nu_0 - \eta))},$$
$$A_\eta := (\mu_0 + \eta)(\nu_0 + \eta) - 1, \qquad B_\eta := (1 - \mu_0 + \eta)(\nu_0 + \eta) - 1,$$

provided that $\mu_0 - \eta, 1 - \mu_0 - \eta, \nu_0 - \eta, w_0 - \eta$ are positive. Hence, for any $\zeta > 0$,

$$\Big[\frac{u_0(x)}{u_G(x)}\Big]^\zeta \le c^*_{\eta,\zeta} \frac{u_0(x)^\zeta}{x^{A_\eta \zeta}(1-x)^{B_\eta \zeta}} =: g^*(x)$$

for a suitable constant $c^*_{\eta,\zeta}$. By assumption (20), there is ζ such that $C_0 := \int g^*(x) u_0(x) dx < +\infty$. Hence for such (η, ζ), by Lemma 7 of Ghosal and van der Vaart (2007),

$$KL(u_0, u_G) \le C_1 d_H^2(u_G, u_0)[1 + \max(0, \log(d_H^{-1}(u_G, u_0)))]$$

where $d_H(u_G, u_0) = (\int (\sqrt{u_G} - \sqrt{u_0})^2 dx)^{1/2}$ is the Hellinger distance between u_0 and u_G. Note that the constant C_1 depends on C_0, η and ζ only. Since $d_H(u_G, u_0)^2 \le |u_G - u_0|_1$ (see, e.g., Corol.1.2.1 in Ghosh and Ramamoorthi (2003)) it follows that

$$KL(u_0, u_G) \le C_2 |u_G - u_0|_1^{1/2}$$

for every $G \in \mathcal{U}^*_{\eta,\eta'}$ when η' is small enough. Now, it is easy to check that

$$|u_G - u_0|_1 \le 2|w_1 - w_0| + |u_{\delta_{\theta_1}} - u_{\delta_{\theta_0}}|_1 + |u_{G_1} - u_{P_0}|_1$$

and that $|u_{\delta_{\theta_1}} - u_{\delta_{\theta_0}}|_1$ goes to zero as $|\theta_1 - \theta_0| \to 0$. Since if $\eta'' \le \eta$ then $\mathcal{U}^*_{\eta'',\eta'} \subset \mathcal{U}^*_{\eta,\eta'}$, using the previous results, for every $\epsilon > 0$, one can find sufficiently small η' and $\eta'' \le \eta$ such that if $G \in \mathcal{U}^*_{\eta'',\eta'}$ then $KL(u_G, u_0) \le \epsilon$. By standard argument (see e.g. the proof of Thm.2 in Tang et al. (2007)) if G_1 is in a sufficiently small weak neighbourhood $V_{\eta'}$ of P_0 then $|u_{P_0} - u_{G_1}|_1 \le \eta'$, hence

$$\{G = w_1\delta_{\theta_1} + (1 - w_1)G_1 : G_1 \in V_{\eta'}, (w_1, \theta_1) \in \mathcal{U}_{\eta''}\} \subset \mathcal{U}^*_{\eta'',\eta'} \subset \{KL(u_G, u_0) \le \epsilon\}.$$

Moreover, $supp(P_0) \subset supp(G_0)$ yields that P_0 belongs to the support of $Dir(\phi, G_0)$ and hence $P(G_1 \in V_{\eta'}) > 0$, while the fact that $\theta_1 \in supp(G_0)$ yields that $P((w_1, \theta_1) \in \mathcal{U}_{\eta''}) > 0$. Using the independence of (w_1, θ_1) and G_1, one concludes

$$P(KL(u_0, u_G) \le \epsilon) \ge P(G \in \mathcal{U}^*_{\eta'',\eta'}) \ge P(G_1 \in V_{\eta'}) P((w_1, \theta_1) \in \mathcal{U}_{\eta''}) > 0.$$

□