2016/50

Bayesian Semiparametric Forecasts of Real Interest Rate Data

PHILIPPE J. DESCHAMPS

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
http://www.uclouvain.be/en-44508.html

BAYESIAN SEMIPARAMETRIC FORECASTS OF REAL INTEREST RATE DATA

Philippe J. Deschamps

Université de Fribourg, Switzerland; Université Catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium

November 2016

Abstract. The non-hierarchical Dirichlet process prior has been mainly used for parameters of innovation distributions. It is, however, easy to apply to all the parameters (coefficients of covariates and innovation variance) of more general regression models. This paper investigates the predictive performance of a simple (non-hierarchical) Dirichlet process mixture of Gaussian autoregressions for forecasting monthly US real interest rate data. The results suggest that the number of mixture components increases sharply over time, and the predictive marginal likelihoods strongly dominate those of a benchmark autoregressive model. Unconditional predictive coverage is vastly improved in the mixture model.

Faculté des Sciences Economiques et Sociales, Boulevard de Pérolles 90, CH-1700 Fribourg, Switzerland. Telephone: +41-26-300-8200. Telefax: +41-26-300-9725.

E-mail address: philippe.deschamps@unifr.ch

JEL classification: C11, C14, C22, C53.

Key words and phrases. Dirichlet process mixture, Bayesian nonparametrics, structural change, real interest rate.

The author thanks Luc Bauwens for helpful comments.

1. Introduction

Discrete mixtures of autoregressions have a long history in econometrics. Most contributions have assumed that the latent mixture indicators follow a first-order Markov process. The pioneering paper by Goldfeld and Quandt (1973) was followed by several frequentist applications (e.g. Hamilton, 1989; Garcia and Perron, 1996) and applications using the Bayesian paradigm. The latter category includes regime-switching models (e.g. Albert and Chib, 1993; Chib, 1996; Deschamps, 2006, 2008) and change-point models, where there is no possibility of reverting to a previous regime; see, for instance, Chib (1998), Pesaran et al. (2006), and Koop and Potter (2007). In this context, Koop and Potter (2007) insist on the importance of a formulation where the number of regimes is not restricted ex ante. They achieve this objective with a Poisson prior on regime durations. In the limit, their model reduces to a standard state space model where the parameters in an autoregression follow random walks; in this case, the number of regimes is equal to the number of time periods.

The present paper follows a different tradition. Rather than assuming that the mixture indicators follow a Markov process, it assumes a Dirichlet process (DP) prior (Ferguson, 1973) on all the parameters of a Gaussian autoregression. The resulting DP mixture model allows for a number of components that evolves over time, but does not exclude reverting to a past regime; in this sense, it may be both more flexible and more parsimonious than previous change-point formulations.

A disadvantage of the DP prior is that the individual mixture components are not identifiable. However, this is of no consequence when the model is used for predictive rather than explanatory purposes. The predictive is a mixture of K Gaussian densities, and of an additional Student density which allows for a possible additional (K + 1st) regime. Since this mixture is labeling-invariant, identification is not an issue when the model is used only for forecasting.

Section 2 of this paper presents the model and briefly discusses the DP prior used in this context. Section 3 gives a complete discussion of the estimation method. This is a straightforward multivariate generalization of the Gibbs sampler described in Escobar and West (1995). Section 4 describes an application to monthly US real interest rate data ranging from 1948 to 2015, and Section 5 concludes.

Several papers have successfully modeled the real interest rate for less extended

time periods, using other approaches in the literature. Three-regime Markov switching autoregressions were used by Garcia and Perron (1996) for quarterly data covering 1961 to 1986; Deschamps (2006) uses similar models with quarterly and monthly data ranging from 1953 to 2002. Giordani and Kohn (2008) use a mixture innovation formulation, where quarterly real interest rate data ranging from 1952 to 2004 are described by a state space model involving discrete shocks on the observation and evolution equation variances. All these contributions have involved models with fixed numbers of mixture components.

More recently, other papers (Fox et al., 2011; Song, 2014; Bauwens et al., 2016) have proposed hidden Markov models where the number of regimes is potentially infinite. Estimating these models requires formulating a hierarchical DP prior (Teh et al., 2006) and involves MCMC methods that are more complicated than the simple Gibbs sampler used here.

Our simple DP mixture model is not intended as an alternative to these more elaborate formulations, but rather as an easily implemented benchmark against which more sophisticated models can be evaluated. As will be seen, the model of Section 2 leads to vastly improved predictive coverage when compared to the usual benchmark, which is a single-regime Gaussian AR or ARMA model.

2. Dirichlet process priors and mixtures of autoregressions

Consider the following Gaussian autoregression with time-varying parameters:

$$\Phi_t(L)y_t = \mu_t + \sigma_t\epsilon_t \quad \text{for } t = 1, \ldots, T \tag{2.1}$$

where $\epsilon_t \sim \text{Niid}(0,1)$, $\Phi_t(L) = 1 - \phi_t^1 L - \cdots - \phi_t^p L^p$ is a polynomial in the lag operator L, and $y_0, y_{-1}, \ldots, y_{-p+1}$ are treated as fixed. Let $\theta_t = (\mu_t, \phi_t^1, \ldots, \phi_t^p, \sigma_t^{-2})$. Our purpose in this section is to investigate some implications of a Dirichlet process (DP) prior on θ_t. This prior is defined by:

$$\begin{aligned} \theta_t \mid G &\sim G \\ G \mid G_0, \alpha &\sim DP(\alpha, G_0) \end{aligned}$$

where G_0 admits the Normal-Gamma density:

$$g_0(\theta \mid m, V, s^2, \nu) = f_N(\mu, \phi^1, \ldots, \phi^p \mid m, \sigma^2 V) f_G(\sigma^{-2} \mid s^{-2}, \nu) \tag{2.2}$$

$f_N(\bullet \mid m, \Sigma)$ being the multivariate Normal density with expectation m and covariance matrix Σ and $f_G(\bullet \mid a, b)$ being the Gamma density with expectation a and variance $2a^2/b$ (see, e.g., Koop, 2003, p. 326).

Following Escobar and West (1995), we add some flexibility to this prior by putting an independent Normal-Wishart hyperprior on m and V^{-1} and a Gamma hyperprior on α, as follows:

$$p(m) = f_N(m \mid \underline{m}, \underline{V}) \tag{2.3}$$

$$p(V^{-1}) = f_W(V^{-1} \mid \underline{\nu}, \underline{H}) \tag{2.4}$$

$$p(\alpha) = f_G(\alpha \mid m_\alpha, \nu_\alpha) \tag{2.5}$$

where $f_W(\bullet \mid \underline{\nu}, \underline{H})$ denotes a Wishart density with $\underline{\nu}$ degrees of freedom and scale matrix $\underline{H}$. The joint prior on (m, V) will be proper if $\underline{\nu} > p + 1$ and $\underline{H}, \underline{V}$ are nonsingular.

The DP prior was introduced by Ferguson (1973). The random distribution G has the property that, for any partition $A_1, \ldots, A_K$ of the parameter space, the vector $(G(A_1), \ldots, G(A_K))$ has a Dirichlet distribution with parameters $\alpha G_0(A_1), \ldots, \alpha G_0(A_K)$. G_0 is known as the base distribution and α is called the total mass. Using the moments of the Dirichlet distribution, it is easy to verify that G converges to G_0 in distribution as $\alpha \to \infty$.

Ferguson (1973) shows that G is almost surely discrete; Sethuraman (1994) shows that if $\theta_t \sim G$, its probability function can be represented as the following infinite mixture with random weights v_j and random knots θ_j^*:

$$g(\theta_t) = \sum_{j=1}^{\infty} v_j \mathcal{I}_{\theta_j^*}(\theta_t) \tag{2.6}$$

where $\mathcal{I}_{\theta_j^*}(\theta_t) = 1$ if $\theta_t = \theta_j^*$, 0 otherwise, the θ_j^* are independently distributed with common distribution G_0, and the v_j are defined as:

$$v_1 = w_1, \quad v_2 = w_2(1 - w_1), \quad \ldots \quad v_j = w_j \prod_{s=1}^{j-1}(1 - w_s)$$

the w_j being independently distributed as Beta$(1, \alpha)$. This is known as the "stick-breaking" representation. It is readily verified that the v_j add up to unity;

since they are geometrically decreasing in a probabilistic sense, the model favors parsimony. The degree of parsimony depends on the parameter α.

Neal (2000) gives an equivalent representation of the DP prior which is perhaps closer to a model with switching regimes. Consider the finite discrete mixture with K components (or classes):

$$g_K(\theta_t) = \sum_{j=1}^{K} p_j \mathcal{I}_{\theta_j^*}(\theta_t)$$

where the θ_j^* are independently distributed with common distribution G_0, and where $(p_1, \ldots, p_K) \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$. Upon integrating with respect to the p_j, Neal (2000) shows that the conditional probability that $\theta_t = \theta_j^*$ is:

$$P(c_t = j \mid c_1, \ldots, c_{t-1}) = \frac{n_{tj} + \frac{\alpha}{K}}{t - 1 + \alpha} \tag{2.7}$$

where $c_t = j$ iff $\theta_t = \theta_j^*$ and n_{tj} is the number of c_s that are equal to j for $s < t$. Let now K_t be the number of non-empty classes at time $t - 1$. Upon letting $K \to \infty$ in (2.7), we obtain:

$$\begin{aligned} P(c_t = j \mid c_1, \ldots, c_{t-1}) &= \frac{n_{tj}}{t - 1 + \alpha} \quad \text{for } 1 \leq j \leq K_t; \\ &= 1 - \frac{\sum_{j=1}^{K_t} n_{tj}}{t - 1 + \alpha} = 1 - \frac{t - 1}{t - 1 + \alpha} \\ &= \frac{\alpha}{t - 1 + \alpha} \quad \text{for } j = K_t + 1; \end{aligned} \tag{2.8}$$

on this derivation, see also Griffiths and Ghahramani (2011).

Equation (2.8) is the Pólya urn representation of the DP prior, originally proved by Blackwell and MacQueen (1973). An observation t is assigned to an existing class j with probability proportional to its "popularity" n_{tj}, and assigned to a new class with probability proportional to α; in the latter case, θ_t is drawn from the Normal-Gamma prior (2.2). Bayesian inference will combine this prior with the likelihood of the autoregressive model (2.1), leading to a potentially infinite mixture of autoregressions. Of course, the estimated model will be a finite mixture since the estimated number of regimes cannot exceed T. Low values of α will favor parsimony.

3. Gibbs sampling

The estimation method follows Escobar and West (1995) and Bush and MacEachern (1996); see also Neal (2000). It relies on the fact that the DP prior is exchangeable, so that (2.8) gives the full conditional prior probability function of any c_t. We draw successively from the full conditional posteriors:

$$\theta_1, \ldots, \theta_T \mid m, V, \alpha, y$$

$$\alpha \mid \theta_1, \ldots, \theta_T$$

$$V \mid m, \theta_1, \ldots, \theta_T$$

$$m \mid V, \theta_1, \ldots, \theta_T.$$

3.1 Drawing $\theta_1, \ldots, \theta_T$.

The full conditional posterior of θ_t is obtained by multiplying the full conditional prior obtained from (2.8) by the likelihood $\prod_{t=1}^{T} f_t(y_t \mid \theta)$, where:

$$f_t(y_t \mid \theta) = f_N(y_t \mid \mu + \sum_{j=1}^{p} \phi^j y_{t-j}, \sigma^2).$$

For ease of notation, we will omit in what follows the parameters of the Normal-Gamma density (2.2) and simply write $g_0(\theta)$. Upon defining:

$$\theta_{-t} = (\theta_1, \ldots, \theta_{t-1}, \theta_{t+1}, \ldots, \theta_T)$$

this yields the discrete-continuous mixture:

$$p(\theta_t \mid \theta_{-t}, \alpha, y) \propto \sum_{s \neq t} f_t(y_t \mid \theta_s) \mathcal{I}_{\theta_s}(\theta_t) + \alpha \left[\int_{\mathbb{R}^{p+1} \times \mathbb{R}_+} f_t(y_t \mid \theta) g_0(\theta) d\theta \right] p(\theta_t \mid y_t) \tag{3.1}$$

where:

$$p(\theta_t \mid y_t) \propto f_t(y_t \mid \theta_t) g_0(\theta_t). \tag{3.2}$$

So, θ_t is drawn from the set of existing θ_s with probability proportional to the likelihood of θ_s at time t, or a new value is drawn from the posterior (3.2) based on the single observation y_t, with probability proportional to the prior predictive density of y_t multiplied by α.

For reasons of efficiency, the algorithm is implemented as follows. Let $\tilde{\theta}_1$, $\ldots$,$\tilde{\theta}_K$ be the K distinct values generated by drawing $\theta_1, \ldots, \theta_T$ from (3.1), and let the indicator variables $c_1, \ldots, c_T$ be such that $c_t = j$ iff $\theta_t = \tilde{\theta}_j$. K values θ_j^* are drawn from the partial sample posteriors that use all observations t such that $c_t = j$:

$$p(\theta_j^* \mid c_1, \ldots, c_T, y) \propto \prod_{t:c_t=j} f_t(y_t \mid \theta_j^*) g_0(\theta_j^*). \tag{3.3}$$

The intermediate values $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$ are then discarded and one sets $\theta_t = \theta_j^*$ if $c_t = j$.

The conjugacy between g_0 and the likelihood implies the following form for (3.3):

$$p(\theta_j^* \mid c_1, \ldots, c_T, y) = \\ f_N\left[\mu_j^*, \phi_j^{*1}, \ldots, \phi_j^{*p} \mid m_j^*, (\sigma_j^*)^2 V_j^*\right] f_G\left[(\sigma_j^*)^{-2} \mid (s_j^*)^{-2}, \nu_j^*\right] \tag{3.4}$$

and implies the following form for (3.2):

$$p(\theta_t \mid y_t) = f_N(\mu_t, \phi_t^1, \ldots, \phi_t^p) \mid \tilde{m}_t, \sigma_t^2 \tilde{V}_t) f_G(\sigma_t^{-2} \mid \tilde{s}_t^{-2}, \tilde{\nu}_t); \tag{3.5}$$

the parameters m_j^*, V_j^*, $\tilde{m}_t$, $\tilde{V}_t$, s_j^*, $\tilde{s}_t$, ν_j^*, $\tilde{\nu}_t$ appearing in (3.4) and (3.5) can be easily inferred from the Bayesian update formulas for Gaussian linear regression (see, e.g., Poirier, 1995, p. 526). On the other hand, the predictive density appearing in (3.1):

$$\int_{\mathbb{R}^{p+1} \times \mathbb{R}_+} f_t(y_t \mid \theta) g_0(\theta) d\theta = p(y_t \mid y_{t-1}, \ldots, y_{t-p})$$

is univariate Student with expectation $m' \tilde{y}_t$, scale parameter $s^2(1 + \tilde{y}_t' V \tilde{y}_t)$, and ν degrees of freedom, where $\tilde{y}_t' = (\,1 \quad y_{t-1} \quad \ldots y_{t-p}\,)$.

3.2 Drawing α.

The total mass parameter α can be estimated by Gibbs sampling if the Gamma prior (2.5) is assumed. Escobar and West (1995) show that this can be done by first drawing an auxiliary variable η from a Beta$(\alpha+1, T)$ distribution, conditional

on the previously drawn value of α. Conditional on η, a new value of α is then drawn from the discrete mixture:

$$p(\alpha \mid K, \eta) = \pi f_G\left(\alpha \,\middle|\, \frac{\nu_\alpha + 2K}{\frac{\nu_\alpha}{m_\alpha} - 2\ln\eta}\,,\, \nu_\alpha + 2K\right) + (1-\pi) f_G\left(\alpha \,\middle|\, \frac{\nu_\alpha + 2K - 2}{\frac{\nu_\alpha}{m_\alpha} - 2\ln\eta}\,,\, \nu_\alpha + 2K - 2\right) \quad (3.6)$$

where K is the current number of clusters (distinct values of $\theta_1, \ldots, \theta_T$) and where:

$$\pi = \frac{\frac{\nu_\alpha}{2} + K - 1}{T\left(\frac{\nu_\alpha}{2m_\alpha} - \ln\eta\right) + \frac{\nu_\alpha}{2} + K - 1}. \quad (3.7)$$

3.3 Drawing m and V.

For $j = 1, \ldots, K$, let θ_j^* be drawn from (3.3) and let:

$$\beta_j = (\,\mu_j^* \quad \phi_j^{*1} \quad \ldots \quad \phi_j^{*p}\,)'$$

so that $\theta_j^* = (\beta_j, (\sigma_j^*)^{-2})$. The Normal density in (2.2) can be written as a seemingly unrelated model with K observations and $p+1$ equations, as follows:

$$(\sigma_j^*)^{-1}\beta_j = (\sigma_j^*)^{-1} m + u_i \quad \text{for } j = 1, \ldots, K \quad (3.8)$$

where $u_i \sim \text{Niid}(0, V)$. The full conditional posteriors of m and V^{-1} under the hyperprior (2.3)–(2.4) can then easily be found to be (see, e.g. Koop, 2003, pp. 137-141):

$$p(V^{-1} \mid \theta_1^*, \ldots, \theta_K^*, m) = f_W(V^{-1} \mid \overline{\nu}, \overline{H}) \quad (3.9)$$

$$p(m \mid \theta_1^*, \ldots, \theta_K^*, V) = f_N(m \mid \overline{m}, \overline{V}) \quad (3.10)$$

with:

$$\overline{V} = \left(\underline{V}^{-1} + V^{-1}\sum_{j=1}^{K}\frac{1}{(\sigma_j^*)^2}\right)^{-1} \quad (3.11)$$

$$\overline{m} = \overline{V}\left(\underline{V}^{-1}\underline{m} + V^{-1}\sum_{j=1}^{K}\frac{\beta_j}{(\sigma_j^*)^2}\right) \quad (3.12)$$

$$\overline{\nu} = K + \underline{\nu} \quad (3.13)$$

$$\overline{H} = \left(\underline{H}^{-1} + \sum_{j=1}^{K}\frac{(\beta_j - m)(\beta_j - m)'}{(\sigma_j^*)^2}\right)^{-1}. \quad (3.14)$$

4. Predicting US real interest rate data

We will consider the monthly time series defined as $y_t = \pi_t - 1200 \times \Delta \ln p_t$, where π_t is the 3-month US treasury bill interest rate (in % per annum) and p_t is the consumer price index. The data for π_t and p_t, available as the series **ftb3** and **pcun** in the Haver Analytics USECON database, cover the period ranging from January 1948 to December 2015 (816 observations). They were not seasonally adjusted. Figure 1 presents a line graph, histogram, and correlogram of y_t. The data are strongly leptokurtic and exhibit significant autocorrelations at low-order and seasonal lags.

Our objective is a forecast evaluation exercise, where the one-step ahead predictive densities implied by the model (2.1)–(2.5) will be compared with those obtained from a benchmark autoregressive model. For this purpose, we will use the data ranging from January 1948 to December 1959 (144 observations) as a training sample, leaving the observations from January 1960 to December 2015 (672 observations) for forecast evaluation.

Geweke and Amisano (2010) propose two complementary tools for comparing and evaluating Bayesian predictions. The first one is based on the predictive densities:

$$p(y_t \mid y_{1:t-1}, M) = \int p(y_t \mid y_{1:t-1}, \theta_M, M) p(\theta_M \mid y_{1:t-1}, M) d\theta_M \tag{4.1}$$

where M is a model index, $p(y_t \mid y_{1:t-1}, \theta_M, M)$ is the conditional density of y_t implied by model M with parameters θ_M, and $p(\theta_M \mid y_{1:t-1}, M)$ is the posterior density of θ_M based on past observations.

The second tool is based on probability integral transforms (Rosenblatt, 1952; Diebold et al, 1998; Berkowitz, 2001). These are defined as:

$$F_t^M = \Phi^{-1} \left[\int_{-\infty}^{y_t^*} p(y_t \mid y_{1:t-1}, M) dy_t \right] \tag{4.2}$$

where y_t^* is the observed value of y_t and Φ^{-1} is the inverse normal integral. Ideally, the F_t^M should be $N(0,1)$ and independent.

A natural benchmark model is a Gaussian autoregression with a Normal-Gamma prior; in this case, the predictive densities in (4.1) are univariate Student (see, e.g. Koop, 2003, p. 46), so that MCMC is not necessary. We use a Gaussian AR(12) model with the prior (2.2), using $m = (0, \ldots, 0)$, $V =$

diag$(1, 0.1, \ldots, 0.1)$, $s^2 = 10$, and $\nu = 10$. The maximum lag order of 12 was chosen according to the Schwarz information criterion; the in-sample OLS residuals did not exhibit significant autocorrelations but were strongly leptokurtic, with a Bera-Jarque statistic of 367.56. The value chosen for s^2 was close to the OLS error variance estimate; the value of ν implies the existence of the first four prior moments of σ^2.

Figure 2 presents a histogram and correlogram of the probability integral transforms for $t = 1960(1)$ to $t = 2015(12)$. The histogram in Figure 2 indicates a very strong deviation from normality, showing that unconditional coverage is not satisfactory, and the correlogram suggests that conditional coverage is less than optimal.

On the other hand, the predictive density obtained from (2.1) with a DP prior is:

$$p(y_t \mid y_{1:t-1}, DP) = \int \left[\int \left[\int f_t(y_t \mid \theta_t) dG(\theta_t) \right] d\mathcal{P}(G \mid \alpha, G_0(m, V, s^2, \nu), y_{1:t-1}) \right] dF(m, V, \alpha \mid y_{1:t-1}) \quad (4.3)$$

where $f_t(y_t \mid \theta_t)$ is the Gaussian density implied by (2.1), F is the posterior distribution of (m, V, α), and $\mathcal{P}$ is the posterior distribution of G; see Basu and Chib (2003). From Ferguson (1973, Theorem 1) the posterior of G is again a Dirichlet process. Upon exploiting the Pólya urn representation of this process, we see that the inner integral in (4.3) is the discrete-continuous mixture:

$$p(y_t \mid m, V, s^2, \nu, \alpha, y_{1:t-1}, \theta_{1:t-1}) = \sum_{j=1}^{K} \frac{n_j}{\alpha + t - 1} f_N(y_t \mid \mu_j^* + \sum_{i=1}^{p} \phi_j^{*i} y_{t-i}, \sigma_j^{*2}) + \frac{\alpha}{\alpha + t - 1} f_{ST}(y_t \mid m' \tilde{y}_t, s^2(1 + \tilde{y}_t' V \tilde{y}_t), \nu) \quad (4.4)$$

where f_{ST} is a Student density, K is the number of distinct values in $(\theta_1, \ldots, \theta_{t-1})$, n_j is the number of these values equal to:

$$\theta_j^* = (\mu_j^*, \phi_j^{*1}, \ldots, \phi_j^{*p}, \sigma_j^{*-2}),$$

and $\tilde{y}_t' = (\,1 \quad y_{t-1} \quad \ldots y_{t-p}\,)$.

The predictive density in (4.3) can be estimated by a posterior average of the mixtures in (4.4) over posterior replications m^ℓ, V^ℓ, α^ℓ, K^ℓ, and $(\theta_j^{*\ell}, n_j^\ell)$ for $j = 1, \ldots, K^\ell$, where $\ell = 1, \ldots, R$ and R is the number of replications. The argument of Φ^{-1} in (4.2) can be estimated by an average of easily computed integrals.

We now discuss the choice of parameters in the hierarchical DP prior (2.2)–(2.5), choosing $p = 12$ as in the Gaussian AR model. For the normal density (2.3), we choose $\underline{m} = (0, \ldots, 0)$ and $\underline{V} = \mathrm{diag}(1, 0.1, \ldots, 0.1)$. In the Gamma density appearing in (2.2), we set $s^2 = 10$ and $\nu = 10$. This choice was the same as that for m, V, s^2 and ν in the Gaussian AR model. For the Wishart prior (2.4), we choose $\underline{\nu} = 15$ in order to guarantee the existence of the expectation of V (see Muirhead, 1982, p. 97) and a scale matrix equal to $\underline{H} = \delta^2 I$. The scale factor δ^2 can be interpreted as a smoothing prior parameter, as will be seen shortly. Finally, the parameters of the Gamma prior (2.5) were $m_\alpha = 0.2$ and $\nu_\alpha = 4$. We now discuss the latter choice.

Antoniak (1974) shows that the DP prior probability of k clusters in a sample of size T when $\alpha = 1$ is given by:

$$P^*(k, T, 1) = \frac{\left|S_T^{(k)}\right|}{\sum_{i=1}^{T} \left|S_T^{(i)}\right|}$$

where $S_T^{(k)}$ is the Stirling number of the first kind; see Abramowitz and Stegun (1972, p. 824). Escobar and West (1995) provide the more general expression for arbitrary α:

$$P^*(k, T, \alpha) = P^*(k, T, 1)\, T!\, \alpha^k\, \frac{\Gamma(\alpha)}{\Gamma(\alpha + T)}.$$

The marginal probability of k clusters in a sample of size T:

$$P(K = k \mid T, m_\alpha, \nu_\alpha) = \int_0^\infty P^*(k, T, \alpha) f_G(\alpha \mid m_\alpha, \nu_\alpha) d\alpha$$

can easily be computed by numerical integration. Table 1 gives several values of this function for $T = 144$ and $T = 815$, corresponding to the lengths of the first and last observation windows, when $m_\alpha = 0.2$ and $\nu_\alpha = 4$.

TABLE 1. *Prior probabilities and expectations ($\bar{k}$) of the numbers of mixture components*

k	1	2	3	4	5	6	7
$P(k)$, $T = 144$ ($\bar{k} = 2.03$)	0.42	0.31	0.16	0.07	0.03	0.01	0.00
$P(k)$, $T = 815$ ($\bar{k} = 2.38$)	0.34	0.29	0.18	0.10	0.05	0.02	0.01

TABLE 2. *Posterior impact of the choice of the smoothing parameter δ^2*

δ^2	1	10	100	1000	10000
$\ln \hat{p}(y_{145:816} \mid y_{1:144})$	−1800.03	−1794.42	−1790.93	−1793.06	−1793.56
KS p-value	0.73	0.84	0.64	0.68	0.73
$\hat{E}(K \mid y_{1:144})$	2.58	3.87	4.57	3.86	3.74
$\hat{E}(K \mid y_{1:815})$	5.51	6.93	7.94	7.29	6.71

The choice of the smoothing parameter δ^2 also deserves some discussion. Table 2 presents, for different values of δ^2, predictive marginal likelihood estimates:

$$\ln \hat{p}(y_{145:816} \mid y_{1:144}, DP) = \sum_{s=145}^{816} \ln \hat{p}(y_s \mid y_{1:s-1}, DP)$$

where $\hat{p}(y_s \mid y_{1:s-1}, DP)$ is a Monte Carlo estimate of (4.1), obtained by averaging (4.4) over posterior replications. The value $\delta^2 = 100$ clearly dominates.

In order to illustrate some consequences of different choices of δ^2, we also report in Table 2 estimates of the posterior expectation of K when $T = 144$ (where estimation covers 1948(1) to 1959(12)) and when $T = 815$ (estimation covers 1948(1) to 2015(11)). The choice of δ^2 has a clear influence on the number of components. However, its impact on predictive coverage appears to be minor. This can be formalized in the second row of Table 2, which reports p-values of

Kolmogorov-Smirnov (KS) tests of the $N(0,1)$ null hypothesis, based on estimates of F_t^{DP} in (4.2), for $t = 145, \ldots, 816$. None of the p-values is significant.

In Figures 3 to 5, we present the forecast evaluation and comparison results, using the DP model with $\delta^2 = 100$ and the Gaussian AR model, both with $p = 12$ lags. The top panel of Figure 3 shows a histogram of the probability integral transforms (estimates of F_t^{DP} for $t = 145, \ldots, 816$). The strong deviation from normality that was evident in the Gaussian AR model has now disappeared. However, some autocorrelations in the bottom panel remain significant, suggesting that the predictive densities still neglect some useful past information.

The top panel of Figure 4 reports the 95% forecast intervals obtained with the DP model, together with the predicted observations; the fraction of observations lying outside of the forecast intervals is 5.2%. The bottom panel of Figure 4 compares the lengths of the forecast intervals in the DP model and in the Gaussian AR model. This provides some insight on the reason why unconditional predictive coverage is satisfactory in the DP model but not in the AR model. Indeed, the interval lengths in the AR model decrease monotonically until 1973, reflecting the additional precision stemming from expanding estimation windows. They stay relatively constant thereafter, with the exception of small transient "bumps" during periods of instability. By contrast, in the DP model, the forecast intervals tend to be shorter than the AR ones for the periods 1960–1973, 1975–1980, and 1985–2003; during periods of instability, however, they can become much wider, especially after the financial crisis of 2008.

In the top panel of Figure 5, we plot the evolution over time of estimates of the posterior expectations of K, given by the averages of the posterior replications. The data strongly favor an increase in the number of components over time, with a sharp jump at the onset of the 2008 financial crisis. The bottom panel displays the evolution of the logarithm of the predictive Bayes factor in favor of of the DP model against the Gaussian AR model:

$$\log_{10} BF(t) = \sum_{s=145}^{t} \log_{10} \hat{p}(y_s \mid y_{1:s-1}, DP) - \sum_{s=145}^{t} \log_{10} \hat{p}(y_s \mid y_{1:s-1}, AR)$$

for $t = 145, \ldots, 816$. The cumulative evidence in favor of the DP mixture model gradually increases from 1960 to 1973, becoming overwhelming after about 60 observations according to the Jeffreys evidence scale (Jeffreys, 1961). After 1973,

it remains approximately constant until the end of 2008, corresponding to the recent financial crisis, when it sharply increases.

5. Discussion and conclusions

This paper has investigated the performance of a DP mixture of autoregressions for forecasting real interest rates. The predictive Bayes factors strongly favor the DP model over a benchmark Gaussian autoregression, and this dominance is apparent over the entire forecast window. Unconditional predictive coverage is satisfactory in the DP model but not in the benchmark model. However, conditional coverage was found wanting in both models.

The DP mixture model in this paper is not a substitute for the models briefly surveyed in the Introduction, since it is used only for forecasting purposes. However, it could serve as an easily implemented benchmark for a comparative investigation of forecasts from different versions of Markov-switching or change-point models. Formulating and estimating these models requires considerable skill and computational investment: in a model with a fixed number of regimes, this number is usually chosen by comparing marginal likelihoods, which is a demanding exercise (see, e.g., Frühwirth-Schnatter, 2004). Also, such a model is unsuitable for forecasting unless the number of regimes does not evolve over time (a strong assumption, as Figure 5 indicates). On the other hand, estimating the model in Koop and Potter (2007) requires $O(T^3)$ operations for a single sweep of their algorithm; and drawing the regime indicators in the hierarchical DP mixture models briefly surveyed in the Introduction requires truncating the maximum number of regimes to a large number or using beam sampling, followed by an application of forward filtering-backward sampling.

By contrast, our DP mixture model makes minimal assumptions on the dynamics of the mixture indicators, and its estimation is relatively inexpensive. One thousand burn-in replications are more than sufficient to guarantee convergence, and a single sweep of the Gibbs sampler in Section 3 only requires 17 milliseconds of computer time (using a 2.8 Ghz workstation and compiled code, with $T = 816$ observations and $p = 12$ lags).

This paper suggests that forecasting the monthly real interest rate is a challenging exercise, but that a simple DP mixture autoregressive model with a hierarchical prior is a practical and reasonably effective tool for this purpose. A

comparative investigation of other models that could be used in this context, however, would be outside the scope of this work.

References

M. Abramowitz and I.A. Stegun, 1972, *Handbook of Mathematical Functions*, Dover Publications, New York.

J. Albert and S. Chib, 1993, Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, *Journal of Business and Economic Statistics* 11, 1-15.

C.E. Antoniak, 1974, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics* 2, 1152-1174.

S. Basu and S. Chib, 2003, Marginal likelihood and Bayes factors for Dirichlet process mixture models, *Journal of the American Statistical Association* 98, 224-235.

L. Bauwens, J-F. Carpantier, and A. Dufays, 2016, Autoregressive moving average infinite hidden Markov-switching models, *Journal of Business and Economic Statistics*, forthcoming.

J. Berkowitz, 2001, Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics* 19, 465-474.

D. Blackwell and J.B. MacQueen, 1973, Ferguson distributions via Pólya urn schemes, *The Annals of Statistics* 1, 353-355.

C.A. Bush and S.N. MacEachern, 1996, A semiparametric Bayesian model for randomized block designs, *Biometrika* 83, 275-285.

S. Chib, 1996, Calculating posterior distributions and modal estimates in Markov mixture models, *Journal of Econometrics* 75, 79-97.

S. Chib, 1998, Estimation and comparison of multiple change-point models, *Journal of Econometrics* 86, 221-241.

P.J. Deschamps, 2006, A flexible prior distribution for Markov switching autoregressions with Student-t errors, *Journal of Econometrics* 133, 153-190.

P.J. Deschamps, 2008, Comparing smooth transition and Markov switching autoregressive models of US unemployment, *Journal of Applied Econometrics* 23, 435-462.

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

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

T.S. Ferguson, 1973, A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* 1, 209-230.

E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky, 2011, A sticky HDP-HMM with application to speaker diarization, *The Annals of Applied Statistics* 5, 1020-1056.

S. Frühwirth-Schnatter, 2004, Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques, *Econometrics Journal* 7, 143-167.

R. Garcia and P. Perron, 1996, An analysis of the real interest rate under regime shifts, *Review of Economics and Statistics* 78, 111-125.

J. Geweke and G. Amisano, 2010, Comparing and evaluating Bayesian predictive distributions of asset returns, *International Journal of Forecasting* 26, 216-230.

P. Giordani and R. Kohn, 2008, Efficient Bayesian inference for multiple change-point and mixture innovation models, *Journal of Business and Economic Statistics* 26, 66-77.

S. Goldfeld and R. Quandt, 1973, A Markov model for switching regression, *Journal of Econometrics* 1, 3-16.

T.L. Griffiths and Z. Ghahramani, 2011, The Indian buffet process: An introduction and review, *Journal of Machine Learning Research* 12, 1185-1224.

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

H. Jeffreys, 1961, *Theory of Probability (3d. edition)*, Oxford University Press, Oxford.

G. Koop, 2003, *Bayesian Econometrics*, Wiley, Chichester.

G. Koop and S.M. Potter, 2007, Estimation and forecasting in models with multiple breaks, *Review of Economic Studies* 74, 763-789.

R. Muirhead, 1982, *Aspects of multivariate statistical theory*, Wiley, New York.

R.M. Neal, 2000, Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* 9, 249-265.

M.H. Pesaran, D. Pettenuzzo, and A. Timmermann, 2006, Forecasting time series subject to multiple structural breaks, *Review of Economic Studies* 73, 1057-1084.

D.J. Poirier, 1995, *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge, Massachusetts.

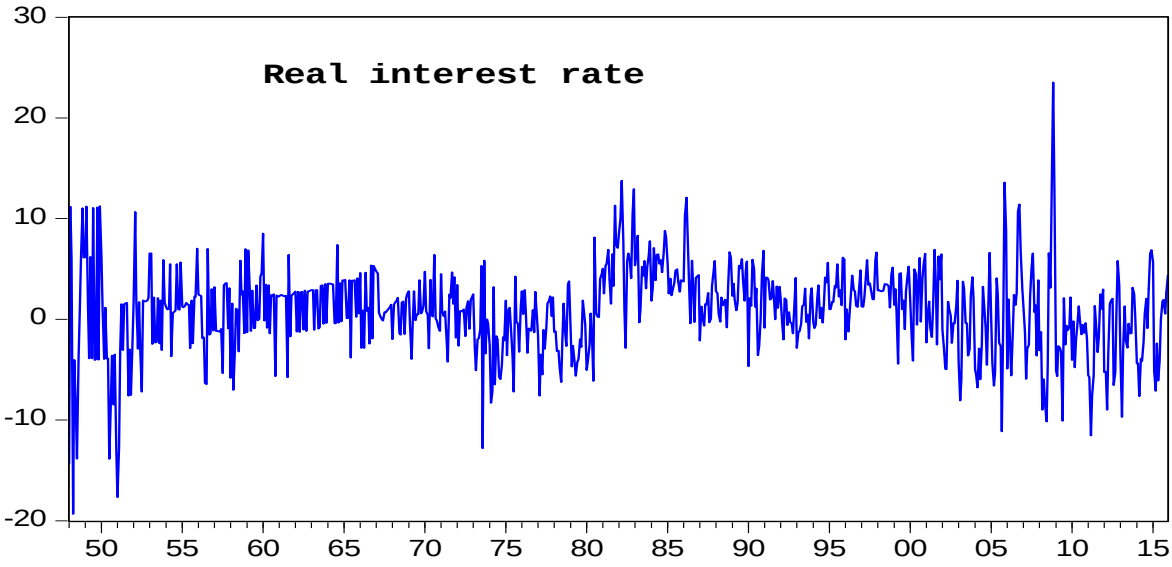
M. Rosenblatt, 1952, Remarks on a multivariate transformation, *Annals of Mathematical Statistics* 23, 470-472.

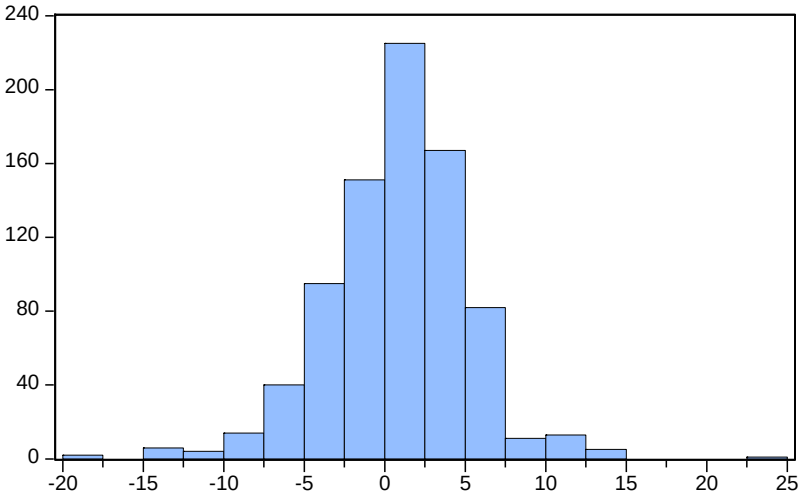
J. Sethuraman, 1994, A constructive definition of Dirichlet priors, *Statistica Sinica* 4, 639-650.

Y. Song, 2014, Modelling regime switching and structural breaks with an infinite hidden Markov model, *Journal of Applied Econometrics* 29, 825-842.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, 2006, Hierarchical Dirichlet processes, *Journal of the American Statistical Association* 101, 1566-1581.

Figure 1. Observations





Series: Real interest rate	
Sample 1948M01 2015M12	
Observations 816	
Mean	0.872510
Median	1.277023
Maximum	23.50643
Minimum	-19.33947
Std. Dev.	4.298113
Skew ness	-0.267712
Kurtosis	5.349121
Jarque-Bera	197.3717
Probability	0.000000

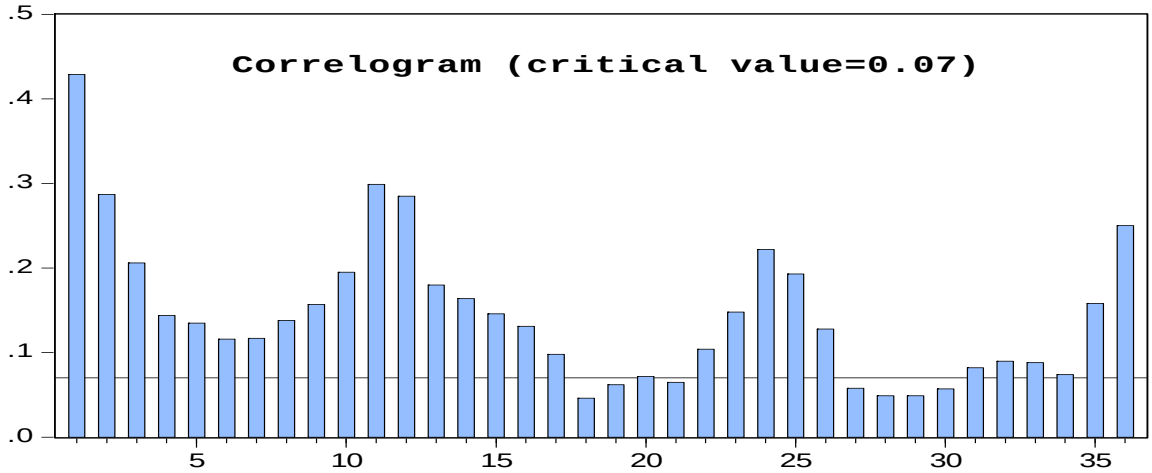
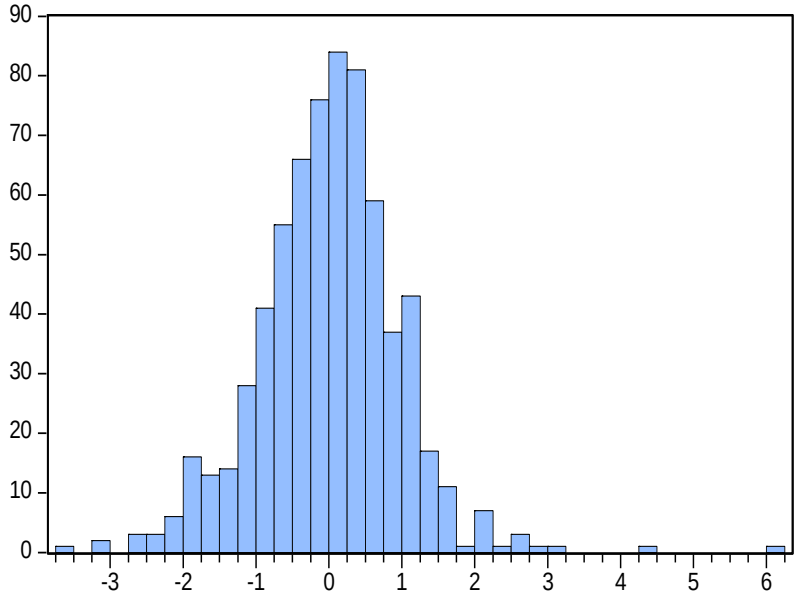


Figure 2. Probability integral transforms (Gaussian AR model)



Sample 1960M01 2015M12	
Observations 672	
Mean	0.000354
Median	0.042623
Maximum	6.024940
Minimum	-3.500268
Std. Dev.	0.961359
Skew ness	0.278546
Kurtosis	6.183343
Jarque-Bera	292.4326
Probability	0.000000

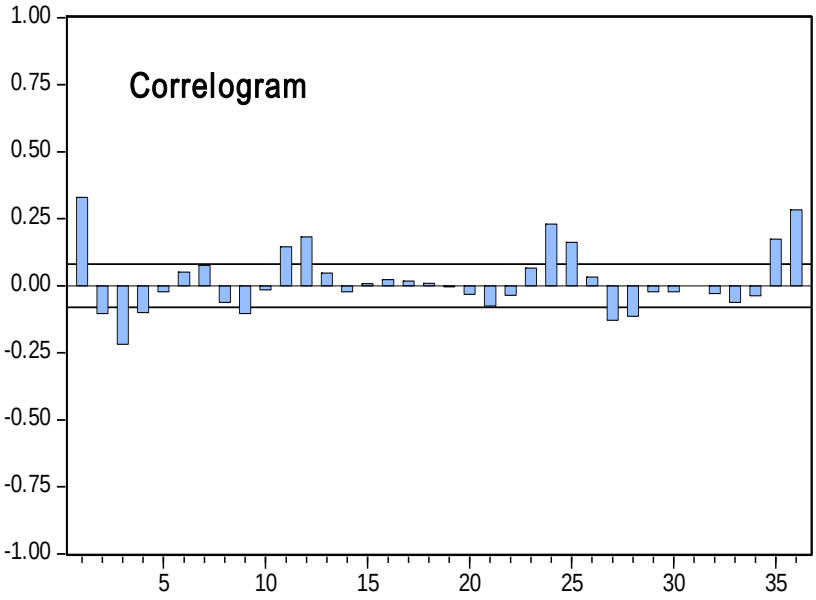
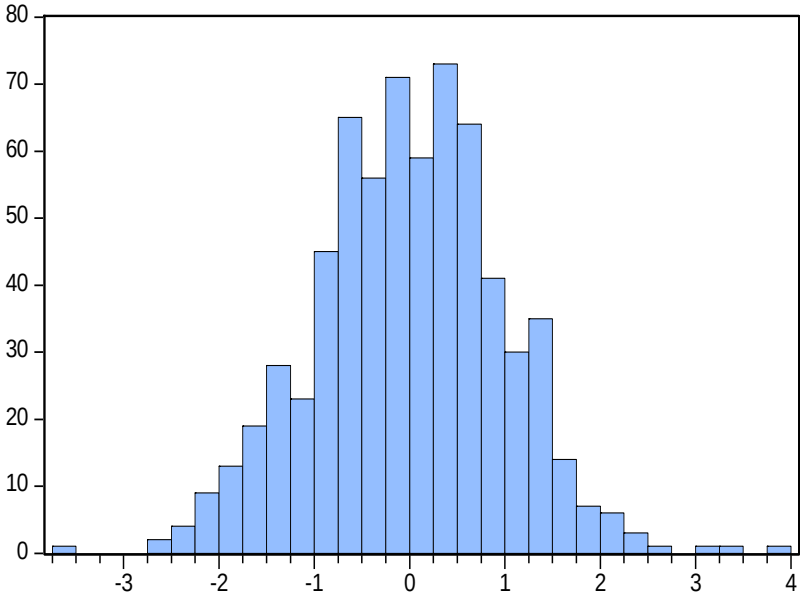


Figure 3. Probability integral transforms (Dirichlet process mixture)



Sample 1960M01 2015M12	
Observations 672	
Mean	-0.011341
Median	-9.80e-05
Maximum	3.845255
Minimum	-3.730168
Std. Dev.	0.983456
Skew ness	-0.010662
Kurtosis	3.411970
Jarque-Bera	4.764872
Probability	0.092325

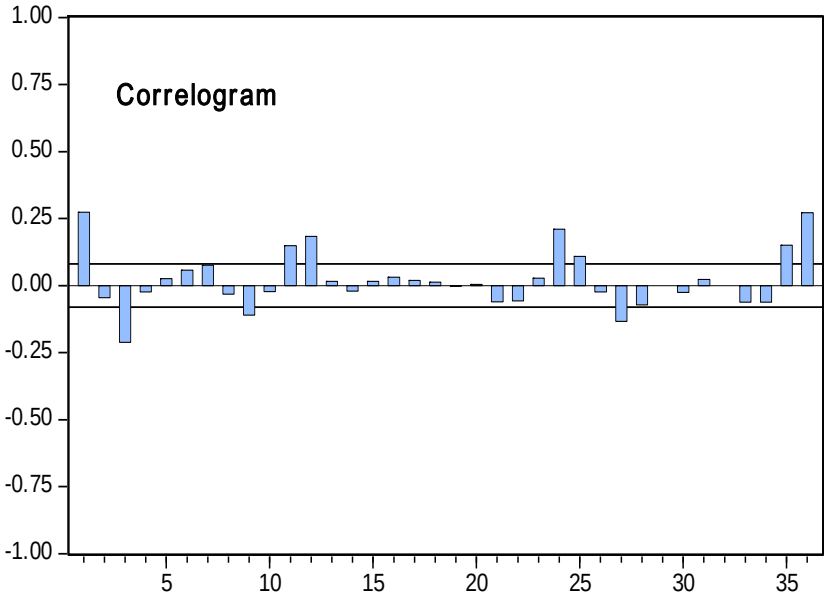


Figure 4. Forecast intervals

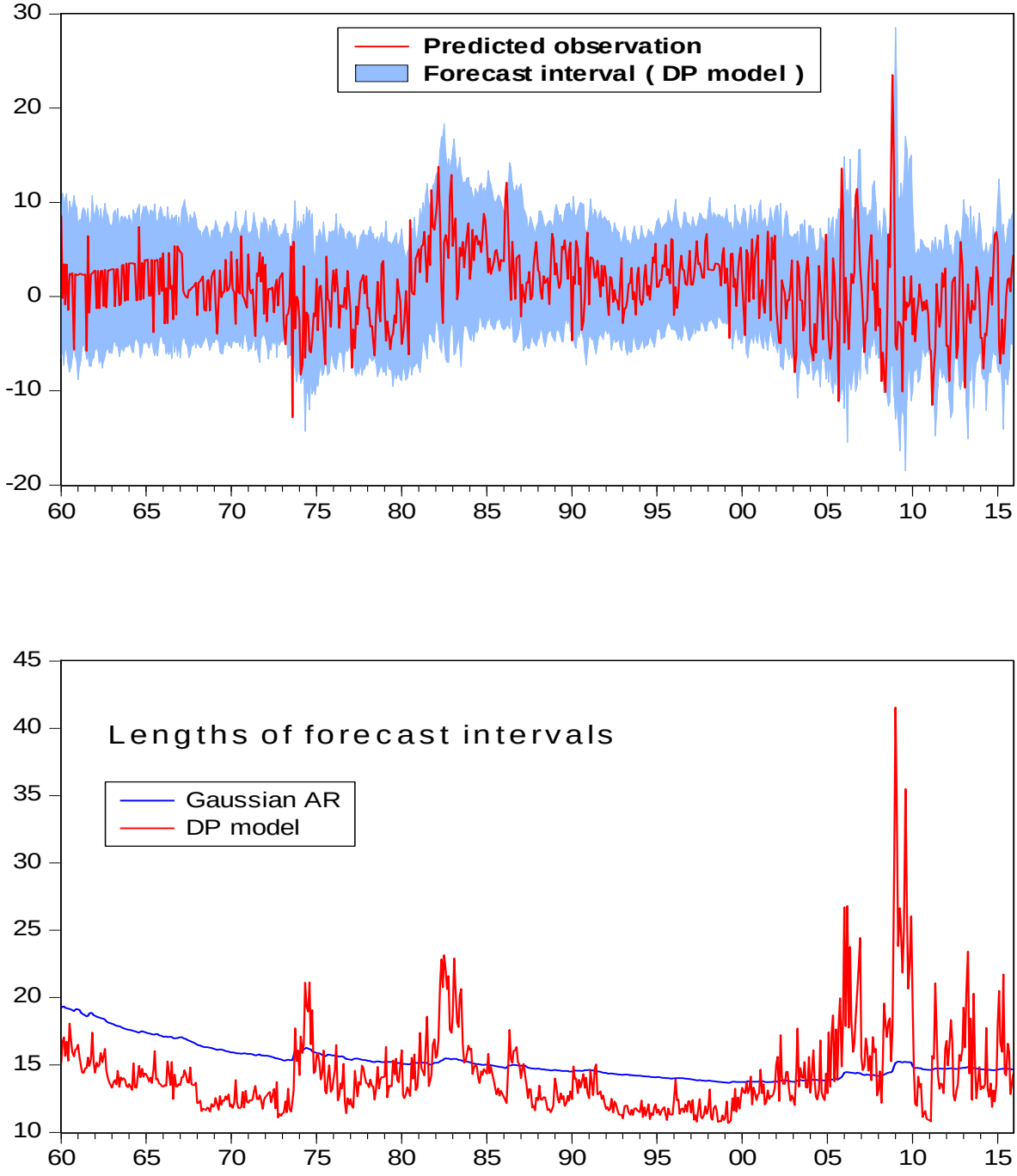


Figure 5. Expected numbers of mixture components and logarithmic Bayes factors