REAL EXCHANGE RATES AND UNIT ROOTS: LEARNING ABOUT THE DISTRIBUTION

GERALD P. DWYER AND MARK FISHER

Incomplete.

Abstract. The goal of this paper is to learn about the distribution of the autoregression coefficient for real exchange rates, including the probability of a unit root. The paper is an exercise in Bayesian statistics. The approach we take allows us to learn not only about the distribution for each specific case for which we have data, but also the generic case for which we have no data as yet. The posterior distribution for the generic case constitutes a well-informed prior distribution for a new case when such data becomes available. The estimation of the distribution for the generic case amounts to indirect density estimation for a latent variables. With this in mind, we adopt a nonparametric Bayesian prior that embodies great flexibility and allows for a unit root as a special case.

Date: July 28, 2017 @ 07:01. Filename: `ppp_distributions`.

1. Introduction [incomplete]

The goal of this paper is to learn about the distribution of the autoregression coefficient for real exchange rates, including the probability of a unit root. The paper is an exercise in Bayesian statistics. The approach we take allows us to learn not only about the distribution for each specific case for which we have data, but also the generic case for which we have no data as yet. The posterior distribution for the generic case constitutes a well-informed prior distribution for a new case when such data becomes available. The estimation of the distribution for the generic case amounts to indirect density estimation for a latent variables. With this in mind, we adopt a nonparametric Bayesian prior that embodies great flexibility and allows for a unit root as a special case. As Poirier (1991, p. 384) says, "I think the case can be made that [the unit root] is sufficiently special to warrant a prior atom."

Literature review. Our approach is novel in a number of ways and has very little in common with earlier Bayesian papers related to the subject such as Schotman and van Dijk (1991) and DeJong and Whiteman (1991). See also Bauwens et al. (1999) for a comprehensive review of the literature.

Background for Baysian inference. Some useful background reading: Koop (2003), Greenberg (2013), Gelman et al. (2014), Kruschke (2011). Textbook introductions to Dirichlet process mixture models can be found in Greenberg (2013) and Gelman et al. (2014). See also Gershman and Blei (2012).

The prior presented here is an extension (allowing for a point mass) of the prior presented in Fisher (2017).

Outline. In Section 2 we introduce the data and the likelihoods and compute Dickey–Fullers tests. In Section 3 we compare two simple models and introduce some basic features of our more general models. In Section 4 we present the hierarchical prior that allows for learning. In Section 5 we specialize the distributions introduced in Section 4. In Section 6 we present an overview of the MCMC sampler. In Section 7 we investigate the data.

There are a number of appendices. Appendix B presents some technical remarks regarding the prior presented in Section 3. Appendix C presents a thumbnail sketch of the measure theory involved in mutually singular measures. In Appendix D we provide additional details regarding the posterior. In Appendix E we describe the scheme we use for sampling from the posterior distribution. In Appendix F we present the SUR likekhood. In Appenxid G we describe a factor structure for the residuals. In Appendix H we show how to implement a more general prior.

2. The data and the likelihood

In this section we describe the data, present the likelihoods, and perform a Dickey–Fuller test.

The data. [Need to treat real exchange rates more fully.]

The real exchange rate between two countries involves the nominal exchange rate between the two currencies and the two price levels. In logs, we have

$$y_{it} = \log(e_{it}) + \log(P_{it}) - \log(P_t^*), \tag{2.1}$$

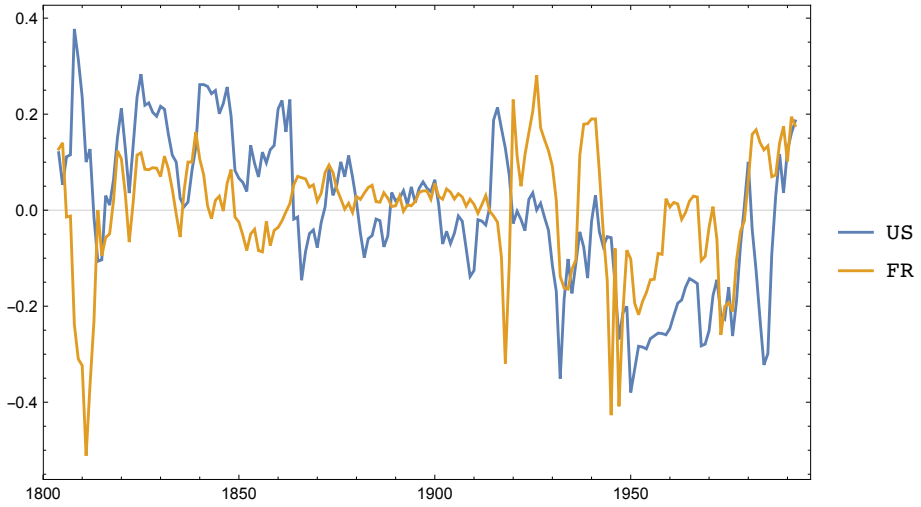

FIGURE 1. LT real exchange rate data: annual, 1804–1992, in logs. Exchange rates are computed relative to GR.

where e_{it} is the nominal exchange rate, P_{it} is the price level in country i and P_t^* is the price level in a "reference" country. If the countries share the same currency, then $e_{it} = 1$ and

$$y_{it} = \log(P_{it}) - \log(P_t^*). \tag{2.2}$$

[Need to describe the data more fully.] See Lothian and Taylor (1996).

The data are shown in Figures 1 and 2. Each series is normalized to have a mean of zero. We break the Euro-related data into two groups, based on when the Euro was adopted. The year-to-year variance in the data is dramatically different between the two eras.

The likelihood. Let y_{it} denote the log of the real exchange rate between to countries, where i indexes the exchange rate and t indexes time. Let $Y_i = (y_{i1}, \ldots, y_{iT_i})$ and $Y_{1:n} = (Y_1, \ldots, Y_n)$. Consider the following model for annual data:

$$y_{it} = \alpha_i + \beta_i\, y_{i,t-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_i^2), \tag{2.3}$$

for $t \geq 2$.

Given (2.3), we can express the model for the data as

$$p(Y_i|\alpha_i, \beta_i, \sigma_i^2) = \prod_{t=2}^{T_i} \mathsf{N}(y_{it}|\alpha_i + \beta_i\, y_{i,t-1}, \sigma_i^2). \tag{2.4}$$

Note we condition on y_{i1}. We integrate out the nuisance parameters (α_i, σ_i^2) using the Jeffreys prior where $p(\alpha_i, \sigma_i^2) \propto p(\sigma_i^2)$ and $p(\sigma_i^2)$ is given in (A.4),[1] producing the marginal likelihood for β_i:

$$p(Y_i|\beta_i) = \mathsf{Student}(\beta_i|m_i, s_i^2, \nu_i) \propto \iint p(Y_i|\alpha_i, \beta_i, \sigma_i)\, p(\sigma_i^2)\, d\sigma_i^2\, d\alpha_i, \tag{2.5}$$

[1]If $a_0 = b_0 = 0$ [see Appendix A], then $p(\alpha_i, \sigma_i^2) \propto 1/\sigma_i^2$, which is the Jefferys prior. The Jeffreys prior is improper (it does not integrate to a finite value) and it is interpreted as uninformative.

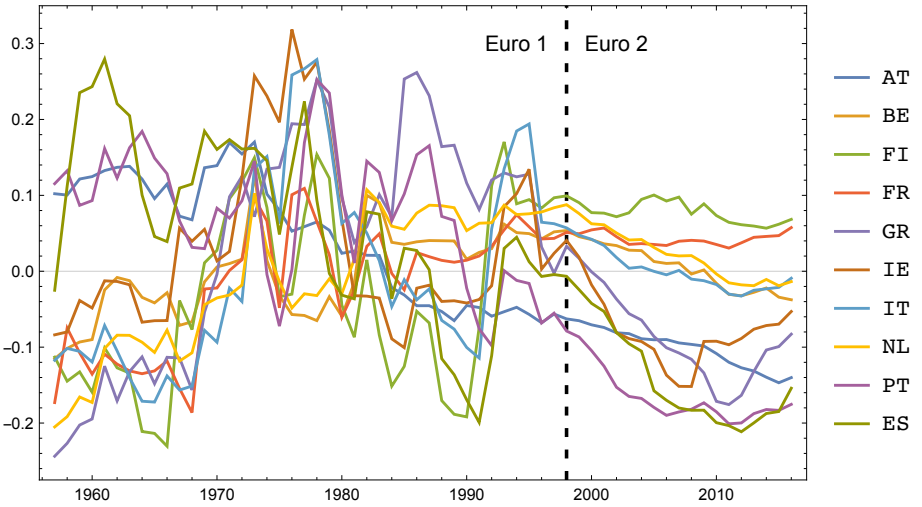

FIGURE 2. Euro-related real exchange rate data: annual, 1957–2016, in logs. Exchange rates are computed relative to DE. The last observation of Euro 1 is 1998, which is also the first observation of Euro 2.

where (m_i, s_i, ν_i) depends on Y_i as shown in Appendix A. See Table 1.

At this stage we have treated the likelihoods as independent in the sense that

$$p(Y_{1:n}|\beta_{1:n}) = \prod_{i=1}^{n} p(Y_i|\beta_i), \tag{2.6}$$

where $\beta_{1:n} = (\beta_1, \ldots, \beta_n)$. This follows from the assumed independence of ε_{it} and ε_{jt}. In Appendix F we will relax this assumption, in which case we will not be able to factor the joint likelihood as in (2.6).

Dickey–Fuller tests. The sole purpose here is to confirm that our Euro-related data are not exceptional with respect to traditional unit root tests.

One version of a Dickey–Fuller test involves the "t statistic"

$$t_{T_i} = \frac{m_i - 1}{s_i} \tag{2.7}$$

to determine whether to reject the null hypothesis of a unit root.[2] In particular, the null hypothesis is rejected if $t_{T_i} < c$ where c is a critical value in the applicable Dickey–Fuller table. The 5% critical values are shown in Table 2.[3]

As indicated in Table 1, the null is rejected for each of the two LT data series, but the null is not rejected for all but one of the Euro-related series. Taking the both Euro-related data sets together, we see that the null is rejected 5% of the time at the 5% level, which is consistent with the null being true for the Euro-related series as a whole. This finding is consistent with what many researchers have found looking at a variety of post-war series

[2]This assumes $a_0 = b_0 = 0$ in the prior for σ_i^2.

[3]They are calculated via linear interpolation for Case 2 in Table B.6 in Hamilton (1994).

TABLE 1. The data in three groups. Note $\nu_i = T_i - 3$. LT exchange rates relative to UK; Euro exchange rates relative to DE.

			m_i	s_i	ν_i	Reject null
LT	1	US	0.901	0.032	186	✓
	2	FR	0.782	0.046	186	✓
Euro 1	3	AT	0.978	0.044	39	
	4	BE	0.733	0.099	39	
	5	FI	0.805	0.098	39	
	6	FR	0.768	0.093	39	
	7	GR	0.882	0.058	39	
	8	IE	0.844	0.082	39	
	9	IT	0.855	0.080	39	
	10	NL	0.877	0.059	39	
	11	PT	0.742	0.115	39	
	12	ES	0.821	0.091	39	
Euro 2	13	AT	0.994	0.042	16	
	14	BE	0.951	0.060	16	
	15	FI	0.708	0.158	16	
	16	FR	0.789	0.183	16	
	17	GR	0.853	0.062	16	
	18	IE	0.743	0.092	16	
	19	IT	0.837	0.057	16	
	20	NL	0.904	0.042	16	
	21	PT	0.804	0.057	16	✓
	22	ES	0.859	0.052	16	

TABLE 2. Critical values for Dickey–Fuller test of Case 2.

T	$c_{.05}$
189	−2.88
42	−2.95
19	−3.02

at a variety of frequencies. We are not interested in parsing these results any further and instead we now switch gears.

3. A simple Bayesian framework

We begin our Bayesian analysis by first examining a simple model that does not involve learning across exchange rates. We compare the special case of a unit root to the simple model for each exchange rate separately as well as for groups of exchange rates. Along the way we introduce a number of assumptions and conventions that we rely on throughout the paper.

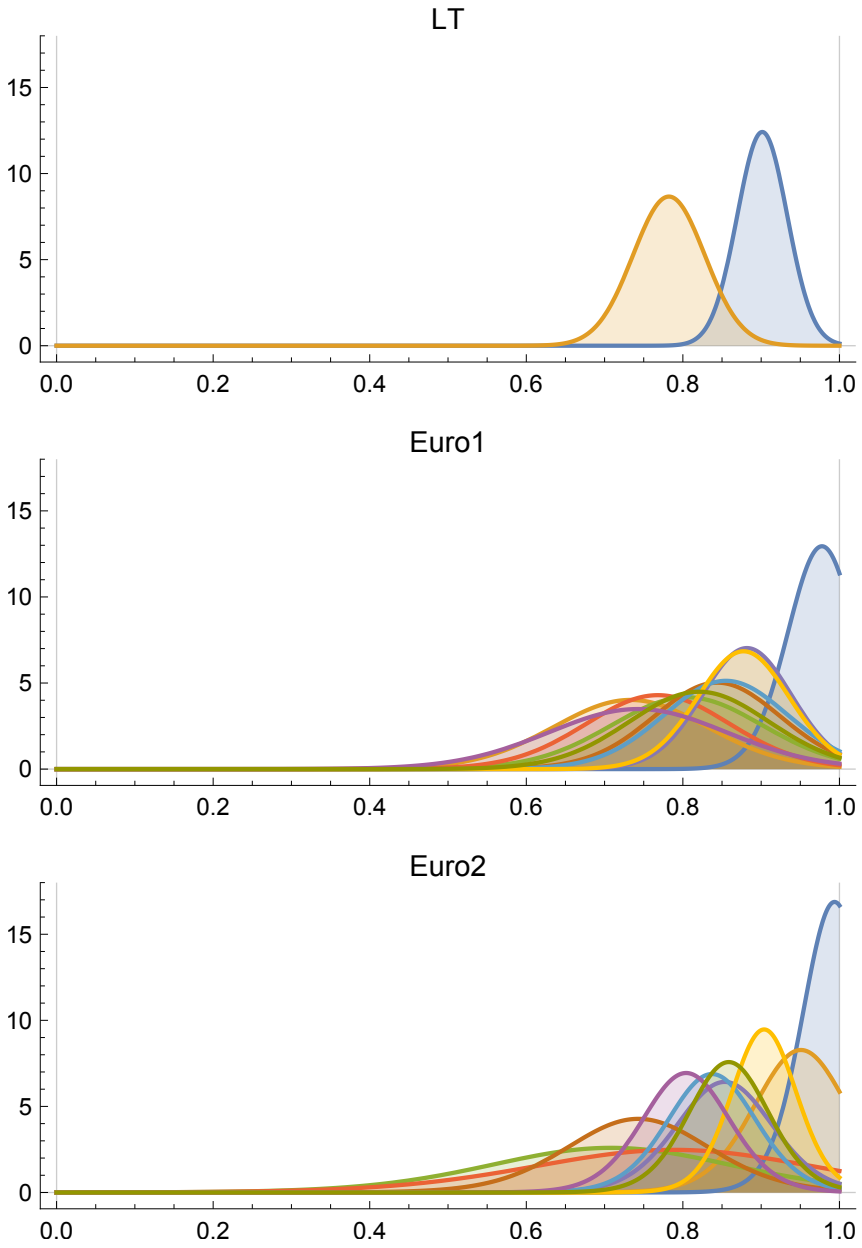

FIGURE 3. Truncated Student t distributions by group.

Bayes' rule. We begin by introducing Bayes' rule:

$$p(\beta_i|Y_i) = \frac{p(Y_i|\beta_i)\, p(\beta_i)}{p(Y_i)} \tag{3.1}$$

Country	Autocorrelation	t-ratio	p-value
1/1957 to 12/1998			
Austria	0.949	-0.905	0.787
Belgium	0.827	-2.338	0.160
Finland	0.752	-2.806	0.058
France	0.755	-2.878	0.049
Greece	0.842	-2.125	0.235
Ireland	0.790	-2.275	0.181
Italy	0.879	-1.702	0.429
Netherlands	0.820	-2.415	0.138
Portugal	0.753	-2.295	0.174
Spain	0.788	-2.233	0.195
1/1999 to 12/2016			
Austria	0.943	-0.626	0.861
Belgium	0.907	-1.157	0.693
Finland	0.607	-2.378	0.149
France	0.662	-1.571	0.495
Greece	0.764	-2.745	0.068
Ireland	0.806	-2.669	0.081
Italy	0.775	-2.457	0.128
Netherlands	0.596	-3.275	0.017
Portugal	0.686	-3.513	0.009
Spain	0.848	-2.139	0.230

TABLE 3. Dickey-Fuller Tests on Real Exchange Rates. The tests are run with the number of lagged changes estimated using the Schwarz Information Criterion. The number of lagged changes included for the period 1/1957 to 12/1998 is zero for all countries other than Belgium, which has one lag, and Ireland, which has 12 lags. The number of lagged changes included for the period 1/1999 to 12/2016 is 12 for all countries other than Austria and France which have 13 lagged changes included in the test regressions.

which expresses the posterior distribution for β_i, $p(\beta_i|Y_i)$, in terms of the likelihood of β_i, $p(Y_i|\beta_i)$ [which is given in Section 2], the prior distribution for β_i, $p(\beta_i)$, and the marginal likelihood of the data (according to the model: the likelihood and the prior)

$$p(Y_i) = \int p(Y_i|\beta_i)\, p(\beta_i)\, d\beta_i. \tag{3.2}$$

Throughout the paper we impose the following restriction: $\beta_i \in [0, 1]$. This restriction rules out both explosive behavior and negative autocorrelation. (The general approach we present does not depend on this restriction.)

TABLE 4. Normalized values: $g_i = \int_0^1 p(Y_i|\beta_i)\,d\beta_i$, $h_i = p(Y_i|\beta_i = 1)$, $z_i = h_i/g_i$, and $d_i = p(Y_i|\beta_i^*)$. Overall product: $\prod_{i=1}^{22} z_i = 1.2 \times 10^{-8}$. Note, the last column displays the maximum evidence against a unit root: $d_i/h_i = 1/\mathcal{B}_i^*$.

		g_i	h_i	$z_i = h_i/g_i$	$\prod_{i=1}^{n} z_i$	d_i	d_i/h_i
LT	1	0.9988	0.1227	0.1229		12.3989	101.0
	2	1.0000	0.0002	0.0002	2.5×10^{-5}	8.6627	42,024.9
Euro 1	3	0.6914	7.8603	11.3685		8.9495	1.1
	4	0.9948	0.1324	0.1331		3.9974	30.2
	5	0.9734	0.5818	0.5976		4.0595	7.0
	6	0.9914	0.2224	0.2243		4.2581	19.1
	7	0.9763	0.8924	0.9142		6.8633	7.7
	8	0.9685	0.8067	0.8329		4.8635	6.0
	9	0.9604	0.9948	1.0358		4.9254	5.0
	10	0.9773	0.8366	0.8560		6.6872	8.0
	11	0.9843	0.3106	0.3155		3.4328	11.1
	12	0.9724	0.6482	0.6666	2.9×10^{-2}	4.3695	6.7
Euro 2	13	0.5591	9.3220	16.6724		9.4357	1.0
	14	0.7867	4.6044	5.8525		6.5115	1.4
	15	0.9581	0.4811	0.5021		2.4851	5.2
	16	0.8674	1.0862	1.2522		2.1477	2.0
	17	0.9843	0.4997	0.5077		6.3222	12.7
	18	0.9933	0.1492	0.1503		4.2480	28.5
	19	0.9940	0.2151	0.2164		6.8392	31.8
	20	0.9816	0.8553	0.8714		9.2941	10.9
	21	0.9984	0.0607	0.0608		6.9295	114.1
	22	0.9922	0.3055	0.3079	1.7×10^{-2}	7.5159	24.6

As a starting point, let the prior for β_i be the uniform distribution over the unit interval:

$$p(\beta_i) = \mathsf{Uniform}(\beta_i|0, 1). \tag{3.3}$$

With this prior, the posterior distribution for β_i is proportional to the likelihood of β_i over the unit interval:

$$p(\beta_i|Y_i) = \frac{p(Y_i|\beta_i)}{\int_0^1 p(Y_i|\beta_i)\,d\beta_i} = \frac{p(Y_i|\beta_i)}{g_i}, \tag{3.4}$$

where $g_i := \int_0^1 p(Y_i|\beta_i)\,d\beta_i$. The posterior distributions are truncated Student t distributions [see Figure 3]. See Table 4 for the specific values of g_i. The marginal likelihood of the data (according to this model) is simply the area under the likelihood (over the unit interval); more generally, $p(Y_i)$ is an average of the likelihood for each value of $\beta_i \in [0, 1]$.

Let us now turn to the special case of a unit root, characterized by $\beta_i = 1$. The likelihood of the special case is simply the likelihood evaluated at $\beta_i = 1$, namely $p(Y_i|\beta_i = 1)$. For reference, let $h_i := p(Y_i|\beta_i = 1)$. See Table 4 for the specific values of h_i.

Model comparison. What can be said about the relative merits of the two models? A standard Bayesian approach involves comparing models via the likelihood of the data according to the model. We take that approach here. First we make case-by-case comparisons, after which we make group-wise comparisons. Before proceeding we note that the Bayes factor for a sharp hypothesis (such as $\beta_i = 1$) depends strongly on the prior for β_i over the unit interval, which we have thus far assumed is flat. Below we will investigate this dependence by entertaining a variety of priors.

Let M_0 denote the model that allows β_i to vary over the unit interval and let M_1 denote the model that restricts β_i to equal one. The fundamental difference between the two models can be characterized in terms of the probability of a unit root:

$$\Pr[\beta_i = 1|M_0] = 0 \qquad \text{and} \qquad \Pr[\beta_i = 1|M_1] = 1. \tag{3.5}$$

The labeling of the models reflects the probability of a unit root and not any notion of *null* or *alternative* hypotheses.

Bayes' rule can be applied at many levels. In (3.1), it is applied at the level of alternative values for the parameter β_i. For the purpose of model comparison, Bayes' rule is applied at the level of the two alternative models M_0 and M_1. Applied to M_0 and M_1, Bayes' rule expresses the posterior probability of model M_j in terms of its likelihood $p(Y_i|M_j)$, its prior probability $p(M_j)$, and the likelihood of the data according to the collection of models under consideration (in the denominator):

$$p(M_j|Y_i) = \frac{p(Y_i|M_j)\, p(M_j)}{p(Y_i|M_0)\, p(M_0) + p(Y_i|M_1)\, p(M_1)}. \tag{3.6}$$

The posterior odds ratio in favor of M_1 relative to M_0 is given by $p(M_1|Y_i)/p(M_0|Y_i)$. Using (3.6), the posterior odds ratio can be expressed in terms of the prior odds ratio, $p(M_1)/p(M_0)$, and the ratio of the likelihoods of the two models:

$$\frac{p(M_1|Y_i)}{p(M_0|Y_i)} = \frac{p(M_1)}{p(M_0)} \times \frac{p(Y_i|M_1)}{p(Y_i|M_0)}. \tag{3.7}$$

The ratio of the likelihoods is called the Bayes factor. Let us assume

$$p(M_0) = p(M_1) = 1/2, \tag{3.8}$$

in which case the prior odds ratio equals one. Then the posterior odds ratio equals the Bayes factor and the posterior probability of the model with a unit root is

$$p(M_1|Y_i) = \frac{p(Y_i|M_1)}{p(Y_i|M_0) + p(Y_i|M_1)}. \tag{3.9}$$

We now apply this framework for model comparison in light of the flat prior we have assumed for β_i. The likelihoods of the two models have already been calculated:

$$p(Y_i|M_0) = \int p(Y_i|\beta_i)\, d\beta_i = g_i \tag{3.10}$$

$$p(Y_i|M_1) = P(Y_i|\beta_i = 1) = h_i. \tag{3.11}$$

The Bayes factor in favor of the unit-root model is[4]

$$\mathcal{B}_i = \frac{p(Y_i|M_1)}{p(Y_i|M_0)} = \frac{P(Y_i|\beta_i = 1)}{\int_0^1 p(Y_i|\beta_i)\, d\beta_i} = \frac{h_i}{g_i}. \tag{3.12}$$

The posterior probability of the model with a unit root is

$$p(M_1) = \frac{h_i}{g_i + h_i}. \tag{3.13}$$

See Table 4 for the specific values of $z_i := h_i/g_i$.

Model averaging. Although we have cast the discussion in terms of model choice, it is not necessary to adopt one of the two models and discard the other. An alternative approach known as Bayesian Model Averaging (BMA) combines the two models. The posterior distribution for β_i becomes a weighted average of the two models, using the posterior model probabilities as the weights:

$$p(\beta_i|Y_i) = p(\beta_i|M_0, Y_i)\, p(M_0|Y_i) + p(\beta_i|M_1, Y_i)\, p(M_1|Y_i). \tag{3.14}$$

For future reference, it is convenient to write this posterior distribution as

$$p(\beta_i|Y_i) = \begin{cases} p(M_1|Y_i) & \beta_i = 1 \\ \big(1 - p(M_1|Y_i)\big)\, p(\beta_i|M_0, Y_i) & \beta_i < 1 \end{cases}, \tag{3.15}$$

where $p(M_1|Y_i)$ is the posterior probability of a unit root and $p(\beta_i|M_0, Y_i)$ is the posterior density over the unit interval. Given the prior in the example [see (3.3) and (3.8)], this becomes

$$p(\beta_i|Y_i) = \begin{cases} \dfrac{h_i}{h_i + g_i} & \beta_i = 1 \\ \dfrac{g_i}{h_i + g_i} \left(\dfrac{p(Y_i|\beta_i)}{g_i} \right) & \beta_i < 1 \end{cases}. \tag{3.16}$$

Thus $h_i/(h_i + g_i)$ is the posterior probability of a unit root for β_i and $p(Y_i|\beta_i)/g_i$ is the posterior density for β_i over the unit interval.

Empirical comparisons. The comparisons in terms of the Bayes factors (and posterior probabilities) are computed from the same Student t distributions as were used for the Dickey–Fuller tests. However, the information that is being extracted from them is quite different. The Bayes factor uses the likelihood of the data, which involves the density at $\beta_i = 1$ and the average density over the unit interval.

For the LT data, the unit-root model is not favored in either case. By contrast, for the Euro 1 data, the unit-root model is favored in 3 of 10 cases, while for the Euro 2 data it is favored in 4 of 10 cases. Thus for the Euro-related data as a whole, the unit-root model is favored in 35% of the cases. The strength of the evidence varies substantially across the cases, ranging from a factor of more than 15 in favor to a factor of more than 13 against.

We now turn to comparing the two models on a group-wise basis. When applied to a group of cases, model M_0 asserts that no exchange rate has a unit root while model M_1 asserts that every exchange rate does. (In later sections we will examine more general models in which some exchange rates in a group may have unit roots while others may not.)

[4]Given the current setup, the Bayes factor equals the density of the posterior at $\beta_i = 1$: $\mathcal{B}_i = p(\beta_i = 1|Y_i)$.

The likelihood of a group of independent cases is simply the product of individual case likelihoods. The Bayes factor in favor of the model with all unit roots relative to the model with no unit roots is given by $\prod_{i=1}^{n} \mathcal{B}_i = \prod_{i=1}^{n} z_i$. We see that with this comparison of two simple models, we arrive at a somewhat different interpretation of the data. For all three data sets, the model of no unit roots is preferred to the model of all unit roots. For the two Euro-related data sets, the ratio is 0.025.

Sensitivity to the prior. As we noted above, the Bayes factor can be quite sensitive to the prior for β_i over the unit interval. Here we examine that issue in some detail.

Consider a prior of the following form:

$$p(\beta_i|a,b) = \mathsf{Beta}(\beta_i|a,b) = 1_{[0,1]}(\beta_i)\,\frac{\beta_i^{a-1}\,(1-\beta_i)^{b-1}}{B(a,b)}, \tag{3.17}$$

where $B(a,b)$ is the beta function, $B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}\,(1-x)^{b-1}\,dx$. We have indexed the prior by the parameters of the beta distribution. The mean of β_i according to this prior is $a/(a+b)$ and the variance is $a\,b/\big((a+b)^2\,(a+b+1)\big)$. The uniform distribution is a special case of this prior: $p(\beta_i|a=1,b=1) = \mathsf{Uniform}(\beta_i|0,1)$. The model indexed by (a,b) will take the role of M_0 in model comparison and model averaging. We will refer to this as the *base model*. Note $\Pr[\beta_i = 1|(a,b)] = 0$.

The posterior for β_i can be expressed conditional on (a,b):

$$p(\beta_i|Y_i,a,b) = \frac{p(Y_i|\beta_i)\,p(\beta_i|a,b)}{p(Y_i|a,b)} \tag{3.18}$$

where the likelihood of the data according to the model (a,b) is[5]

$$p(Y_i|a,b) = \int p(Y_i|\beta_i)\,p(\beta_i|a,b)\,d\beta_i. \tag{3.19}$$

The Bayes factor in favor of the model with a unit root relative to the base model indexed by (a,b) is:

$$\mathcal{B}_i(a,b) := \frac{p(Y_i|\beta_i=1)}{p(Y_i|a,b)} = \frac{h_i}{p(Y_i|a,b)}. \tag{3.20}$$

The first thing to note is that the numerator of the Bayes factor is fixed at $p(Y_i|\beta_i = 1)$. Consequently the Bayes factor varies inversely with the likelihood of the base model (a,b). Therefore, a moderately-well informed prior will produce a higher likelihood for the base model and will reduce the Bayes factor in favor a unit root; similarly, an ill-informed prior that reduces the likelihood of the base model will increase the Bayes factor in favor of a unit root.

We do not have space to provide a complete investigation. Instead we touch on a few special and/or interesting cases. First, since (as noted above) the uniform distribution is a special case of the beta distribution, $p(Y_i|1,1) = g_i$ and $\mathcal{B}_i(1,1) = z_i$.

Next we consider two limiting cases: as $a \to \infty$ all of the probability of $\mathrm{Beta}(\beta_i|a,b)$ becomes concentrated on $\beta_i = 1$, while as $b \to \infty$ all of the probability becomes concentrated on $\beta_i = 0$. In the first case, the base model converges to the unit-root model and

[5]The likelihood $p(Y_i|a,b)$ will appear prominently in the more general model described below.

consequently

$$\lim_{a\to\infty} \mathcal{B}_i(a,b) = \frac{p(Y_i|\beta_i=1)}{p(Y_i|\beta_i=1)} = 1. \tag{3.21}$$

In the second case, the base model converges to a model with no serial correlation:

$$\lim_{b\to\infty} \mathcal{B}_i(a,b) = \frac{p(Y_i|\beta_i=1)}{p(Y_i|\beta_i=0)}, \tag{3.22}$$

This comparison favors a unit root for all but one of our exchange rates (in Euro 2), typically by astronomical amounts. [Need to include these numbers.]

Finally, consider the following prior:[6] Let $\beta_i = \beta_i^*$ where

$$\beta_i^* := \operatorname*{argmax}_{\beta_i} \; p(Y_i|\beta_i) \qquad \text{subject to } \beta_i \in [0,1]. \tag{3.23}$$

For our data, $\beta_i^* = m_i$, where m_i is given in Table 1. Let $d_i := p(Y_i|\beta_i^*)$. [See Table 4 for specific values for d_i.] The model based on this prior provides the maximum evidence against the unit-root model because it concentrates its prior on the maximum (allowed) likelihood value. The Bayes factor in favor of the unit-root model relative to this model is

$$\mathcal{B}_i^* = \frac{p(Y_i|\beta_i=1)}{p(Y_i|\beta_i=\beta_i^*)} = \frac{h_i}{d_i}. \tag{3.24}$$

Table 4 displays values for $1/\mathcal{B}_i^*$, the maximum evidence against a unit root.

Summary and limitations. Given the current setup, we are able to learn about β_i from Y_i. The prior for β_i plays an important role in what we learn: It affects the posterior distribution over the unit interval and it affects the Bayes factor in favor of a unit root. In addition, the prior odds ratio plays an important role in determining to posterior odds ratio in favor of a unit root. We have seen, for example, that ill-informed priors can produce strong evidence in favor of a unit root.

Presumably one should adopt a moderately well-informed prior. But how does one acquire such a prior? Unfortunately within the current setup it is not possible: There is no way to learn about β_j from Y_i. Such learning must come from the dependence between β_i and β_j in the prior. In the following section we provide a framework that has this feature and therefore allows one to apply what one learns about one exchange rate to another and thereby acquire a moderately well-informed prior.

4. Framework for learning

Bayesian inference involves learning. In Section 3 we saw how to learn about the coefficient β_i from the data Y_i. But, as we noted, there was no learning about β_j from Y_i and consequently it was not possible to obtain a reasonably well-informed prior for β_{n+1} given the data $Y_{1:n}$. In this section, we show how to extend the framework to allow for this broader type of learning — learning across regimes. The framework involves prior dependence between β_{n+1} and $\beta_{1:n}$.

[6] This prior can be constructed as a limiting value of a beta prior as well.

Density estimation. At the heart of our approach is the idea that we can learn about one coefficient from other coefficients. For example, if we observed a collection of coefficients then we could produce a reasonably-well informed prior for another — as yet unobserved — coefficient. Of course the coefficients are not observed directly; they are observed only indirectly and with error. Nevertheless, it is useful to begin as if we did indeed observe a collection of coefficients, $\beta_{1:n}$. Let β_{n+1} denote the (as yet unobserved) coefficient for which we seek a reasonably-well informed prior. Given suitable assumptions (described below), we can form a distribution for β_{n+1} based on $\beta_{1:n}$,

$$p(\beta_{n+1}|\beta_{1:n}). \tag{4.1}$$

Since we are assuming the coefficients $\beta_{1:n}$ are observed, (4.1) would be called the *posterior predictive* distribution. Computing this predictive distribution amounts to an exercise in Bayesian *density estimation*. For density estimation, it is natural to adopt assumptions that allow for a wide range of possible distributions, including the possibility of multi-modality.

Indirect density estimation. Having indicated a framework for learning about β_{n+1} when a collection of coefficients is observed, we now expand the framework to account for the fact that the coefficients are not actually observed directly — they are *latent variables*. It is useful to think of the information available regarding the coefficients as a collection of "noisy signals" for each $\beta_i \in \beta_{1:n}$.[7] The noisy signals constitute the (joint) likelihood $p(Y_{1:n}|\beta_{1:n})$. Within this setup, distribution (4.1) provides the link between the coefficient for which we have no signal and those for which we do have signals:

$$p(\beta_{n+1}|Y_{1:n}) = \int p(\beta_{n+1}|\beta_{1:n})\, p(\beta_{1:n}|Y_{1:n})\, d\beta_{1:n}, \tag{4.2}$$

where $p(\beta_{1:n}|Y_{1:n})$ is the posterior distribution for $\beta_{1:n}$ given $Y_{1:n}$. Note that $p(\beta_{n+1}|Y_{1:n})$ is the reasonably-informed prior we seek. It is the result of *indirect* density estimation.

In its role in (4.2) as a link, the distribution $p(\beta_{n+1}|\beta_{1:n})$ has the interpretation of a conditional prior. The necessity of prior dependence (among the coefficients) for leaning is clear in (4.2). In particular, if β_{n+1} were independent of $\beta_{1:n}$ in the prior, then nothing would be learned about β_{n+1} since in that case $p(\beta_{n+1}|Y_{1:n}) = p(\beta_{n+1})$.

Generic and specific cases. When we assumed $\beta_{1:n}$ was observed, there was an obvious asymmetry between $\beta_i \in \beta_{1:n}$ and β_{n+1}. This asymmetry carries over to the situation where $\beta_{1:n}$ is latent. In particular, we observe signals for $\beta_i \in \beta_{1:n}$ but we do not observe a signal for β_{n+1}. Based on this distinction, we refer to β_i as a *specific case* because we have a (specific) signal for it and we refer to β_{n+1} as the *generic case* because it applies to *any* coefficient for which we as yet have no signal (and for which we judge $\beta_{1:n}$ to provide an appropriate basis for inference). The specific cases are the ones that appear in the likelihood:

$$p(Y_{1:n}|\beta_{1:n+1}) = p(Y_{1:n}|\beta_{1:n}). \tag{4.3}$$

The generic case does not appear and therefore is not identified.

Equation (4.2) gives a representation of the posterior distribution for the generic case in terms of the posterior distribution for $\beta_{1:n}$. In addition, we are interested in the marginal

[7] In passing, note that as the number of observations relevant to β_i increases [i.e., as T_i increases], the noise in the measurement gets smaller and in the limit we end up where β_i is effectively observed.

posterior distributions for each of the specific cases. Here we present an intuitive representation that relies on the likelihood factoring as in (2.6). The general case is somewhat complicated and is treated in Appendix D.

When the likelihood factors, the posterior distribution for a specific case can be written as follows:

$$p(\beta_i|Y_{1:n}) = \frac{p(Y_i|\beta_i)\, p(\beta_i|Y_{1:n}^{-i})}{p(Y_i|Y_{1:n}^{-i})}, \tag{4.4}$$

where $Y_{1:n}^{-i}$ excludes Y_i so that β_i is generic with respect to $Y_{1:n}^{-i}$. Since the posterior is an average of the likelihood and the prior, the posterior for β_i is "shrunk" towards the generic distribution based on all the other data.

Density estimation. The model for density estimation adopted here is a Dirichlet Process Mixture (DPM) model. The DPM model can be expressed in terms of a stick-breaking prior.[8] The foundational details do not concern us here. A textbook treatment of the DPM may be found in Gelman et al. (2014).

The density estimate (4.1) can be expressed as

$$p(\beta_{n+1}|\beta_{1:n}) = \int p(\beta_{n+1}|\psi)\, p(\psi|\beta_{1:n})\, d\psi, \tag{4.5}$$

where ψ is a (hyper)parameter. The model is completed by specifying the prior $p(\psi)$ and the likelihood

$$p(\beta_{1:n}|\psi) = \prod_{i=1}^{n} p(\beta_i|\psi). \tag{4.6}$$

The individual likelihood $p(\beta_i|\psi)$ is an infinite-order mixture:

$$p(\beta_i|\psi) = \sum_{c=1}^{\infty} v_c\, f(\beta_i|\theta_c), \tag{4.7}$$

where $\psi = (v, \theta)$ and $v = (v_1, v_2, \ldots)$ denotes an infinite collection of nonnegative mixture weights that sum to one and $\theta = (\theta_1, \theta_2, \ldots)$ denotes a corresponding collection of mixture-component parameters. The density $f(\,\cdot\,|\,\cdot\,)$ is called the *kernel*.

The prior for ψ can be expressed as

$$p(\psi) = p(v)\, p(\theta) = p(v) \prod_{c=1}^{\infty} p(\theta_c). \tag{4.8}$$

The prior for θ_c is called the *base distribution*. We present the kernel and base distribution in Section 5.

[8]See Ishwaran and James (2001) for a general characterization of stick-breaking priors and associated Gibbs samplers.

Prior for the mixture weights. The standard prior for v is given by

$$v|\eta \sim \mathsf{Stick}(\eta), \tag{4.9}$$

where $\mathsf{Stick}(\eta)$ denotes the stick-breaking distribution given by[9]

$$v_c = u_c \prod_{\ell=1}^{c-1}(1-u_\ell) \qquad \text{where } u_c \overset{\text{iid}}{\sim} \mathsf{Beta}(1,\eta). \tag{4.10}$$

The parameter η is called the *concentration parameter*; it controls the rate at which the weights decline on average. In particular, the weights decline geometrically in expectation:

$$E[v_c|\eta] = \eta^{c-1}\,(1+\eta)^{-c}. \tag{4.11}$$

Note $E[v_1|\eta] = 1/(1+\eta)$ and $E\big[\sum_{c=m+1}^{\infty} v_c|\eta\big] = \big(\eta/(1+\eta)\big)^m$.

The number of components in (4.7) required to represent a given distribution depends on the number of modes and the shape of the modes. The concentration parameter plays an important role in determining the effective number of components. If η is small, then the first few weights will dominate and the number of mixture components with nontrivial probabilities will be small. In the limit as $\eta \to 0$, the mixture collapses to a single component. By contrast if η is large, then no single component (or small collection of components) will dominate and more components will be available (on average).

Prior for the concentration parameter. Because the concentration parameter plays an important role in determining the flexibility of the prior for a given finite sample size n, it is be important to allow the data to help determine its magnitude. We adopt the following prior:

$$p(\eta) = \mathsf{Log\text{-}Logistic}(\eta|1,1) = \frac{1}{(1+\eta)^2}. \tag{4.12}$$

This distribution does not have a finite mean; its median equals one. This prior implies $(\eta/(1+\eta))^m \sim \mathsf{Beta}(1/m,1)$.

Taking the prior for the concentration parameter into account, we restate the marginal prior for the mixture weights as $p(v) = \int p(v|\eta)\,p(\eta)\,d\eta$.

5. Kernel and base distribution

In this section we specify the kernel $f(\beta_i|\theta_c)$ and the base distribution $p(\theta_c)$, each of which introduces some novelty. The kernel and the base distribution are based on Fisher (2017) with extensions that accommodate a point mass.

[9]Start with a stick of length one. Break off the fraction u_1 leaving a stick of length $1-u_1$. Then break off the fraction u_2 of the remaining stick leaving a stick of length $(1-u_1)\,(1-u_2)$. Continue in this manner. Alternative stick-breaking distributions can be constructed by changing the distribution for u_c.

Kernel. We specify the conditional prior for β_i (i.e., the kernel). Let $\theta_c = (a_c, b_c, w_c)$ and let

$$f(\beta_i|\theta_c) = f(\beta_i|a_c, b_c, w_c) = \begin{cases} w_c & \beta_i = 1 \\ (1 - w_c)\,\mathsf{Beta}(\beta_i|a_c, b_c) & \beta_i < 1 \end{cases}. \tag{5.1}$$

This distribution includes the probability of a unit root, $w_c \in [0, 1]$, and a density over the unit interval, $\mathsf{Beta}(\beta_i|a_c, b_c)$, where a_c and b_c are positive.[10] This distribution is sometimes called a *one-inflated* beta distribution.

Base distribution. We begin by assuming prior independence between (a_c, b_c) and w_c:

$$p(a_c, b_c, w_c) = p(a_c, b_c)\, p(w_c). \tag{5.2}$$

Let[11]

$$p(w_c) = \mathsf{Beta}\big(w_c|\zeta\,\phi, \zeta\,(1 - \phi)\big). \tag{5.3}$$

Note $E[w_c] = \int_0^1 w_c\, p(w_c)\, dw_c = \phi$. In addition, $E[(w_c - \phi)^2] = \phi\,(1 - \phi)/(1 + \zeta)$. The uniform distribution is delivered by $\phi = 1/2$ and $\zeta = 2$. In the limit as $\zeta \to \infty$ the prior for w collapses to a point mass located at ϕ, in which case there is no learning about w_c. At the other extreme, as $\zeta \to 0$ the prior degenerates to a pair of point masses located at 0 and 1.[12]

In order to specify the prior for (a_c, b_c) it is convenient to change variables to

$$(j_c, k_c) = (a_c, a_c + b_c - 1). \tag{5.4}$$

We adopt the prior $p(j_c, k_c) = p(j_c|k_c)\, p(k_c)$, where

$$p(j_c|k_c) = \begin{cases} 1/k_c & j_c \in \{1, \ldots, k_c\} \\ 0 & \text{otherwise} \end{cases}. \tag{5.5}$$

Let $k_c - 1 \sim \mathsf{Geometric}(\xi)$, so that $p(k_c) = \xi\,(1 - \xi)^{k_c - 1}$ for k_c.[13] Then the prior for (a_c, b_c) is

$$p(a_c, b_c) = p(j_c, k_c|I)|_{j_c = a_c, k_c = a_c + b_c - 1} = \frac{\xi\,(1 - \xi)^{a_c + b_c - 2}}{a_c + b_c - 1}. \tag{5.6}$$

Whenever it is convenient we will adopt the parameterization $\theta_c = (j_c, k_c, w_c)$ in place of $\theta_c = (a_c, b_c, w_c)$. A key feature of the prior is that the expectation of the density component

[10] See Appendix C for a brief discussion of the representation of densities involving mutually singular measures.

[11] The unit root component can be removed via the restriction $w_i \equiv 0$.

[12] If used in conjunction with $\eta = 0$ (so that all cases share the same w_c), this latter prior would be appropriate for an all-or-none view of unit roots: Either all cases have unit roots or no case has.

[13] Note $E[k_c] = 1/\xi$.

given k_c is uniform:

$$\begin{aligned} E[\mathsf{Beta}(\beta_i|j_c, k_c - j_c + 1)|k_c] &= \sum_{j_c=1}^{k_c} \mathsf{Beta}(\beta_i|j_c, k_c - j_c + 1)\, p(j_c|k_c) \\ &= \frac{1}{k_c} \sum_{j_c=1}^{k_c} \mathsf{Beta}(\beta_i|j_c, k_c - j_c + 1) \\ &= \mathsf{Uniform}(\beta_i|0, 1). \end{aligned} \tag{5.7}$$

A generalization that allows for alternative prior predictive distributions is discussed in Appendix H.

We are now equipped to express the marginal prior for β_i:

$$\begin{aligned} p(\beta_i) = \int f(\beta_i|\theta_c)\, p(\theta_c)\, d\theta_c &= \begin{cases} E[w_c] & \beta_i = 1 \\ (1 - E[w_c])\, \mathsf{Uniform}(\beta_i|0, 1) & \beta_i < 1 \end{cases} \\ &= \begin{cases} 1/2 & \beta_i = 1 \\ 1/2 & \beta_i < 1 \end{cases}. \end{aligned} \tag{5.8}$$

The second line of (5.8) assumes $\phi = 1/2$. The marginal prior over the unit interval is flat, which follows from (5.7). Non-flat priors can be obtained by modifying the kernel as shown in (H.1).

Comment. It is instructive to examine the posterior distribution for $\beta_i|Y_i$ in isolation (i.e., without any other regimes from which to learn). Referring to (4.4), we have

$$p(\beta_i|Y_i) = \frac{p(Y_i|\beta_i)\, p(\beta_i)}{\int p(Y_i|\beta_i)\, p(\beta_i)\, d\beta_i}, \tag{5.9}$$

where $p(\beta_i)$ is given in (5.8). This prior [i.e., (5.8)] is equivalent to the prior used in the Bayesian Model Averaging example in Section 3 as [see (3.3) and (3.8)]. Therefore, $p(\beta_i|Y_i)$ is given in (3.16). Thus, for a single coefficient in isolation, we obtain the same inferences as from the BMA example in Section 3.

Features of the prior. It may be useful to understand some features of the prior. The prior encodes both a willingness to learn (via dependence) and open-mindedness (via flexibility).

Dependence. Dependence in the prior among the betas is the key to the ability to learn about β_{n+1} from $\beta_{1:n}$. Without this dependence there is no learning. We now examine how this dependence is structured within the prior by focusing on the joint prior distribution for (β_1, β_2), which we derive in stages. We begin with

$$\begin{aligned} p(\beta_1, \beta_2|v) &= \int p(\beta_1|v, \theta)\, p(\beta_2|v, \theta)\, p(\theta)\, d\theta \\ &= \left(\sum_{c=1}^{\infty} v_c^2\right) \int f(\beta_1|\theta_c)\, f(\beta_2|\theta_c)\, p(\theta_c)\, d\theta_c + \left(1 - \sum_{c=1}^{\infty} v_c^2\right) p(\beta_1)\, p(\beta_2), \end{aligned} \tag{5.10}$$

TABLE 5. Various probabilities according the joint prior predictive distribution (5.13).

Condition	Probability
$\beta_1 = 1 \ \wedge \ \beta_2 = 1$	$1/2\,(1/3) + 1/2\,(1/4) = 7/24$
$\beta_1 < 1 \ \wedge \ \beta_2 < 1$	$1/2\,(1/3) + 1/2\,(1/4) = 7/24$
$\beta_1 < 1 \ \wedge \ \beta_2 = 1$	$1/2\,(1/6) + 1/2\,(1/4) = 5/24$
$\beta_1 = 1 \ \wedge \ \beta_2 < 1$	$1/2\,(1/6) + 1/2\,(1/4) = 5/24$

where $\sum_{c=1}^{\infty} v_c^2$ is the probability that β_1 and β_2 share the same component. As noted above, $E[\sum_{c=1}^{\infty} v_c^2|\eta] = 1/(1+\eta)$. Consequently,

$$\begin{aligned} p(\beta_1, \beta_2|\eta) &= \int p(\beta_1, \beta_2|v)\, p(v|\eta)\, dv \\ &= \frac{1}{1+\eta} \int f(\beta_1|\theta_c)\, f(\beta_2|\theta_c)\, p(\theta_c)\, d\theta_c + \frac{\eta}{1+\eta}\, p(\beta_1)\, p(\beta_2). \end{aligned} \tag{5.11}$$

Note that β_1 and β_2 become independent as $\eta \to \infty$.

Given our prior for η [see (4.12)], the unconditional probability that β_1 and β_2 share the same component is

$$\int_0^{\infty} \frac{1}{1+\eta}\, p(\eta)\, d\eta = \int_0^{\infty} \frac{1}{(1+\eta)^3}\, d\eta = \frac{1}{2}. \tag{5.12}$$

Therefore,

$$p(\beta_1, \beta_2) = \frac{1}{2} \int f(\beta_1|\theta_c)\, f(\beta_2|\theta_c)\, p(\theta_c)\, d\theta_c + \frac{1}{2}\, p(\beta_1)\, p(\beta_2). \tag{5.13}$$

See Table 5 for various probabilities according to (5.13). Assuming both β_1 and β_2 are less than one, the joint density is

$$p(\beta_1, \beta_2) = \sum_{k_c=1}^{\infty} p(\beta_1, \beta_2|k_2)\, p(k_c). \tag{5.14}$$

A closed-form expression is not available. Instead we can examine a single term in which

$$\begin{aligned} p(\beta_1, \beta_2|k_c) &= \frac{1}{2} + \frac{1}{2\,k_c} \sum_{j_c=1}^{k_c} \mathsf{Beta}(\beta_1|j_c, k_c - j_c + 1)\, \mathsf{Beta}(\beta_2|j_c, k_c - j_c + 1) \\ &= \frac{1}{2} + \frac{k_c}{2} \left((1-\beta_1)\,(1-\beta_2)\right)^{k_c - 1}\, {}_2F_1\left(1 - k_c, 1 - k_c; 1; \frac{\beta_1\,\beta_2}{(1-\beta_1)\,(1-\beta_2)}\right), \end{aligned} \tag{5.15}$$

where ${}_2F_1$ is the hypergeometric function. For $k_c = 1$ the distribution is uniform on the unit square. For $k_c > 1$, β_1 and β_2 are positively related; the strength of the dependence is increasing in k_c. In all cases, the marginals are uniform: $\int_0^1 p(\beta_1, \beta_2|k_c)\, d\beta_1 = 1$. See Figure 4 for a plot of $p(\beta_1, \beta_2|k_c = 10)$.

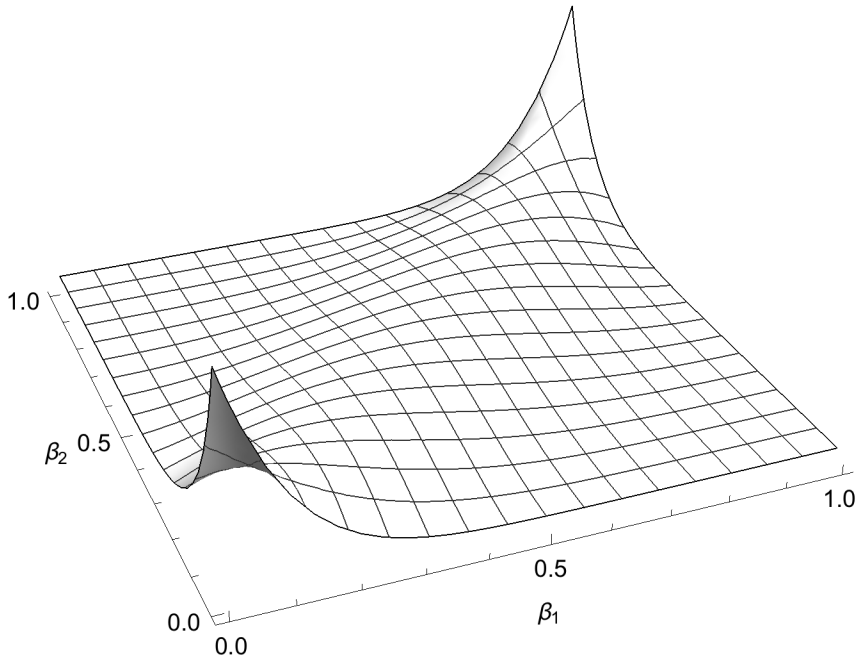

FIGURE 4. $p(\beta_1, \beta_2 | k_c = 10)$ given $\beta_1 < 1 \wedge \beta_2 < 1$.

Open-mindedness. An open-minded prior allows for substantial variation around the prior predictive distribution. The prior predictive distribution is given in (5.8). Variation around it can be examined as follows. Make draws $\{\psi^{(r)}\}_{r=1}^{R}$ from the prior, where $\psi^{(r)} \overset{\text{iid}}{\sim} p(\psi)$. The prior predictive can be approximated by

$$p(\beta_i) \approx \frac{1}{R} \sum_{r=1}^{R} p(\beta_i | \psi^{(r)}). \tag{5.16}$$

For a subset of the draws, plot $p(\beta_i|\psi^{(r)})$ to examine the amount and sort of variation. In Figure 5, we display ten draws of the density $p(\beta_i|\psi)$ given $\beta_i < 1$.

6. MCMC sampler

In the previous sections we described the likelihood and the prior. In this section we show how to compute the posterior distribution using a a Markov Chain Monte Carlo (MCMC) sampler. We adopt a Gibbs-sampler approach, decomposing an unweildly high-dimensional joint distribution into a collection of manageable lower-dimensional conditional distributions.

The unknowns are $\beta_{1:n}$, ψ, and η. To this list we add another. As is typical when dealing with mixture models, it is convenient to introduce classification variables $z_{1:n} = (z_1, \ldots, z_n)$,

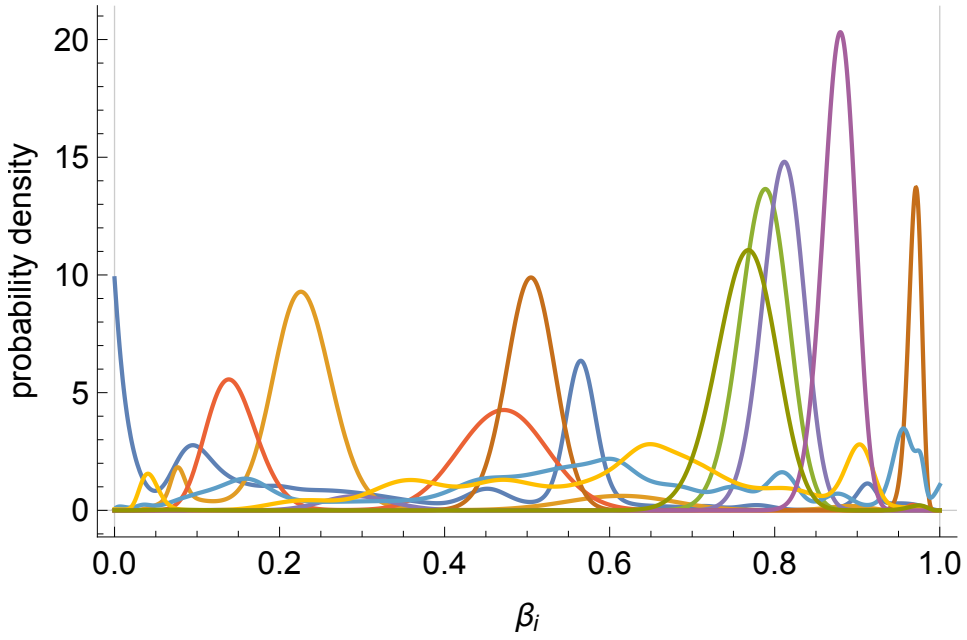

FIGURE 5. Illustrating one aspect of an open-minded prior: $p(\beta_i|\psi)$ given $\beta_i < 1$ is plotted for each of ten draws from $p(\psi)$. Note that $p(\beta_i) = 1$.

where $z_i = c$ signifies β_i is assigned to cluster c. The joint posterior distribution for the augmented collection of unknowns is

$$p(\beta_{1:n}, \psi, \eta, z_{1:n}|Y_{1:n}). \tag{6.1}$$

This distribution is completely characterized by the following two full conditional distributions:

$$p(\psi, \eta, z_{1:n}|Y_{1:n}, \beta_{1:n}) = p(\psi, \eta, z_{1:n}|\beta_{1:n}) \tag{6.2a}$$

$$p(\beta_{1:n}|Y_{1:n}, \psi, \eta, z_{1:n}) = \prod_{i=1}^{n} p(\beta_i|Y_i, \psi, z_{1:n}). \tag{6.2b}$$

Note that $\beta_{1:n}$ is treated as known in (6.2a) while it is treated as unknown (6.2b). The Gibbs sampler alternates between (6.2a) and (6.2b). The right-hand sides of (6.2) indicate simplifications that play important roles in the sampler. In particular, on the right-hand side of (6.2a) the data are absent and on the right-hand side of (6.2b) the conditional distributions for β_i are independent.[14]

Step 1. We begin with (6.2a).

As noted above, the prior we have adopted is equivalent to a Dirichlet Process Mixture (DPM) model. As such draws from (6.2a) may be computed via any number of existing algorithms. The simplest algorithm to describe and implement is the blocked Gibbs sampler

[14]This latter factorization relies on the the joint likelihood factoring. The case where the likelihood does not factor is treated in the Appendices.

described in Gelman et al. (2014, pp. 552–553). This sampler relies on approximating $p(\beta_i|\psi)$ with a finite sum: Choose m large enough to make $\big(\eta/(1+\eta)\big)^m$ close enough to zero (on average) and set $u_m = 1$.

The distribution $p(\psi, \eta, z_{1:n}|\beta_{1:n})$ itself is completely characterized by the following four full conditional distributions:

$$p(z_{1:n}|\beta_{1:n}, v, \theta, \eta) = \prod_{i=1}^{n} p(z_i|\beta_i, v, \theta) \tag{6.3a}$$

$$p(\theta|\beta_{1:n}, z_{1:n}, v, \eta) = \prod_{c=1}^{m} p(\theta_c|B_c) \tag{6.3b}$$

$$p(v|\beta_{1:n}, z_{1:n}, \theta, \eta) = p(v|z_{1:n}, \eta) \tag{6.3c}$$

$$p(\eta|\beta_{1:n}, z_{1:n}, v, \theta) = p(\eta|z_{1:n}), \tag{6.3d}$$

where $B_c = \{\beta_i \in \beta_{1:n} : z_i = c\}$, the collection of observations for which $z_i = c$. Let $n_c = |B_c|$, the number of times c occurs in $z_{1:n}$. Note $\sum_{c=1}^{m} n_c = n$. (As a diagnostic, m should be large enough that $n_c = 0$ for some c. If not, then increase m.)

The Gibbs sampler cycles through the four distributions in (6.3). We comment briefly on each in turn. Omitted details can be found in Appendix E.

The distribution for $z_i = c \in \{1, \ldots, m\}$ is categorical, where the category probabilities are given by

$$p(z_i = c|\beta_{1:n}, v, \theta) \propto v_c\, f(\beta_i|\theta_c), \qquad \text{for } c = 1, \ldots, m. \tag{6.4}$$

The cluster parameters, $\theta_c|B_c$, are updated as in a finite mixture model, with the parameters for the unoccupied components (for which $n_c = 0$) sampled directly from the prior $p(\theta_c)$.

The weights v can be updated by updating the stick-breaking weights u via

$$u_c|z_{1:n} \sim \mathsf{Beta}\big(1 + n_c, \eta + \textstyle\sum_{c'=c+1}^{m} n_{c'}\big) \qquad \text{for } c = 1, \ldots, m-1. \tag{6.5}$$

(Recall $u_m = 1$.) Finally, regarding the concentration parameter η, refer to Appendix E.

Step 2. We now turn to (6.2b).

We have

$$p(\beta_i|Y_{1:n}, \psi, z_{1:n}) = p(\beta_i|Y_i, \theta_i) \propto p(Y_i|\beta_i)\, f(\beta_i|\theta_i), \tag{6.6}$$

where θ_i is shorthand notation for θ_{z_i}. These draws may be made using the Metropolis–Hastings scheme. See Appendix E for details on how to make draws in this case and other cases where the likelihood does not factor.

Posterior distributions. Given draws $\big\{(\beta_{1:n}^{(r)}, \psi^{(r)})\big\}_{r=1}^{R}$ from the posterior distribution, we can compute approximate posterior distributions for the generic and specific cases.

The generic distribution can be approximated via

$$\begin{aligned} p(\beta_{n+1}|Y_{1:n}) &= \int p(\beta_{n+1}|\psi)\, p(\psi|Y_{1:n})\, d\psi \\ &\approx \frac{1}{R}\sum_{r=1}^{R} p(\beta_{n+1}|\psi^{(r)}) \\ &\approx \frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} v_c^{(r)}\, f\big(\beta_{n+1}|\theta_c^{(r)}\big), \end{aligned} \tag{6.7}$$

where the choice of m is described in Step 1 above. In (6.7) the weight on the point mass is

$$\widehat{\pi} = \frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} v_c^{(r)}\, w_c^{(r)} \tag{6.8}$$

and the density component is

$$\frac{\frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} v_c^{(r)}\,(1-w_c^{(r)})\,\mathsf{Beta}(\beta_{n+1}|j_c^{(r)}, k_c^{(r)} - j_c^{(r)} + 1)}{1-\widehat{\pi}}. \tag{6.9}$$

Turning to specific distributions, an approximation of the posterior distribution for the i-th specific case can be obtained from a histogram of the draws $\{\beta_i^{(r)}\}_{r=1}^{R}$. A lower-variance approximation can be computed as follows (via Rao–Blackwellization):

$$p(\beta_i|Y_{1:n}) = \int p(\beta_i|Y_i,\theta_i)\, p(\theta_i|Y_{1:n})\, d\theta_i \approx \frac{1}{R}\sum_{r=1}^{R} p(\beta_i|Y_i,\theta_i^{(r)}). \tag{6.10}$$

Note that (6.10) assumes the likelihood factors as in (2.6). For the case where the likelihood does not factor see Appendix D [which see also for an explicit representation of $p(\beta_i|Y_i,\theta_i)$].

7. Empirical section [incomplete]

In this section, we compute the predictive distributions. But first we discuss some odds and ends.

Comment on robustness. An earlier version of this paper adopted a different prior and obtained very similar results. That prior can be expressed within the framework of Section 4 as follows. The kernel involves a truncated normal distribution:

$$f(\beta_i|\theta_c) = \begin{cases} w_c & \beta_i = 1 \\ (1-w_c)\,\mathsf{N}_{[0,1]}(\beta_i|\mu_c,\sigma_c^2) & \beta_i < 1. \end{cases} \tag{7.1}$$

The prior distribution $p(\mu_c,\sigma_c)$ is defined in terms of (m_c, s_c), the mean and standard deviation of the truncated distribution. The variables (m_c, s_c) are given a flat prior over the finite region of their support. (These specifications for the kernel and the base distribution produce a prior predictive distribution for β_i that is relatively flat over the unit interval.) The concentration parameter is set to zero ($\alpha = 0$), producing complete sharing. This prior allows for learning but imposes unimodality on the posterior predictive distributions. As it turns out, this restriction does not materially affect the results.

Nevertheless we now choose to adopt the more flexible approach described in Section 4 (in which $\alpha > 0$ is allowed). The reason for changing the kernel (from truncated normal to beta) is related in part to numerical issues. From a pratical standpoint the prior on the truncated normal is difficult to use because the transformation from (m_c, s_c) to (μ_c, σ_c) is numerically unstable for some important regions of the parameter space. Anyone who wishes to replicate our results will find the current prior much easier to work with.

Posterior distributions with independent likelihoods. Here we present the results given independent likelihoods.

Let $\xi = 1/200$ (so the prior mean for k_c is 200), and let $\zeta = 2$ and $\phi = 1/2$ (so the prior distribution for w_c is uniform). See Figures 6, 7, and 8 and Table 6.

Figure 6 shows the generic distributions computed from each of the three groups (LT, EURO 1, and EURO 2). They are remarkably similar to each other. Recall that the LT group is composed of two series each with 189 observations, while the EURO 2 group is composed of ten series each with 14 observations. The similarities in posterior densities and the point-mass probabilities indicate that the information about the auto-regressive coefficient in the two groups is similar. The generic distribution based on EURO 1 (composed of ten series with 42 observations) appears to contain more information than either of the other two (the density is more peaked and the point-mass is less probable), but is otherwise consistent them. These results suggest that it is sensible to aggregate all three into one large group of 22 series. The generic distribution based on this group (ALL) is shown in the figure are well.

Figure 7 shows 95% highest posterior density (HPD) intervals for the density component for each of the 22 specific distributions. For each specific distribution, there are three HPD intervals shown that differ according to what information from other regimes was used in the prior. As a baseline, the first HPD interval is computed from the truncated Student t distribution for the given regime without regard to any other regime. The second interval is computed based on the regime's group, and the third interval is computed using information from all regimes. Comparing the second and third intervals to the first, one sees a substantial amount of shrinkage.

The posterior probabilities for a unit root are shown in Figure 8 and displayed in Table 6. Regime 2 is special, in that the probability of a unit root ALONE is sufficiently small that it is difficult to assess the change in probability due to adding information from other regimes (at the resolutions in the figure and the table). In all but one of the other regimes (regime 15), the probability of a unit root decreases when adding the group information. Further reduction occurs in all regimes (still excluding regime 2) when the posterior is based on ALL.

An additional comment is in order. In Section 3 we noted that for three exchange rates there can exist no prior that provides evidence against a unit root. Yet we see that one of these (regime 14) has a posterior probability of a unit root less that 1/2. To understand this result, refer to (4.4). Relative to the prior $p(\beta_i|Y_{1:n}^{-i})$, these exchange rates provide evidence in *favor* of a unit root; but this prior — which is computed from all other data — has a probability of a unit root that is substantially less than 1/2. This case illustrates what can happen on the journey from an open-minded prior, $p(\beta_i|\beta_{1:n}^{-i})$, to a well-informed prior, $p(\beta_i|Y_{1:n}^{-i})$.

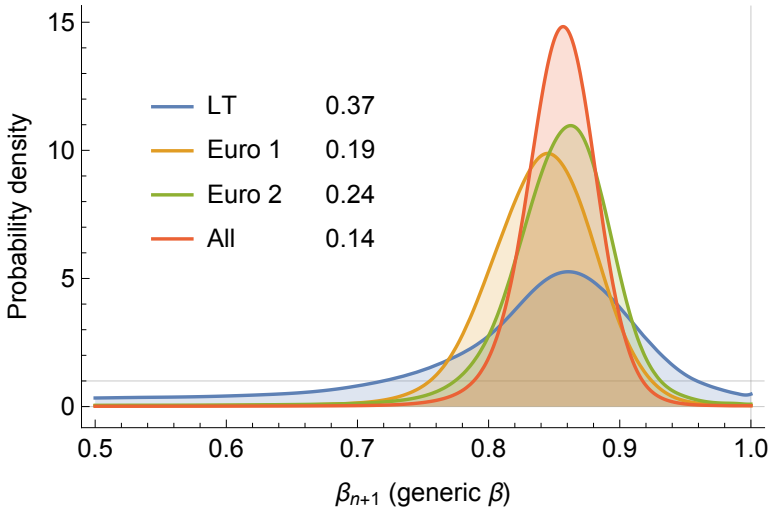

FIGURE 6. Generic posterior distributions by group. Posterior probability of a unit root is indicated in the legend.

Results using dependent likelihoods. [This subsection is incomplete.]

Here we present some results using the SUR likelihood. We have had some difficulty estimating the model for EURO 2 and consequently we report result for LT and EURO 1. Compare Figure **??** with Figure 6. The only noticeable difference is that the probability of a unit root has increased from 0.21 to 0.27. Indeed, for the LT data a 90% HPD interval for the correlation coefficient is about $[-0.05, 0.19]$.

APPENDIX A. SUFFICIENT STATISTICS FOR β_i

In this section, we derive the sufficient statistics for the marginal likelihood for β_i [see (2.5)]. For this purpose it is convenient to adopt the following notation:

$$Y = X\xi + \varepsilon, \tag{A.1}$$

where Y is $\mathcal{T} \times 1$, X is $\mathcal{T} \times \mathcal{K}$, ξ is $\mathcal{K} \times 1$, and ε is $\mathcal{T} \times 1$. Let

$$\varepsilon \sim \mathsf{N}(0, \sigma^2 I_{\mathcal{T}}), \tag{A.2}$$

so that (suppressing X from the notation on the left-hand side)

$$p(Y|\xi, \sigma^2) = \mathsf{N}(Y|X\xi, \sigma^2 I_{\mathcal{T}}). \tag{A.3}$$

Assume $p(\xi, \sigma^2) = p(\xi)\, p(\sigma^2)$ and let the prior for σ^2 be given by

$$p(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{(a_0/2)+1} \exp\left(-\frac{(b_0/2)}{\sigma^2}\right), \tag{A.4}$$

where $a_0 \geq 0$ and $b_0 \geq 0$. If $a_0 > 0$ and $b_0 > 0$ then

$$p(\sigma^2) = \mathsf{Inv\text{-}Gamma}(\sigma^2|a_0/2, b_0/2). \tag{A.5}$$

Otherwise the prior is improper. If $a_0 = b_0 = 0$, then $p(\sigma^2) \propto 1/\sigma^2$.

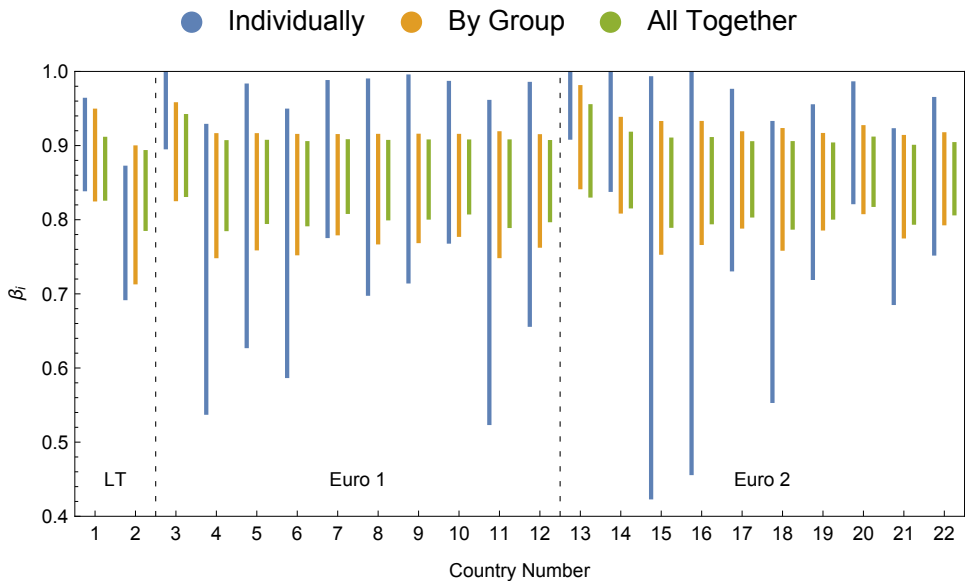

FIGURE 7. 95% highest posterior density intervals for specific cases: individually (from truncated Student t distributions), by group, and all together.

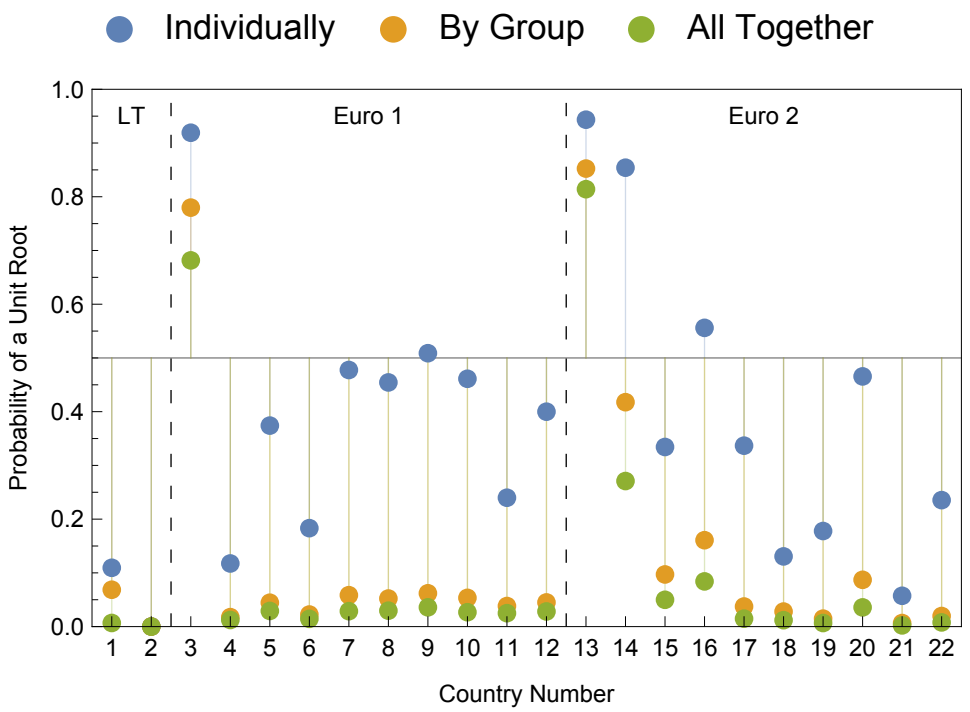

FIGURE 8. Posterior probabilities of a unit root for each specific case: individually (from truncated Student t distributions), by group, and all together.

TABLE 6. Unit root probabilites for specific cases (rounded to two decimals).

			Individually	By Group	All Together
LT	1	US	0.11	0.07	0.01
	2	FR	0.00	0.00	0.00
Euro 1	3	AT	0.92	0.78	0.68
	4	BE	0.12	0.02	0.01
	5	FI	0.37	0.04	0.03
	6	FR	0.18	0.02	0.01
	7	GR	0.48	0.06	0.03
	8	IE	0.45	0.05	0.03
	9	IT	0.51	0.06	0.04
	10	NL	0.46	0.05	0.03
	11	PT	0.24	0.04	0.02
	12	ES	0.40	0.04	0.03
Euro 2	13	AT	0.94	0.85	0.58
	14	BE	0.85	0.42	0.27
	15	FI	0.33	0.10	0.05
	16	FR	0.56	0.16	0.08
	17	GR	0.34	0.04	0.01
	18	IE	0.13	0.03	0.01
	19	IT	0.18	0.01	0.01
	20	NL	0.47	0.09	0.04
	21	PT	0.06	0.01	0.00
	22	ES	0.24	0.02	0.01

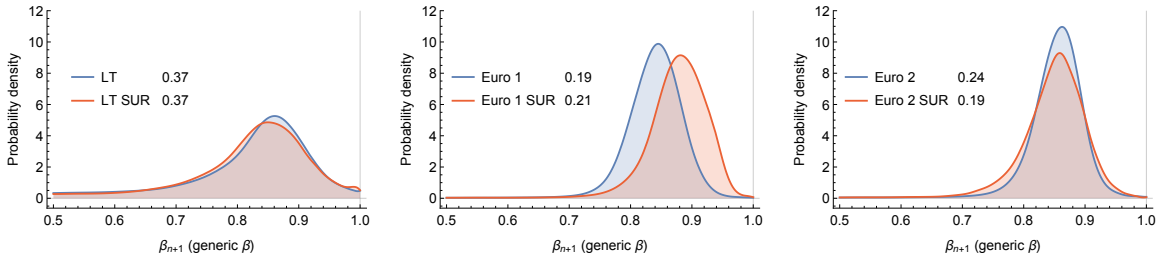

FIGURE 9. Generic posterior distributions by group using SUR likelihood compared with generic distributions using independent likelihoods. Posterior probability of a unit root is indicated in the legend.

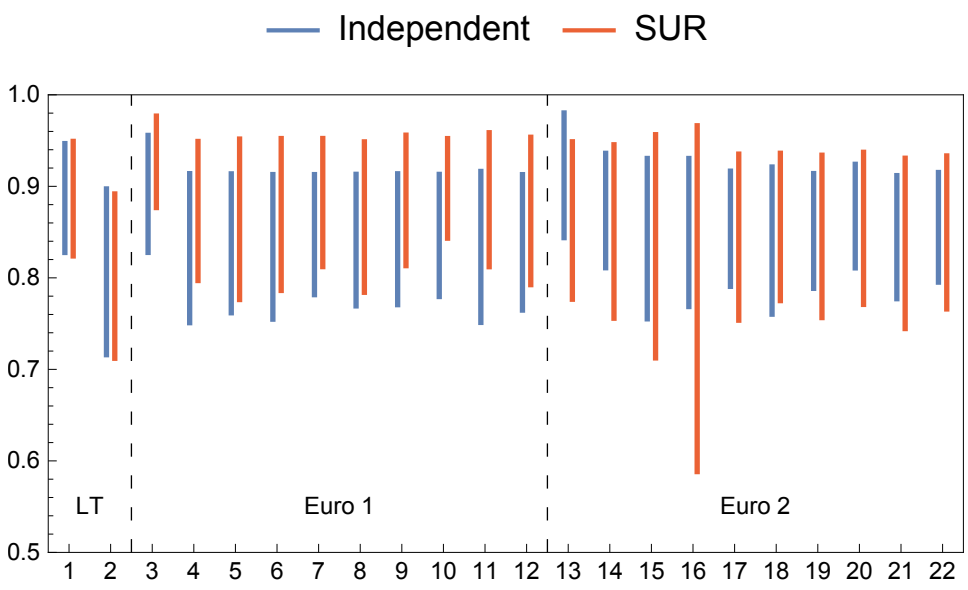

FIGURE 10. Specific posterior distributions by group using SUR likelihood compared with generic distributions using independent likelihoods.

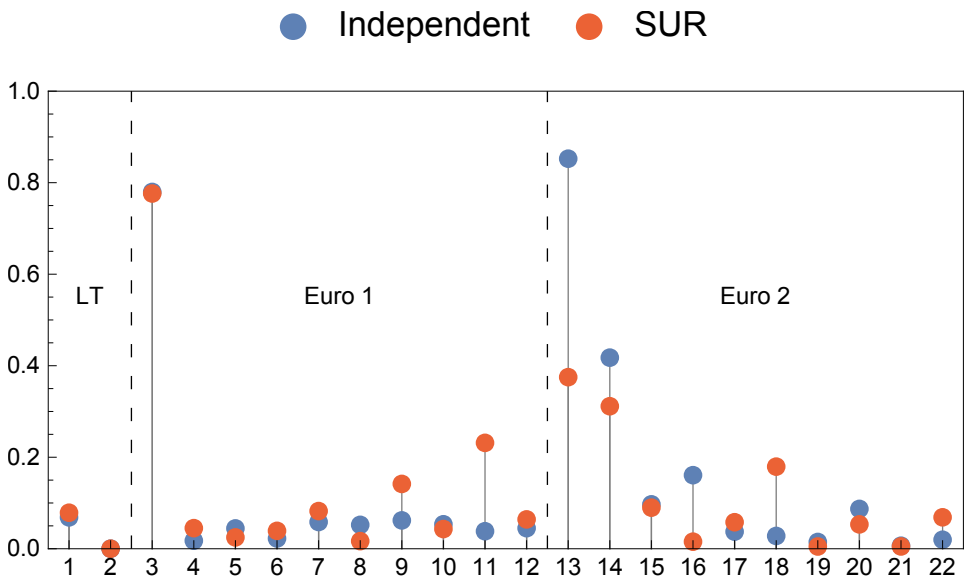

FIGURE 11. Specific posterior unit-root probabilities by group using SUR likelihood compared with generic distributions using independent likelihoods.

The marginal likelihood for ξ is given by

$$p(Y|\xi) = \int_0^\infty p(Y|\xi,\sigma^2)\,p(\sigma^2)\,d\sigma^2 \propto \mathsf{Student}(\xi|\widehat{\xi},\widehat{\Sigma},\nu), \tag{A.6}$$

where

$$\widehat{\xi} = (X^\top X)^{-1}X^\top Y \tag{A.7}$$

$$\widehat{\Sigma} = \widehat{\sigma}^2\,(X^\top X)^{-1} \tag{A.8}$$

$$\nu = a_0 + \mathcal{T} - \mathcal{K}, \tag{A.9}$$

and where

$$\widehat{\sigma}^2 = \frac{b_0 + (Y - X\,\widehat{\xi})^\top (Y - X\,\widehat{\xi})}{\nu}. \tag{A.10}$$

Assume $p(\xi) = p(\xi^{-j})\,p(\xi_j)$ and let $p(\xi^{-j}) \propto 1$. The marginal likelihood for ξ_j follows immediately:

$$p(Y|\xi_j) = \mathsf{Student}(\xi_j|\widehat{\xi}_j,\widehat{\Sigma}_{jj},\nu). \tag{A.11}$$

In passing, note the conditional posterior distribution for σ^2 is

$$p(\sigma^2|Y,\xi) = \mathsf{Inv\text{-}Gamma}(\sigma^2|a/2,b/2), \tag{A.12}$$

where

$$a = a_0 + \mathcal{T} \tag{A.13}$$

$$b = b_0 + (Y - X\,\xi)^\top (Y - X\,\xi). \tag{A.14}$$

Let us apply this framework to (2.5). In this case $\xi = (\alpha_i,\beta_i)^\top$ so that $\mathcal{K} = 2$. In addition, $\mathcal{T} = T_i - 1$. Therefore,

$$m_i = \widehat{\xi}_2 \qquad \text{and} \qquad s_i^2 = \widehat{\Sigma}_{22}. \tag{A.15}$$

In addition, if $a_0 = 0$, then $\nu_i = T_i - 3$.

Appendix B. Technical remarks regarding the prior in Section 3

[This section is incomplete.]

Here we discuss a number of technical details regarding the prior presented in Section 3.

Data frequency. Our data are of annual frequency and the priors we discuss apply to the AR coefficients of such data. For some other frequency of observations (such as monthly), the AR coefficient would require that the prior be transformed.

For example, let $\rho_i = \beta_i^{1/12}$. Then

$$p(\rho_i) = 12\,\rho_i^{11}\,p(\beta_i)|_{\beta_i=\rho_i^{12}}. \tag{B.1}$$

If, in addition, $p(\beta_i) = \mathsf{Beta}(\beta_i|1,1)$, then

$$p(\rho_i) = \mathsf{Beta}(\rho_i|12,1). \tag{B.2}$$

See Appendix H for how this prior (and others) can be incorporated into our framework.

Half-lives. The half-life in years is related to the AR coefficient by $\lambda_i = \log(1/2)/\log(\beta_i)$. Note that

$$p(\lambda_i) = \log(2)\, 2^{-1/\lambda_i}\, \lambda_i^{-2}\, p(\beta_i)|_{\beta_i = 2^{-1/\lambda_i}}. \tag{B.3}$$

If $p(\beta_i) = \mathsf{Beta}(\beta_i|1,1)$, then $p(\lambda_i) = 2^{-1/\lambda_i}\, \log(2)/\lambda_i^2$.

Also, if $\rho_i = \beta_i^{1/12}$, then $\lambda_i = \big(\log(1/2)/\log(\rho_i)\big)/12$.

Previous Bayesian approaches to unit roots and PPP. There have some previous uses of Bayesian inference. For example, see DeJong and Whiteman (1991) and Schotman and van Dijk (1991).

Prior for intercept: Schotman and van Dijk. The "meaning" of α_i depends on whether $\beta_i < 1$ or $\beta_i = 1$. Note that if $\beta_i < 1$ then $\alpha_i = (1-\beta_i)\,\mu_i$ where $\mu_i = E[y_{it}]$. On the other hand, if $\beta_i = 1$ then $\alpha_i = E[\Delta y_{it}]$. Here is what SvD say (using our notation) [p. 205]:

> If $\beta_i = 1$, the interpretation of the constant term changes. For $\beta_i < 1$, the constant term conveys information about the mean of Y_i; for $\beta_i = 1$, it determines the drift of Y_i. To exclude a random walk with drift under the null, when a trend is not present under the alternative, the parameter α_i should shrink to zero if $\beta_i \to 1$. Such a restriction must be incorporated in the prior.

Schotman and van Dijk (1991, SvD) consider (among other things) the model characterized by (2.3) and (2.4). They wish to rule out a priori the possibility of $\alpha_i \neq 0 \ \wedge \ \beta_i = 1$.

SvD are interested in testing the joint restriction

$$\alpha_i = 0 \ \wedge \ \beta_i = 1. \tag{B.4}$$

SvD choose to model in terms of μ_i rather than α_i. The likelihood can be expressed in terms of μ_i:

$$p(Y_i|\mu_i, \beta_i, \sigma_i) = p(Y_i|\alpha_i, \beta_i, \sigma_i)|_{\alpha_i = (1-\beta_i)\,\mu_i}. \tag{B.5}$$

SvD observe that $\lim_{\beta_i \to 1} (1-\beta_i)\,\mu_i = 0$. In their view, the prior for μ_i should enforce the implication $\beta_i = 1 \implies \alpha_i = 0$, in which case a test of $\beta_i = 1$ amounts to a test of the joint restriction (B.4). In addition, SvD argue the prior uncertainty for μ_i should increase to infinity as $\beta_i \to 1$. They propose a prior that embodies both of these features.

Here is their prior for μ_i:

$$p(\mu_i|\beta_i, \sigma_i) = \mathsf{N}\left(\mu_i \,\middle|\, y_{i1}, \frac{\sigma_i^2}{1-\beta_i^2}\right). \tag{B.6}$$

Changing variables, we obtain the equivalent prior for α:

$$p(\alpha_i|\beta_i, \sigma_i) = \mathsf{N}\left(\alpha_i \,\middle|\, (1-\beta_i)\, y_{i1}, \sigma_i^2 \left(\frac{1-\beta_i}{1+\beta_i}\right)\right). \tag{B.7}$$

The dependence in prior for μ_i on β_i is a two-way street in the following sense: Information about μ_i will affect the marginal likelihood for β_i. Let $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T_i} y_{it}$ be the sample mean. Referring to (B.6), we see the extent to which $y_{i1} \neq \overline{y}_i$, there will be pressure on β_i to increase toward one in order to make the variance for the prior for μ_i larger to accommodate the divergence.

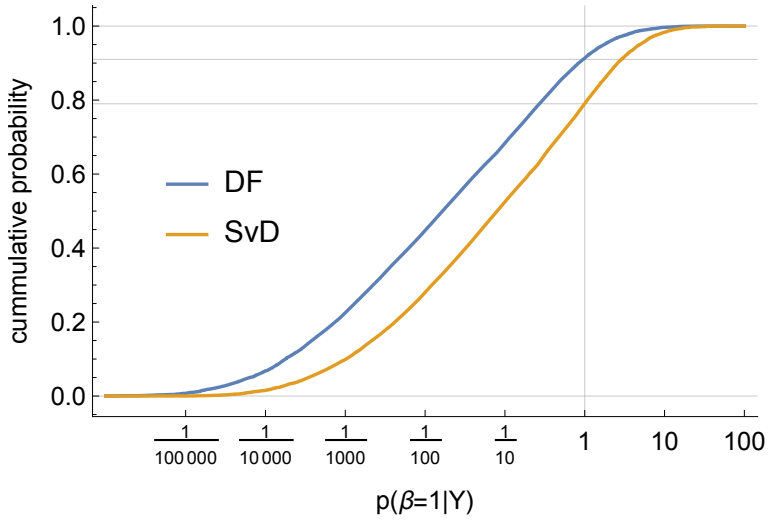

FIGURE 12. Empirical CDFs for $\log_{10}(z_j^{(r)})$.

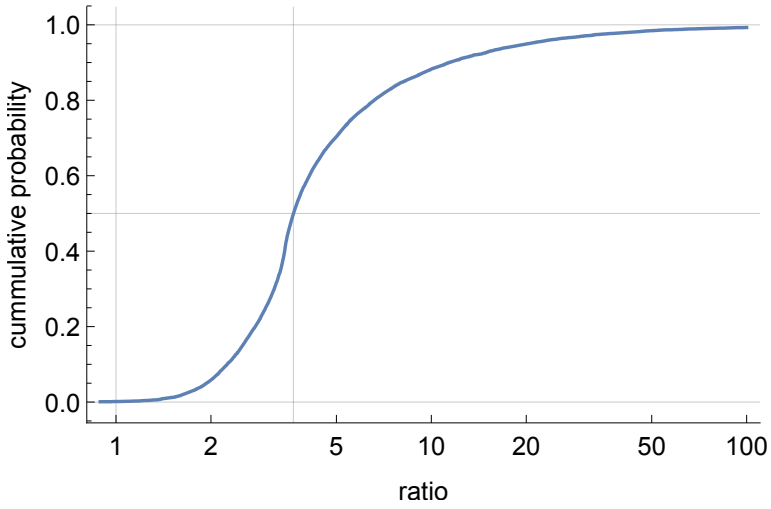

FIGURE 13. Empirical CDF for $\log_{10}(z_2^{(r)}/z_1^{(r)})$.

Simulation. In order to illustrate the effect of the SvD prior relative to our prior, we ran a simulation. (We drop the dependence on i here.) We generated the data $Y^{(r)}$ for each simulation r as follows. We set $\mu = 0$, $\sigma = 1$, and $T = 31$, and we let $\beta \sim \mathsf{Uniform}(0, 1)$. We adopted the prior $p(\beta, \sigma) \propto 1/\sigma$ for $(\beta, \sigma) \in [0, 1] \times [0, \infty)$. We ran 10^4 simulations of $Y^{(r)}$ and computed the marginal posterior distribution $p(\beta|Y^{(r)}, M_j)$ for each simulation using our prior $p(\alpha) \propto 1$ (for $j = 1$) and (B.7) (for $j = 2$). We summarize the effects of the two priors for α via $z_j^{(r)} := p(\beta = 1|Y^{(r)}, M_j)$. See Figures 12 and 13. Figure 14 shows how often the Bayes factor prefers M_2 to M_1 as a function of β.

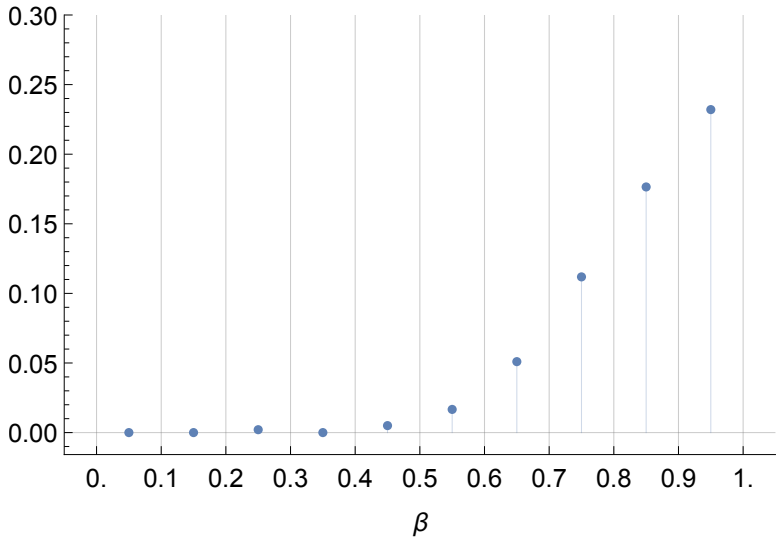

FIGURE 14. Fraction of times Bayes factor favors M_2 over M_1 as a function β.

APPENDIX C. A SMALL AMOUNT OF MEASURE THEORY

This appendix contains a brief discussion of the representation of densities involving mutually singular measures. The material here is taken from Gottardo and Raftery (2008), which see for more detail.

Let the dominating measure be $\nu = \delta_1 + \lambda$, where δ_1 is the Dirac mass at one and λ is one-dimensional Lebesgue measure. The measures δ_1 and λ are mutually singular. Partition the closed unit interval into $[0,1] = [0,1) \cup \{1\}$. Note $\nu([0,1]) = \delta_1([0,1]) + \lambda([0,1]) = 2$ since $\delta_1([0,1]) = \delta_1([0,1)) + \delta_1(\{1\}) = 0 + 1 = 1$ and $\lambda([0,1]) = \lambda([0,1)) + \lambda(\{1\}) = 1 + 0 = 1$.

Consider the measure $\mu = w\,\delta_1 + (1-w)\,\lambda$ where $w \in [0,1]$. Then $\mu([0,1]) = w\,\delta_1([0,1]) + (1-w)\,\lambda([0,1]) = 1$. The density of μ with respect to ν is $\frac{d\mu}{d\nu}(x) = w\,1_{\{1\}}(x) + (1-w)\,1_{[0,1)}(x)$, where

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}. \tag{C.1}$$

More generally, we can write a density with respect to ν as

$$\frac{d\Pi}{d(\delta_1 + \lambda)}(x) = w\,\frac{d\Pi_1}{d\delta_1}(x)\,1_{\{1\}}(x) + (1-w)\,\frac{d\Pi_2}{d\lambda}(x)\,1_{[0,1)}(x), \tag{C.2}$$

or equivalently as

$$\pi(x) = w\,\pi_1(x)\,1_{\{1\}}(x) + (1-w)\,\pi_2(x)\,1_{[0,1)}(x), \tag{C.3}$$

where $\pi_1(x)$ is a density with respect to the Dirac mass at one and $\pi_2(x)$ is a density with respect to Lebesgue measure.

Thus we can express (5.1) as

$$f(\beta_i|\theta_c) = w_c\,1_{\{1\}}(\beta_i) + (1-w_c)\,\mathsf{Beta}(\beta_i|a_c,b_c)\,1_{[0,1)}(\beta_i). \tag{C.4}$$

Equation (C.4) is a density with respect to the dominating measure $\delta_1 + \lambda$ over the sets $\{1\}$ and $[0, 1)$, where δ_1 denotes a Dirac mass located at 1 and λ denotes Lebesgue measure.

Appendix D. More on the posterior for the specific case

The posterior distribution for the specific case is conveniently expressed in terms of $\theta_{1:n}$:

$$\begin{aligned} p(\beta_i|Y_{1:n}) &= \int p(\beta_{1:n}|Y_{1:n})\, d\beta_{1:n}^{-i} \\ &= \int \left(\int p(\beta_{1:n}|Y_{1:n}, \theta_{1:n})\, p(\theta_{1:n}|Y_{1:n})\, d\theta_{1:n} \right) d\beta_{1:n}^{-i} \\ &= \iint p(\beta_i|Y_{1:n}, \beta_{1:n}^{-i}, \theta_{1:n})\, p(\beta_{1:n}^{-i}, \theta_{1:n}|Y_{1:n})\, d\theta_{1:n}\, d\beta_{1:n}^{-i}, \end{aligned} \tag{D.1}$$

where[15]

$$p(\beta_i|Y_{1:n}, \beta_{1:n}^{-i}, \theta_{1:n}) = \frac{p(Y_{1:n}|\beta_i, \beta_{1:n}^{-i})\, p(\beta_i|\theta_i)}{\int p(Y_{1:n}|\beta_i, \beta_{1:n}^{-i})\, p(\beta_i|\theta_i)\, d\beta_i}. \tag{D.2}$$

If the likelihood factors then

$$p(\beta_i|Y_{1:n}, \beta_{1:n}^{-i}, \theta_{1:n}) = \frac{p(Y_i|\beta_i)\, p(\beta_i|\theta_i)}{\int p(Y_i|\beta_i)\, p(\beta_i|\theta_i)\, d\beta_i} = p(\beta_i|Y_i, \theta_i), \tag{D.3}$$

and consequently

$$p(\beta_i|Y_{1:n}) = \int p(\beta_i|Y_i, \theta_i)\, p(\theta_i|Y_{1:n})\, d\theta_i. \tag{D.4}$$

Here is an explicit expression for $p(\beta_i|Y_i, \theta_i)$. The posterior distribution for β_i conditional on θ_i is given by

$$\begin{aligned} p(\beta_i|Y_i, \theta_i) &= \frac{p(Y_i|\beta_i)\, f(\beta_i|\theta_i)}{\int p(Y_i|\beta_i)\, f(\beta_i|\theta_i)\, d\beta_i} \\ &= \begin{cases} w_i' & \beta_i = 1 \\ (1 - w_i') \left(\dfrac{p(Y_i|\beta_i)\, \mathsf{Beta}(\beta_i|a_i, b_i)}{p(Y_i|a_i, b_i)} \right) & \beta_i < 1 \end{cases}, \end{aligned} \tag{D.5}$$

where the posterior probability of a unit root is

$$w_i' = \frac{w_i\, h_i}{w_i\, h_i + (1 - w_i)\, p(Y_i|a_i, b_i)} \tag{D.6}$$

and $p(Y_i|a_i, b_i) = \int p(Y_i|\beta_i)\, \mathsf{Beta}(\beta_i|a_i, b_i)\, d\beta_i$ [as in (3.19)] and $h_i = p(Y_i|\beta_i = 1)$.

Appendix E. Details regarding the sampler

Here we provide details for the Gibbs sampler outlined in Section 6.

[15] See Appendix F for a related expression in the case of a seemingly unrelated regressions (SUR) setting.

Step 1: Drawing $\boldsymbol{\theta}$. We address how to draw θ_c given B_c, where (recall)

$$B_c = \{\beta_i \in \beta_{1:n} : z_i = c\}, \tag{E.1}$$

and where $n_c = |B_c|$ is the number of elements in cluster c. Within cluster c, let

$$S_c = \{\beta_i \in B_c : \beta_i < 1\}, \tag{E.2}$$

and $s_c = |S_c|$. Note $n_c - s_c$ is the number of unit roots in cluster c.

First, the draw of $w_c|B_c$ is straightforward because the likelihood for $w_c|B_c$ is binomial:

$$w_c|B_c \sim \mathsf{Beta}\big(\zeta\,\phi + (n_c - s_c), \zeta\,(1-\phi) + s_c\big). \tag{E.3}$$

If $n_c = s_c = 0$, then this amounts to a draw from the prior.

We now turn to the draw of $(j_c, k_c)|B_c$. If $n_c = 0$, then simply draw (j_c, k_c) from its prior. Otherwise, note that

$$p(j_c, k_c|B_c) \propto p(S_c|j_c, k_c)\, p(j_c, k_c). \tag{E.4}$$

Here is one possible Metropolis–Hastings scheme: Let the proposal be given by

$$k_c' - 1 \sim \mathsf{Poisson}(k_c) \tag{E.5}$$

$$j_c' - 1 \sim \mathsf{Binomial}(k_c' - 1, \overline{\beta}^c), \tag{E.6}$$

where $\overline{\beta}^c$ is the sample mean of those β_i in S_c. Consequently, the proposal density is

$$q\big((j_c', k_c')|(j_c, k_c)\big) = \mathsf{Binomial}(j_c' - 1|k_c' - 1, \overline{\beta}^c)\,\mathsf{Poisson}(k_c' - 1|k_c). \tag{E.7}$$

Letting (j_c, k_c) stand for $(j_c^{(r-1)}, k_c^{(r-1)})$, the acceptance rule is

$$(j_c^{(r)}, k_c^{(r)}) = \begin{cases} (j_c', k_c') & \mathcal{M} \geq u^{(r)} \\ (j_c, k_c) & \text{otherwise} \end{cases}, \tag{E.8}$$

where $u^{(r)} \sim \mathsf{Uniform}(0,1)$ and

$$\mathcal{M} = \frac{p(S_c|j_c', k_c')\, p(j_c', k_c')}{p(S_c|j_c, k_c)\, p(j_c, k_c)} \times \frac{q\big((j_c, k_c)|(j_c', k_c')\big)}{q\big((j_c', k_c')|(j_c, k_c)\big)}. \tag{E.9}$$

Step 1: Drawing $\boldsymbol{\eta}$. Note that

$$p(\eta|z_{1:n}) \propto p(z_{1:n}|\eta)\, p(\eta). \tag{E.10}$$

Recall $p(\eta)$ is given in (4.12). The likelihood for η is given by

$$p(z_{1:n}|\eta) = p(z_1|\eta) \prod_{i=1}^{n-1} p(z_{i+1}|z_{1:i}, \eta) \propto \frac{\eta^d\, \Gamma(\eta)}{\Gamma(n+\eta)}, \tag{E.11}$$

where d is the number of occupied clusters (i.e., clusters for which $n_c > 0$).[16]

Draws of η can be made using a Metropolis–Hastings scheme.[17] Let the proposal be

$$\eta' \sim \mathsf{LogLogistic}(\eta^{(r)}, h), \tag{E.12}$$

[16]The sampling distribution $p(z_{i+1}|z_{1:i}, \eta)$ can be obtained from the Chinese Restaurant Process.

[17]Alternatively, draws of η can be made using a Metropolis scheme. Make a random-walk proposal of $\lambda' \sim \mathrm{N}(\lambda^{(r)}, s^2)$, where $\lambda^{(r)} = \eta^{(r)}/(1+\eta^{(r)})$ and s^2 is a suitable scale. Then evaluate the likelihood ratio for $\eta' = \lambda'/(1-\lambda')$ relative to $\eta^{(r)}$ to determine whether or not to accept the proposal η'.

where[18]

$$\mathsf{LogLogistic}(x|m,h) = \frac{h\,x^{h-1}\,m^h}{(x^h+m^h)^2}. \tag{E.13}$$

Note that $\eta \sim \mathsf{LogLogistic}(1,1)$. The inverse-CDF method can be used to make draws from $\mathsf{LogLogistic}(m,h)$: Draw $u \sim \mathsf{Uniform}(0,1)$ and set $x = m\left(u/(1-u)\right)^{1/h}$. For determining whether or not to accept the proposal, we require the "Hastings ratio,"

$$\frac{\mathsf{LogLogistic}(\eta^{(r)}|\eta',h)}{\mathsf{LogLogistic}(\eta'|\eta^{(r)},h)} = \frac{\eta'}{\eta^{(r)}}. \tag{E.14}$$

Step 2: Drawing $\boldsymbol{\beta_{1:n}}$. The draws of $\beta_i|Y_{1:n},\theta_{1:n}$ and $\beta_j|Y_{1:n},\theta_{1:n}$ are independent. (Recall $\theta_i = \theta_{z_i}$.) In particular,

$$p(\beta_i|Y_{1:n},\theta_{1:n}) = p(\beta_i|Y_i,\theta_i) \propto p(Y_i|\beta_i)\,f(\beta_i|\theta_i). \tag{E.15}$$

The draws of β_i involve both a point mass at one and a density with respect to Lebesgue measure over the unit interval. To account for these mutually singular measures in the sampler, we adopt the framework of Gottardo and Raftery (2008). The proposal density has the following form:

$$q(\beta_i,\beta_i') = \begin{cases} \gamma_1\,1_{\{1\}}(\beta_i') + (1-\gamma_1)\,q^*(\beta_i,\beta_i')\,1_{[0,1)}(\beta_i') & \beta_i = 1 \\ \gamma_0\,1_{\{1\}}(\beta_i') + (1-\gamma_0)\,q^*(\beta_i,\beta_i')\,1_{[0,1)}(\beta_i') & \beta_i < 1 \end{cases} \tag{E.16}$$

where γ_ℓ is the probability of proposing a move to $\{1\}$ from component ℓ and $q^*(\beta_i,\beta_i')$ is the proposal when $\beta_i' < 1$. The Metropolis–Hastings sampling scheme is characterized by

$$\beta_i^{(r+1)} = \begin{cases} \beta_i' & \mathcal{M}_i^{(r)} \geq u^{(r+1)} \\ \beta_i^{(r)} & \text{otherwise} \end{cases}, \tag{E.17}$$

where $u^{(r+1)} \sim \text{Uniform}(0,1)$ and

$$\mathcal{M}_i^{(r)} = \frac{p(Y_i|\beta_i')\,f(\beta_i'|\theta_i)}{p(Y_i|\beta_i^{(r)})\,f(\beta_i^{(r)}|\theta_i)}\,\frac{q(\beta_i',\beta_i^{(r)})}{q(\beta_i^{(r)},\beta_i')}. \tag{E.18}$$

We choose

$$q(\beta_i,\beta_i') = f(\beta_i'|\theta_i), \tag{E.19}$$

in which case

$$\mathcal{M}_i^{(r)} = \frac{p(Y_i|\beta_i')}{p(Y_i|\beta_i^{(r)})}. \tag{E.20}$$

[18] If $x \sim \mathsf{LogLogistic}(m,h)$, then $\log(x) \sim \mathsf{Logistic}(\log(m), 1/h)$.

Appendix F. SUR likelihood

Thus far we have assumed that the noisy signals are independent (in the sense that the likelihood factors). But if they are not, the the actual amount of information in the joint likelihood about the unobserved coefficients may be more or less that what has been tacitly assumed and the implicit density may not properly represent the underlying signals. In this section, we allow for dependence between ε_{it} and ε_{jt}, adopting an SUR setting for the likelihood.

Refer to (2.3) for a description of the data-generating process. Within each of our three groups of data, it is reasonable to take into account any contemporaneous correlation across the series innovations. Possible dependency is modeled using the seemingly unrelated regression (SUR) model.[19] It is convenient to express (2.3) as

$$y_{it} = x_{it}^\top \xi_i + \varepsilon_{it}, \tag{F.1}$$

where $x_{it} = (1, y_{i,t-1})$ and $\xi_i = (\alpha_i, \beta_i)$. Let $y_{\cdot t} = (y_{1t}, \dots, y_{nt})^\top$, $\xi = (\xi_1^\top, \dots, \xi_n^\top)^\top$, $\varepsilon_{\cdot t} = (\varepsilon_{1t}, \dots, \varepsilon_{nt})^\top$, and

$$X_{\cdot t} = \begin{bmatrix} x_{1t}^\top & 0 & \cdots & 0 \\ 0 & x_{2t}^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{nt}^\top \end{bmatrix}. \tag{F.2}$$

Then the data-generating process can be expressed as

$$y_{\cdot t} = X_{\cdot t}\, \xi + \varepsilon_{\cdot t} \qquad \text{where } \varepsilon_{\cdot t} \overset{\text{iid}}{\sim} \mathsf{N}(0_n, \Sigma), \tag{F.3}$$

where Σ is an $n \times n$ covariance matrix with

$$\Sigma_{ij} = \begin{cases} \sigma_i^2 & i = j \\ \sigma_i\, \sigma_j\, \rho_{ij} & i \neq j \end{cases} \tag{F.4}$$

where $-1 < \rho_{ij} < 1$.

Given (F.3), the SUR-based likelihood can be expressed as

$$\begin{aligned} p(Y_{1:n} | \xi, \Sigma) &= p(\{y_{\cdot t}\}_{t=p+1}^T | \{y_{\cdot t}\}_{t=1}^p, \xi, \Sigma) \\ &= \prod_{t=p+1}^T \mathsf{N}(y_{\cdot t} | X_{\cdot t}\, \xi, \Sigma) \\ &\propto \frac{1}{|\Sigma|^{(T-p)/2}} \exp\left(-\frac{1}{2} \sum_{t=p+1}^T (y_{\cdot t} - X_{\cdot t}\, \xi)^\top \Sigma^{-1} (y_{\cdot t} - X_{\cdot t}\, \xi) \right), \end{aligned} \tag{F.5}$$

where p is the order of the AR process; in our case, $p = 1$. The dependence in the data-generating process is captured by the off-diagonal elements in Σ.[20] The Bayesian SUR model is completed by specifying prior distributions for ξ and Σ. Given $\xi \sim \mathsf{N}(\xi_0, \Xi_0)$ and

[19] We adopt the approach in Greenberg (2013, pp. 169–172), which see for omitted details and references.

[20] Specifying Σ to be diagonal (i.e., $\rho_{ij} = 0$ for all i and j) delivers independent likelihoods.

$\Sigma^{-1} \sim \mathsf{Wishart}(\nu_0, R_0)$, the posterior distribution of (ξ, Σ) is characterized the a pair of conditional distributions:

$$\xi|y, \Sigma \sim \mathsf{N}(\overline{\xi}, \Xi_1) \qquad \text{and} \qquad \Sigma^{-1}|y, \xi \sim \mathsf{Wishart}(\nu_1, R_1), \tag{F.6}$$

where

$$\Xi_1 = \left(\Xi_0^{-1} + \sum_t X_{\cdot t}^\top \Sigma^{-1} X_{\cdot t}\right)^{-1} \tag{F.7a}$$

$$\overline{\xi} = \Xi_1 \left(\Xi_0^{-1}\xi_0 + \sum_t X_{\cdot t}^\top \Sigma^{-1} y_{\cdot t}\right) \tag{F.7b}$$

$$\nu_1 = \nu_0 + (T - p) \tag{F.7c}$$

$$R_1 = \left(R_0^{-1} + \sum_t (y_{\cdot t} - X_{\cdot t}\xi)\,(y_{\cdot t} - X_{\cdot t}\xi)^\top\right)^{-1}. \tag{F.7d}$$

We adopt an uninformative (Jeffreys) prior ($\Xi_0^{-1} = 0$, $R_0^{-1} = 0$, and $\nu_0 = 0$).[21,22,23]

Sampler. We modify the sampler presented in Section 6 and Appendix E. Only Step 2 is changed. Instead of having analytically integrated out $\alpha_{1:n}$ and Σ, we integrate them out numerically via the sampler. The draws of Σ^{-1} are made according to (F.6). For $\alpha_{1:n}$ we have

$$\alpha_{1:n}|Y_{1:n}, \beta_{1:n}, \Sigma \sim \mathsf{N}(\overline{\xi}^\alpha, \Xi_1^\alpha), \tag{F.8}$$

where $(\overline{\xi}^\alpha, \Xi_1^\alpha)$ are the appropriate conditioning parameters computed from $(\overline{\xi}, \Xi_1)$. Finally, we make draws of β_i for $i = 1, \ldots, n$ as described in Appendix E with the exception that (E.20) is replaced with

$$\mathcal{M}_i^{(r)} = \frac{p(Y_{1:n}|\alpha_{1:n}, \beta_i', \beta_{1:n}^{-i}, \Sigma)}{p(Y_{1:n}|\alpha_{1:n}, \beta_i^{(r)}, \beta_{1:n}^{-i}, \Sigma)}, \tag{F.9}$$

where the likelihoods are computed using (F.5) and $(\alpha_{1:n}, \beta_{1:n}^{-i}, \Sigma)$ have the appropriate values.

Rao-Blackwellizing specific cases. The conditional likelihood for β_i can be expressed as

$$p(Y_{1:n}|\alpha_{1:n}, \beta_i, \beta_{1:n}^{-i}, \Sigma) \propto \mathsf{N}(\beta_i|m_i, v_i) \tag{F.10}$$

[21]For $n = 1$, this prior reduces to $p(\sigma_1^2) \propto \sigma_1^{-2}$ which is equivalent to $p(\sigma_1) \propto \sigma_1^{-1}$, the prior adopted in the independent likelihoods case.

[22]In order to compute the Bayes factor for the SUR-based model with the independent likelihoods model we would need an informed prior for (the off-diagonal elements of) Σ. Thus we are not able to make inferences regarding which model is more likely.

[23]With $\Xi_0^{-1} = 0$, the value for ξ_0 is irrelevant.

for some m_i and v_i that depend on $(Y_{1:n}, \alpha_{1:n}, \beta_{1:n}^{-i}, \Sigma)$. Let $\theta_i = \theta_{z_i}$ and define

$$\begin{aligned} Z_i &= \int_0^1 \mathsf{N}(\beta_i|m_i, v_i)\, f(\beta_i|\theta_i)\, d\beta_i \\ &= w_i\, \mathsf{N}(\beta_i = 1|m_i, v_i) + (1 - w_i) \int_0^1 \mathsf{N}(\beta_i|m_i, v_i)\, \mathsf{Beta}(\beta_i|a_i, b_i)\, d\beta_i. \end{aligned} \tag{F.11}$$

Then the conditional posterior distribution for β_i is

$$\begin{aligned} p(\beta_i|Y_{1:n}, \alpha_{1:n}, \beta_{1:n}^{-i}, \Sigma, \theta_i) &= \mathsf{N}(\beta_i|m_i, v_i)\, f(\beta_i|\theta_i)/Z_i \\ &= \begin{cases} w_i\, \mathsf{N}(\beta_i = 1|m_i, v_i)/Z_i & \beta_i = 1 \\ (1 - w_i)\, \mathsf{N}(\beta_i|m_i, v_i)\, \mathsf{Beta}(\beta_i|a_i, b_i)/Z_i & \beta_i < 1 \end{cases}. \end{aligned} \tag{F.12}$$

The marginal posterior for β_i is obtained via integration involving $p(\alpha_{1:n}, \beta_{1:n}^{-i}, \Sigma, \theta_i|Y_{1:n})$, which can be approximated using draws from the posterior:

$$\frac{1}{R} \sum_{r=1}^{R} p(\beta_i|Y_{1:n}, \alpha_{1:n}^{(r)}, (\beta_{1:n}^{-i})^{(r)}, \Sigma^{(r)}, \theta_i^{(r)}). \tag{F.13}$$

Consequently, the probability of a unit root for the specific case is approximated by

$$\widehat{\pi}_i = \frac{1}{R} \sum_{r=1}^{R} \frac{w_i^{(r)}\, \mathsf{N}(\beta_i = 1|m_i^{(r)}, v_i^{(r)})}{Z_i^{(r)}} \tag{F.14}$$

and the corresponding density over the unit interval is approximated by

$$\frac{1}{(1 - \widehat{\pi}_i)\, R} \sum_{r=1}^{R} \frac{(1 - w_i^{(r)})\, \mathsf{N}(\beta_i|m_i^{(r)}, v_i^{(r)})\, \mathsf{Beta}(\beta_i|a_i^{(r)}, b_i^{(r)})}{Z_i^{(r)}}, \tag{F.15}$$

where $Z_i^{(r)}$ is given by Z_i in (F.11) with the parameters $(a_i, b_i, w_i, m_i, v_i)$ replaced by $(a_i^{(r)}, b_i^{(r)}, w_i^{(r)}, m_i^{(r)}, v_i^{(r)})$.

Appendix G. Residual factor structure

In this section we model a residual factor structure. We follow Jones and Shanken (2005) who rely on Geweke and Zhou (1996).

Let $G = (g_2, \ldots, g_T)$ be a common factor. We begin by conditioning on G. Let

$$y_{it} = \alpha_i + \beta_i\, y_{i,t-1} + \delta_i\, g_t + \varepsilon_{it}, \tag{G.1}$$

where

$$\varepsilon_{it} \sim \mathsf{N}(0, \sigma_i^2). \tag{G.2}$$

Let $\phi_i = (\alpha_i, \delta_i, \sigma_i^2)$. Then

$$p(Y_i|G, \beta_i, \phi_i) = \prod_{t=2}^{T} p(y_{it}|y_{i,t-1}, g_t, \beta_i, \phi_i), \tag{G.3}$$

where

$$p(y_{it}|y_{i,t-1}, g_t, \beta_i, \phi_i) = \mathsf{N}(y_{it}|\alpha_i + \beta_i\, y_{i,t-1} + \delta_i\, g_t, \sigma_i^2). \tag{G.4}$$

Integrating out ϕ_i, where $p(\phi_i) \propto 1/\sigma_i^2$, the marginal likelihood for β_i conditional on G is given by

$$p(Y_i|G, \beta_i) = \mathsf{Student}(\beta_i|m_{iG}, s_{iG}^2, \nu_{iG}). \tag{G.5}$$

The parameters $(m_{iG}, s_{iG}^2, \nu_{iG})$ depend only on Y_i and G; they can be computed following the steps discussed in Appendix A.

We now treat G as a latent residual factor. Let $g_t \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ and let $\xi_t = (\xi_{1t}, \ldots, \xi_{nt})$, where

$$\xi_{it} = \delta_i \, g_t + \varepsilon_{it}. \tag{G.6}$$

Then

$$\xi_t \sim \mathsf{N}(0, \delta \, \delta^\top + \Sigma), \tag{G.7}$$

where $\delta = (\delta_1, \ldots, \delta_n)$ and Σ is an $n \times n$ diagonal matrix where $\Sigma_{ii} = \sigma_i^2$.

The Gibbs sampler. Here we describe the structure of the Gibbs sampler when a latent factor is involved. The joint posterior distribution for the unknowns is

$$p(\beta_{1:n}, \psi, \eta, z_{1:n}, \phi_{1:n}, G|Y_{1:n}). \tag{G.8}$$

This joint distribution can be characterized by the following full conditional distributions:

$$p(\psi, \eta, z_{1:n}|Y_{1:n}, G, \beta_{1:n}, \phi_{1:n}) = p(\psi, \eta, z_{1:n}|\beta_{1:n}) \tag{G.9a}$$

$$p(\beta_{1:n}, \phi_{1:n}|Y_{1:n}, G, \psi, \eta, z_{1:n}) = \prod_{i=1}^{n} p(\beta_i, \phi_i|Y_i, G, \theta_i) \tag{G.9b}$$

$$p(G|Y_{1:n}, \beta_{1:n}, \phi_{1:n}, \psi, \eta, z_{1:n}) = \prod_{t=2}^{T} p(g_t|\mathcal{Y}_t, \mathcal{Y}_{t-1}, \beta_{1:n}, \phi_{1:n}), \tag{G.9c}$$

where $\mathcal{Y}_t = (y_{1t}, \ldots, y_{nt})$. The right-hand side of (G.9a) is unchanged from the case with no latent factor and so Step 1 is unchanged. A new version of Step 2 is embodied in the right-hand sides of (G.9b–G.9c).

Regarding (G.9b), we can factor the distribution for (β_i, ϕ_i) into conditional and marginal distributions as follows:

$$p(\beta_i, \phi_i|Y_i, G, \theta_i) = p(\phi_i|Y_i, G, \beta_i) \, p(\beta_i|Y_i, G, \theta_i). \tag{G.10}$$

The marginal posterior distribution for β_i conditional on G can be expressed as

$$p(\beta_i|Y_i, G, \theta_i) \propto p(Y_i|\beta_i, G) \, f(\beta_i|\theta_i), \tag{G.11}$$

where the marginal likelihood for β_i conditional on G is given by (G.5).[24] The draws can be made according to (E.17).

[24]By contrast, consider the conditional likelihood for β_i, which can be expressed as

$$p(Y_i|\alpha_i, \beta_i, \delta_i, \sigma_i^2, G) \propto \mathsf{N}(\beta_i|m_i, v_i),$$

where

$$m_i = \frac{\sum_{t=2}^{T} y_{i,t-1} \, (y_{it} - \alpha_i - \delta_i \, g_t)}{\sum_{t=2}^{T} y_{i,t-1}^2} \quad \text{and} \quad v_i = \frac{\sigma_i^2}{\sum_{t=2}^{T} y_{i,t-1}^2}.$$

The variance v_i can be extremely small, making a sampler based on this likelihood inefficient.

Next, we can characterize $p(\phi_i|Y_i, G, \beta_i)$ in terms of the following full conditional distributions:

$$p(\alpha_i|Y_i, G, \beta_i, \delta_i, \sigma_i^2), \quad p(\delta_i|Y_i, G, \alpha_i, \beta_i, \sigma_i^2), \quad \text{and} \quad p(\sigma_i^2|Y_i, G, \alpha_i, \beta_i, \delta_i). \tag{G.12}$$

Note the $p(\sigma_i^2|Y_i, G, \alpha_i, \beta_i, \delta_i)$ is the conditional distribution for σ_i^2 rather than the marginal distribution as typically would be the case. Finally, it is straightforward to draw from (G.9c). The details are presented below.

In summary, Step 1 is unchanged. Step 2 begins the same as before, except that the sufficient statistics for $\beta_{1:n}$ are computed conditional on G. Then $\phi_{1:n}$ is drawn conditional on $(\beta_{1:n}, G)$ and finally G is drawn conditional on $(\beta_{1:n}, \phi_{1:n})$.

Rao–Blackwellization for specific cases is based on (D.5) using $p(Y_i|\beta_i, G)$ in place of $p(Y_i|\beta_i)$.

Details. In order to preserve the marginal–conditional factorization of the distribution for (β_i, ϕ_i) the draw of β_i must precede the draw of ϕ_i within a given sweep and the draw of G must not intervene.

First, for $i = 1, \ldots, n$,

$$\alpha_i \,|\, Y_i, \beta_i, \sigma_i, \delta_i, G \sim \mathsf{N}(\widetilde{m}_i, \widetilde{s}_i^2), \tag{G.13}$$

where

$$\widetilde{m}_i = \frac{\sum_{t=2}^T y_{it} - \beta_i\, y_{i,t-1} - \delta_i\, g_t}{T-1} \qquad \text{and} \qquad \widetilde{s}_i^2 = \frac{\sigma_i^2}{T-1}. \tag{G.14}$$

Second,

$$\delta_i \,|\, Y_i, \alpha_i, \beta_i, \sigma_i, G \sim \mathsf{N}(\widehat{m}_i, \widehat{s}_i^2) \tag{G.15}$$

where

$$\widehat{m}_i = \frac{\sum_{t=2}^T g_t\,(y_{i,t} - \alpha_i - \beta_i\, y_{i,t-1})}{\sum_{t=2}^T g_t^2} \quad \text{and} \quad \widehat{s}_i^2 = \frac{\sigma_i^2}{\sum_{t=2}^T g_t^2} \tag{G.16}$$

Third,

$$\sigma_i^2 \,|\, Y_i, \alpha_i, \beta_i, \delta_i, G \sim \mathsf{Inv\text{-}Gamma}(a_i/2, b_i/2), \tag{G.17}$$

where

$$a_i = a_0 + T_i - 1 \qquad \text{and} \qquad b_i = b_0 + \sum_{t=2}^{T_i} \big(y_{it} - \alpha_i - \beta_i\, y_{i,t-1} - \delta_i\, g_t\big)^2. \tag{G.18}$$

Finally, for $t = 2, \ldots, T$,

$$g_t \,|\, \mathcal{Y}_t, \mathcal{Y}_{t-1}, \alpha, \beta, \delta, \sigma \sim \mathsf{N}(M, S^2), \tag{G.19}$$

where[25]

$$M = \delta^\top(\delta\,\delta^\top + \Sigma)^{-1}(\mathcal{Y}_t - \alpha - \beta\,\mathcal{Y}_{t-1}) \qquad \text{and} \qquad S^2 = 1 - \delta^\top(\delta\,\delta^\top + \Sigma)^{-1}\delta. \tag{G.20}$$

[25]Note $(\delta\,\delta^\top + \Sigma)^{-1} = \Sigma^{-1} - (\Sigma^{-1}\delta\,\delta^\top\Sigma^{-1})/(1 + \delta^\top\Sigma^{-1}\delta)$.

No learning. Let examine the case with the latent factor but where there is no learning. We can use this case as a baseline for comparison. The joint posterior distribution for the unknowns is

$$p(\beta_{1:n}, \phi_{1:n}, G|Y_{1:n}), \tag{G.21}$$

which can be characterized by the following full conditional distributions:

$$p(\beta_{1:n}, \phi_{1:n}|Y_{1:n}, G) = \prod_{i=1}^{n} p(\beta_i, \phi_i|Y_i, G) \tag{G.22a}$$

$$p(G|Y_{1:n}, \beta_{1:n}, \phi_{1:n}) = \prod_{t=2}^{T} p(g_t|\mathcal{Y}_t, \mathcal{Y}_{t-1}, \beta_{1:n}, \phi_{1:n}). \tag{G.22b}$$

Draws for ϕ_i and G are made as before. Draws of β_i can be made directly from the conditional posterior distribution:

$$p(\beta_i|Y_i, G) = \begin{cases} \frac{h_i^G}{g_i^G + h_i^G} & \beta_i = 1 \\ \frac{g_i^G}{g_i^G + h_i^G} \left(\frac{p(Y_i|\beta_i, G)}{g_i^G} \right) & \beta_i < 1 \end{cases}, \tag{G.23}$$

where $p(Y_i|\beta_i, G)$ is given in (G.5) and

$$g_i^G = \int_0^1 p(Y_i|\beta_i, G)\, d\beta_i \qquad \text{and} \qquad h_i^G = p(Y_i|\beta_i = 1, G). \tag{G.24}$$

Rao–Blackwellization can also be based on (G.23):

$$\widehat{\pi}_i \approx \frac{1}{R} \sum_{r=1}^{R} \frac{h_i^{G^{(r)}}}{g_i^{G^{(r)}} + h_i^{G^{(r)}}} \tag{G.25}$$

and the density over the unit interval is approximated by

$$\frac{1}{(1 - \widehat{\pi}_i)\, R} \sum_{r=1}^{R} \frac{p(Y_i|\beta_i, G^{(r)})}{g_i^{G^{(r)}} + h_i^{G^{(r)}}}. \tag{G.26}$$

Appendix H. More general prior

Here we sketch how to engineer a more general marginal prior over the unit interval (or any interval). We generalize (5.1) as follows:

$$\widetilde{f}(\beta_i|a_i, b_i, w_i) = \begin{cases} w_i & \beta_i = 1 \\ (1 - w_i)\, \mathsf{Beta}(\Upsilon(\beta_i)|a_i, b_i)\, \upsilon(\beta_i) & \beta_i < 1 \end{cases}, \tag{H.1}$$

where $\upsilon(x) \geq 0$, $\int_{-\infty}^{\infty} \upsilon(x)\, dx = 1$, and $\Upsilon(x) := \int_{-\infty}^{x} \upsilon(t)\, dt$. Note

$$\int_{-\infty}^{\infty} \mathsf{Beta}(\Upsilon(x)|a, b)\, \upsilon(x)\, dx = 1. \tag{H.2}$$

In addition,

$$\frac{1}{k} \sum_{j=1}^{k} \mathsf{Beta}(\Upsilon(x)|a, b)\, \upsilon(x) = \upsilon(x), \tag{H.3}$$

and therefore the prior predictive distribution over the unit interval is

$$E[\mathsf{Beta}(\Upsilon(\beta_i)|a_c, b_c)\, \upsilon(\beta_i)] = \upsilon(\beta_i). \tag{H.4}$$

For example, $\upsilon(x) = \mathsf{Beta}(x|\widetilde{a}, \widetilde{b})$. The prior in the body of the paper is a special case with $\upsilon(x) = \mathsf{Uniform}(x|0, 1)$ and $\Upsilon(x) = x$.

Sampler. In order to accommodate the more general marginal prior displayed in (H.1), a modification to the sampler is required. It is convenient to work directly with the transformed coefficients, where $\widehat{\beta}_i = \Upsilon(\beta_i)$. This allows for the use of the sampler described above (with $\widehat{\beta}_{1:n}$ replacing $\beta_{1:n}$ in the formulas) with a single modification to account for the likelihood for $\widehat{\beta}_i$. In particular, replace (E.20) with

$$\widehat{\mathcal{M}}_i^{(r)} = \frac{\widehat{p}(Y_i|\widehat{\beta}_i')}{\widehat{p}(Y_i|\widehat{\beta}_i^{(r)})}, \tag{H.5}$$

where

$$\widehat{p}(Y_i|\widehat{\beta}_i) := p(Y_i|\beta_i)\big|_{\beta_i = \Upsilon^{-1}(\widehat{\beta}_i)}. \tag{H.6}$$

The draws $\{\widehat{\beta}_{1:n}^{(r)}\}_{r=1}^R$ can be transformed via $\beta_i^{(r)} = \Upsilon^{-1}(\widehat{\beta}_i^{(r)})$ if so desired.

Given the draws of $\{\psi_n^{(r)}\}_{r=1}^R$, we turn to computing the Rao–Blackwellized approximations to the generic and specific distributions. Starting with the generic case, define

$$\varphi_{n+1}(x) := p(\widehat{\beta}_{n+1}|Y_{1:n})|_{\widehat{\beta}_{n+1}=x} \qquad \text{for } x \neq 1. \tag{H.7}$$

Then for $\beta_{n+1} \neq 1$

$$p(\beta_{n+1}|Y_{1:n}) = \varphi_{n+1}\big(\Upsilon(\beta_{n+1})\big)\, \upsilon(\beta_{n+1}). \tag{H.8}$$

Turning to the specific cases, define

$$\varphi_i(x) := p(\widehat{\beta}_i|Y_{1:n})|_{\widehat{\beta}_i=x} \qquad \text{for } x \neq 1, \tag{H.9}$$

where $p(\widehat{\beta}_i|Y_{1:n})$ is calculated using $\widehat{p}(Y_i|\widehat{\beta}_i)$ in place of $p(Y_i|\widehat{\beta}_i)$. In particular,

$$\widehat{p}(Y_i|a_i, b_i) = \int \widehat{p}\big(Y_i|\widehat{\beta}_i\big)\, f(\widehat{\beta}_i|a_i, b_i)\, d\widehat{\beta}_i. \tag{H.10}$$

Then for $\beta_i \neq 1$

$$p(\beta_i|Y_{1:n}) = \varphi_i\big(\Upsilon(\beta_i)\big)\, \upsilon(\beta_i). \tag{H.11}$$

References

Bauwens, L., M. Lubrano, and J.-F. Richard (1999). *Bayesian inference in dynamic econometric models*. Oxford University Press.

DeJong, D. J. and C. H. Whiteman (1991). Reconsidering 'trends and random walks in macroeconomic time series'. *Journal of Monetary Economics 28*, 221–254.

Fisher, M. (2017). Nonparametric density estimation using a mixture of order-statistic distributions. Technical report, Federal Reserve Bank of Atlanta.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis* (Third ed.). CRC Press.

Gershman, S. J. and D. M. Blei (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology 56*, 1–12.

Geweke, J. and G. Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies 9*(2), 557–587.

Gottardo, R. and A. E. Raftery (2008). Markov chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics 17*, 949–975.

Greenberg, E. (2013). *Introduction to Bayesian Econometrics* (Second ed.). Cambridge University Press.

Hamilton, J. D. (1994). *Times series analsis*. Princeton, NJ: Princeton University Press.

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

Jones, C. S. and J. Shanken (2005). Mutual fund performance with learning across funds. *Journal of Financial Economics 78*, 507–552.

Koop, G. (2003). *Bayesian Econometrics*. John Wiley & Sons.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press.

Lothian, J. R. and M. P. Taylor (1996). Real exchange rate behavior: The recent float from the perspective of the past two centuries. *Journal of Political Economy 104*(3), 488–509.

Poirier, D. J. (1991). A comment on 'to criticise the critics: An objective Bayesian analysis of stochastic trends'. *Journal of Applied Econometrics 6*, 381–386.

Schotman, P. and H. K. van Dijk (1991). A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics 49*, 195–238.

(Fisher) Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street N.E., Atlanta, GA 30309–4470

E-mail address: mark.fisher@atl.frb.org

URL: http://www.markfisher.net

(Dwyer) Department of Economics, Clemson University, Clemson, SC