PRIORS FOR THE LONG RUN

DOMENICO GIANNONE, MICHELE LENZA, AND GIORGIO E. PRIMICERI

Abstract. We propose a class of prior distributions that discipline the long-run behavior of Vector Autoregressions (VARs). These priors can be naturally elicited using economic theory, which provides guidance on the joint dynamics of macroeconomic time series in the long run. Our priors for the long run are conjugate, and can thus be easily implemented using dummy observations and combined with other popular priors. In VARs with standard macroeconomic variables, a prior based on the long-run predictions of a wide class of dynamic stochastic general equilibrium models yields substantial improvements in the forecasting performance.

1. Introduction

In this paper we propose a class of prior distributions that disciplines the long-run behavior of economic time series in Vector Autoregressions (VARs). Bayesian inference with informative priors has a long tradition for VARs, given that these models typically include many free parameters to accommodate general forms of autocorrelations and cross-correlations among variables. Therefore, with flat priors, such flexibility is likely to lead to in-sample overfitting and poor out-of-sample forecasting accuracy.

Our prior is motivated by a specific form of overfitting of flat-prior VARs, which is their tendency to attribute an implausibly large share of the variation in observed time series to a deterministic—and thus entirely predictable—component (Sims, 1996, 2000). In these models, inference is conducted by taking the initial observations of the variables as nonrandom. Therefore, the likelihood does not penalize parameter values implying that the variables' steady state (for stationary series, their trend for nonstationary ones) is distant from their initial observations. Complex transient dynamics from these initial conditions to the steady state are thus implicitly regarded as reasonable. As a consequence, they end up

Date: First version: December 2014. This version: August 2016.
We thank Gianni Amisano, Francesco Bianchi, Todd Clark, Gary Koop, Dimitris Korobilis, Ulrich Müller, Chris Sims, Jim Stock, Harald Uhlig, Herman van Dijk, Mark Watson, as well as seminar and conference participants for comments and suggestions. The views expressed in this paper are those of the authors and are not necessarily reflective of views at the European Central Bank, the Eurosystem, the Federal Reserve Bank of New York, or the Federal Reserve System.

explaining an implausibly large share of the low frequency variation of the data, yielding inaccurate out-of-sample forecasts.

Modifying inference to explicitly incorporate the density of the initial conditions may not be the right solution to this problem, since most macroeconomic time series are nonstationary and it is not obvious how to specify the distribution of their initial observations (examples of studies trying to address this issue include Phillips, 1991a,b, Kleibergen and van Dijk, 1994, Uhlig, 1994a,b and, more recently, Mueller and Elliott, 2003, and Jarocinski and Marcet, 2015, 2011). Following Sims and Zha (1998), an alternative route is to formulate a prior that expresses disbelief in an excessive explanatory power of the deterministic component of the model, by specifying that initial conditions should not be an important predictor of the subsequent evolution of the series. However, there are a variety of specific ways to implement this idea, especially in a multivariate setting.

Our main insight is that economic theory should play a central role for the elicitation of such a prior. Consider, for example, a simple bivariate VAR with the logarithm of GDP and investment. A wide class of theoretical models in macroeconomics predicts that these two variables should share a common stochastic trend, while the (log) investment-to-GDP ratio should be stationary. As a consequence, we might want to formulate a prior according to which the initial level of the common stochastic trend should explain very little of the subsequent dynamics of the system, while the initial conditions of the investment-to-GDP ratio should have a higher predictive power. In fact, if this variable is really mean reverting, then it is reasonable that initial conditions should shape the low frequency dynamics in the early part of the sample, while the variable converges back to its equilibrium value.

Our prior for the long run (PLR) is a formalization of this general concept. Its key ingredient is the choice of two orthogonal vector spaces, corresponding to the set of linear combinations of the model variables that are a-priori likely to be stationary and nonstationary. It is exactly for the identification of these two orthogonal spaces that economic theory plays a crucial role. The PLR essentially consists of shrinking the VAR coefficients towards values that imply little predictive power of the initial conditions of all these linear combinations of the variables, but particularly so for those that are likely to be nonstationary.

This idea of imposing priors informed by the long-run predictions of economic theory is reminiscent of the original insight of cointegration. However, our methodology differs from the classic literature on cointegration along two main dimensions. First of all, our fully probabilistic approach does not require to take a definite stance on the cointegration relations, but only on their plausible existence, thus avoiding the pre-testing and hard restrictions that typically plague error-correction models. More important, the focus of the cointegration literature is on identifying nonstationary linear combinations of the model variables, and dogmatically imposing that they cannot affect the short-run dynamics of the model, while remaining completely agnostic about the impact of the stationary combinations. On the contrary, we argue that shrinking the effect of these stationary combinations—albeit more gently—towards zero is at least as important as disciplining the impact of the common trends.

While we postpone the detailed description of our proposal to the main body of the paper, here we stress that our PLR is conjugate, and can thus be easily implemented using dummy observations and combined with existing popular priors, such as the Minnesota prior (Litterman, 1979). Moreover, conjugacy allows the closed-form computation of the marginal likelihood, which can be used to select the tightness of our PLR following an empirical Bayes approach, or conduct fully Bayesian inference on it based on a hierarchical interpretation of the model (Giannone et al., 2015).

We apply these ideas to the estimation of three VARs with an increasing number of standard macroeconomic variables. The first is a small-scale model with real variables such as output, consumption and investment. The second, medium-scale VAR also includes two labor market variables, i.e. real wages and hours worked. The third, larger-scale VAR also contains some nominal variables, such as inflation and the short-term interest rate. In each case, we set up our PLR based on the robust lessons of a wide class of dynamic stochastic general equilibrium (DSGE) models. Roughly speaking, these theories typically predict the existence of a common stochastic trend for the real variables, and possibly another trend for the nominal variables, while the ratios are likely to be stationary. We show that a PLR set up in accordance with these theoretical predictions is successful in reducing the explanatory power of the deterministic component implied by flat-prior VARs. To the extent that such explanatory power is spurious, this is a desirable feature of the model. In fact, a VAR with the PLR improves over more traditional BVARs in terms of out-of-sample forecasting performance, especially at long horizons.

The rest of the paper is organized as follows. Section 2 explains in what sense flat-prior VARs attribute too much explanatory power to initial conditions and deterministic trends. Section 3 illustrates our approach to solve this problem, i.e. our PLR. Section 4 puts our contribution in the context of a vast related literature, which is easier to do after having discussed the details of our procedure. Section 5 describes the results of our empirical application. Section 6 discusses some limitations of our approach and possible extensions to address them. Section 7 concludes.

2. Initial Conditions and Deterministic Trends

In this section, we show that flat-prior VARs tend to attribute an implausibly large share of the variation in observed time series to a deterministic—and thus entirely predictable—component. This problem motivates the specific prior distribution proposed in this paper. Most of the discussion in this section is based on the work of Sims (1996, 2000), although our recipe to address this pathology differs from his, as we will see in section 3.

To illustrate the problem, let us begin by considering the simple example of an AR(1) model,

$$y_t = c + \rho y_{t-1} + \varepsilon_t. \tag{2.1}$$

Equation (2.1) can be iterated backward to obtain

$$y_t \quad = \quad \underbrace{\rho^{t-1} y_1 \; + \; \sum_{j=0}^{t-2} \rho^j c}_{DC_t} \; + \; \underbrace{\sum_{j=0}^{t-2} \rho^j \varepsilon_{t-j}}_{SC_t}, \tag{2.2}$$

which shows that the model separates the observed variation of the data into two parts. The first component of (2.2)—denoted by DC_t—represents the counterfactual evolution of y_t in absence of shocks, starting from the initial observation y_1. Given that AR and VAR models are typically estimated treating the initial observation as given and non-random, DC_t corresponds to the deterministic component of y_t. The second component of (2.2)—denoted by SC_t—depends instead on the realization of all the shocks between time 2 and t, and thus corresponds to the unpredictable or stochastic component of y_t.

To analyze the properties of the deterministic component of y_t, it is useful to rewrite DC_t as

$$DC_t = \begin{cases} y_1 + (t-1)\,c & if\ \rho = 1 \\ \frac{c}{1-\rho} + \rho^{t-1}\left(y_1 - \frac{c}{1-\rho}\right) & if\ \rho \neq 1 \end{cases} .$$

If $\rho = 1$, the deterministic component is a simple linear trend. If instead $\rho \neq 1$, DC_t is an exponential, and has a potentially more complex shape as a function of time. The problem is that, when conducting inference, this potentially complex deterministic dynamics arising from estimates of $\rho \neq 1$ can be exploited to fit the low frequency variation of y_t, even when such variation is mostly stochastic. This peculiar "overfitting" behavior of the deterministic component is clearly undesirable. According to Sims (2000), it is due to two main reasons. First, the treatment of initial observations as non-stochastic removes any penalization in the likelihood for parameter estimates that imply a large distance between y_1 and $\frac{c}{1-\rho}$ (the unconditional mean of the process in the stationary case) and, as such, magnifies the effect of the ρ^{t-1} term in DC_t. Second, the use of a flat prior on (c, ρ) implies an informative prior on $\left(\frac{c}{1-\rho}, \rho\right)$, with little density in the proximity of $\rho = 1$, and thus on an approximately linear behavior of DC_t.

Sims (2000) illustrates this pathology by simulating artificial data from a random walk process, and analyzing the deterministic component implied by the flat-prior parameter estimates of an AR(1) model. By construction, all the variation in the simulated data is stochastic. Nevertheless, the estimated model has the tendency to attribute a large fraction of the low frequency behavior of the series to the deterministic component, i.e. to a path of convergence from unlikely initial observations to the unconditional mean of the process. In addition, Sims (2000) argues that the fraction of the sample variation due to the deterministic component converges to a non-zero distribution, if the data-generating process is a random walk without drift. We formally prove this theoretical result in appendix A, and show that it also holds when the true data-generating process is local-to-unity. Put it differently, if the true data-generating process exhibits a high degree of autocorrelation, estimated AR models will imply a spurious explanatory power of the deterministic component even in arbitrarily large samples.

The problem is much worse in VARs with more variables and lags, since these models imply a potentially much more complex behavior of the deterministic trends. For example, the deterministic component of an n-variable VAR with p lags is a linear combination of $n \cdot p$

exponential functions plus a constant term. As a result, it can reproduce rather complicated low-frequency dynamics of economic time series.

To illustrate the severity of the problem in real applications, consider the popular 7-variable VAR with log-real GDP, log-real consumption, log-real investment, log-real wages, log hours worked, inflation and a short-term nominal interest rate (Smets and Wouters, 2007, Del Negro et al., 2007, Justiniano et al., 2010). Suppose that a researcher is estimating this model at the end of 1994, using forty years of quarterly data and 5 lags. Figure 2.1 plots the deterministic components implied by the flat-prior (OLS) estimates for six representative time series, along with their actual realization between 1955:I and 1994:IV. First of all, notice that these deterministic trends are more complex at the beginning of the sample. For example, the predictable component of the investment-to-GDP ratio fluctuates substantially between 1955 and 1970, more so than in the rest of the sample.

In addition to exhibiting this marked temporal heterogeneity (Sims, 2000), the deterministic component also seems to explain a large share of the variation of these time series. Consistent with theory, this feature is most evident for the case of persistent series without (or with little) drift, such as hours, inflation, the interest rate or the investment-to-GDP ratio. For instance, the estimated model implies that most of the hump-shaped low-frequency behavior of the federal funds rate was due to deterministic factors, and was thus predictable since as far as 1955 for a person with the knowledge of the VAR coefficients. And so was the fact that interest rates would hit the zero lower bound around 2010.

Most economists would be skeptical of this likely spurious explanatory power of deterministic trends, and may want to downplay it when conducting inference. In principle, “one way to accomplish this is to use priors favoring pure unit-root low frequency behavior” (Sims, 2000, pp. 451), according to which implausibly precise long-term forecasts are unlikely. However, it is not obvious how to formulate such a prior. For example, the undesirable properties of the deterministic component persist even when using the popular Minnesota prior with conventional tightness (Litterman, 1979, Sims and Zha, 1998, see appendix C), as shown in figure 2.1. In the next section we detail our specific proposal regarding how to address this problem.

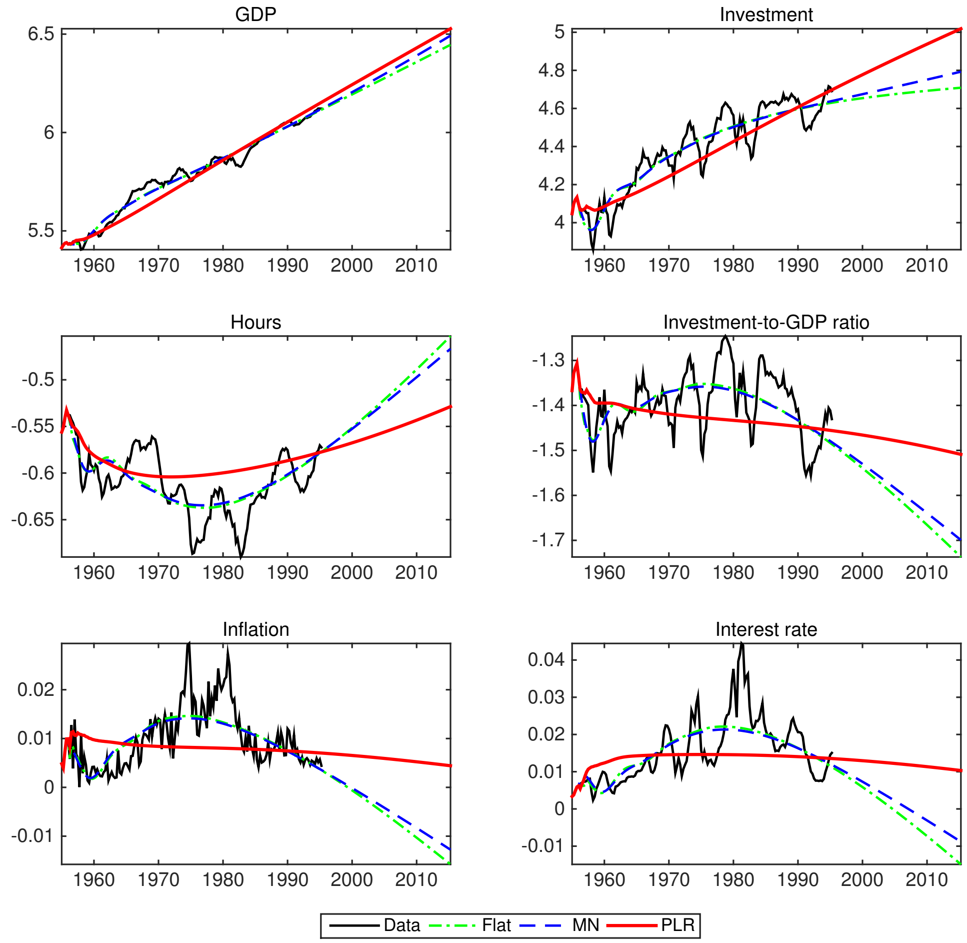

FIGURE 2.1. Deterministic component for selected variables implied by various 7-variable VARs. Flat: BVAR with a flat prior; MN: BVAR with the Minnesota prior; PLR: BVAR with the prior for the long run.

3. Elicitation of a Prior for the Long Run

Consider the VAR model

$$y_t = c + B_1 y_{t-1} + .. + B_p y_{t-p} + \varepsilon_t \tag{3.1}$$

$$\varepsilon_t \sim \text{i.i.d. } N(0, \Sigma),$$

where y_t is an $n \times 1$ vector of endogenous variables, ε_t is an $n \times 1$ vector of exogenous shocks, and c, B_1,..., B_p and Σ are matrices of suitable dimensions containing the model's unknown parameters. The model can be rewritten in terms of levels and differences

$$\Delta y_t = c + \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} ... + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{3.2}$$

where $\Pi = (B_1 + \ldots + B_p) - I_n$ and $\Gamma_j = -(B_{j+1} + \ldots + B_p)$, with $j = 1, ..., p - 1$.

The aim of this paper is to elicit a prior for Π. To address the problems described in the previous section, we consider priors that are centered around zero. As for the prior covariance matrix on the elements of Π, our main insight is that its choice must be guided by economic theory, and that alternative—automated or "theory-free"—approaches are likely to lead to a prior specification with undesirable features.

To develop this argument, let H be any invertible $n-$dimensional matrix, and rewrite (3.2) as

$$\Delta y_t = c + \Lambda \tilde{y}_{t-1} + \Gamma_1 \Delta y_{t-1} ... + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{3.3}$$

where $\tilde{y}_{t-1} = H y_{t-1}$ is an $n \times 1$ vector containing n linearly independent combinations of the variables y_{t-1}, and $\Lambda = \Pi H^{-1}$ is an $n \times n$ matrix of coefficients capturing the effect of these linear combinations on Δy_t. In this transformed model, the problem of setting up a prior on Π corresponds to choosing a prior for Λ, conditional on the selection of a specific matrix H. What is a reasonable prior for Λ will then depend on the choice of H. For example, consider an H matrix whose $i-$th row contains the coefficients of a linear combination of y that is a priori likely to be mean reverting. Then it would surely be unwise to place a prior on the elements of the $i-$th column of Λ that is excessively tight around zero. In fact, following the standard logic of cointegration, if the elements of the $i-$th column of Λ were all zero, there would not be any "error-correction" mechanism at play to preserve the stationarity of this linear combination of y. A similar logic would suggest that, if a raw of H contains the coefficients of an a-priori likely nonstationary linear combination of y, one can afford more shrinkage on the elements of the corresponding column of Λ.

This simple argument suggests that it is important to set up different priors on the loadings associated with linear combinations of y with different degree of stationarity. This objective can be achieved by formulating a prior on Λ, conditional on a choice of H that combines the data in a way that a-priori likely stationary combinations are separated from the nonstationary ones.

Interestingly, in many contexts, economic theory can provide useful information for choosing a matrix H with these characteristics. For example, according to the workhorse macroeconomic model, output, consumption and investment are likely to share a stochastic trend, while both the consumption-to-output and the investment-to-output ratios should be stationary variables. Similarly, standard economic theory would predict that the price of different goods might be trending, while relative prices should be mean reverting (in absence of differential growth in the production technology of these goods).[1] If these statements were literally true, the corresponding VARs would have an exact error-correction representation, as in Engle and Granger (1987), with a reduced-rank Π matrix. In practice, it is difficult to say with absolute confidence whether certain linear combinations of the data are stationary or integrated. It might therefore be helpful to work with a prior density that is based on some robust insights of economic theory, while also allowing the posterior estimates to deviate from them, based on the likelihood information.

We operationalize these ideas by specifying the following prior distribution on the loadings Λ (as opposed to Π), conditional on a specific choice of the matrix H:

$$\Lambda_{\cdot i} | H_{i\cdot}, \Sigma \sim N\left(0, \tilde{\phi}_i\left(H_{i\cdot}\right)\Sigma\right), \qquad i = 1, ..., n, \tag{3.4}$$

where $\Lambda_{\cdot i}$ denotes the i-th column of Λ, and $\tilde{\phi}_i\left(H_{i\cdot}\right)$ is a scalar hyperparameter that is allowed to depend on $H_{i\cdot}$, the i-th row of H. For tractability, we also assume that these priors are scaled by the variance of the error Σ, are independent across i's and Gaussian, which guarantees conjugacy. Notice, however, that the assumption that the priors on the columns of Λ are independent from each other does not rule out (and will in general imply) that the priors on the columns of Π are correlated, with a correlation structure that depends on the choice of H and $\tilde{\phi}$.

The tightness of the prior in (3.4) is controlled by the hyperparameter $\tilde{\phi}_i\left(H_{i\cdot}\right)$. One way to choose its value is based on subjective considerations. An alternative (empirical Bayes) strategy is to set $\tilde{\phi}_i\left(H_{i\cdot}\right)$ by maximizing the marginal likelihood, which is the likelihood of the model only as a function of the hyperparameters. Thanks to the conjugacy of the prior, the marginal likelihood is available in closed form and is thus very easy to compute.

[1] Economic theory usually identifies the set of nonstationary combinations of the model variables, and the space spanned by the stationary combinations. To form the H matrix, our baseline PLR requires the selection of one specific set of linear combinations belonging to this space. In section 6, we also develop an extension of our methodology that is invariant to rotations within this space.

A third option, in between these two extremes, is to adopt a hierarchical interpretation of the model, and set $\tilde{\phi}_i\left(H_{i\cdot}\right)$ based on its posterior distribution, which combines the marginal likelihood with a hyperprior (Giannone et al., 2015). This is the approach that we adopt in our empirical applications. In the next subsection, we describe a reference parameterization that facilitates the choice of hyperpriors or subjective values for $\tilde{\phi}_i\left(H_{i\cdot}\right)$.

3.1. **Reference value for $\tilde{\phi}_i$.** A crucial element of the density specified in (3.4) is the fact that $\tilde{\phi}_i$ can be a function of $H_{i\cdot}$, which is consistent with the intuition that the tightness of the prior on the loadings $\Lambda_{\cdot i}$ should depend on whether these loadings multiply a likely stationary or nonstationary linear combination of y from an a-priori perspective. To capture this important insight, we propose a reference parameterization of $\tilde{\phi}_i$ as follows:

$$\tilde{\phi}_i(H_{i\cdot}) = \frac{\phi_i^2}{\left(H_{i\cdot}\bar{y}_0\right)^2}, \tag{3.5}$$

where ϕ_i is a scalar hyperparameter (controlling the standard deviation of the prior on the elements of $\Lambda_{\cdot i}$), and $\bar{y}_0$ is a column vector containing the average of the initial p observations of each variable of the model (these are the observations taken as given in the computation of the likelihood function). Therefore, the denominator of (3.5) corresponds to the square of the initial value of the linear combination at hand.

There are a few reasons why this reference formulation is appealing. First of all, substituting (3.5) into (3.4) makes it clear that the prior variance of $\Lambda_{\cdot i}$ has a scaling that is similar to that of the likelihood, with the variance of the error at the numerator, and the (sum of) squared regressor(s) at the denominator (recall for instance the form of the variance of the OLS estimator). Second, expression (3.5) captures the insight that tighter priors are more desirable for the loadings of nonstationary linear combinations of y, which are likely to have larger initial values (assuming that the data generating process has been in place for a long enough period of time before the observed sample).[2] Third, scaling the prior variance by $1/\left(H_{i\cdot}\bar{y}_0\right)^2$ is more attractive than any alternative scale meant to capture the same idea, because in this way the prior setup does not rely on any information that is also used to construct the likelihood function, avoiding any type of “double counting” of the data.

[2] More specifically, suppose that the true data-generating process of $H_{i\cdot}y_t$ has been in place for a number of periods T_0, where T_0 is proportional to the observed sample size T. In the stationary case, $\left(H_{i\cdot}\bar{y}_0\right)^2$ is bounded in probability. It is instead of order T^{κ} in the integrated or local-to-unity case, where κ is equal to 1 or 3/2 depending on the presence of the drift.

In the reference parameterization (3.5), the prior tightness is controlled by the hyperparameter ϕ_i, which is just a monotone transformation of $\tilde{\phi}_i$. Therefore, from a theoretical point of view, the problem of choosing ϕ_i is identical to selecting $\tilde{\phi}_i$, and can also be based on subjective considerations, the maximization of the marginal likelihood, or a combination of the two.[3] In practice, however, the choice of a specific subjective value—or a hyperprior—is easier for ϕ_i than for $\tilde{\phi}_i$, because it has a more direct connection with the problem of the initial conditions and deterministic trends that we have highlighted in section 2. We clarify this point in the next subsection, where we explain how to implement this prior using simple dummy observations, and provide some additional insights into its interpretation.

3.2. **Implementation with dummy observations.**

The prior in (3.4) can be rewritten in a more compact form as

$$vec\left(\Lambda\right)|H,\Sigma\sim N\left(0,\tilde{\Phi}_H\otimes\Sigma\right), \tag{3.6}$$

with $\tilde{\Phi}_H = diag\left(\left[\tilde{\phi}_1\left(H_{1\cdot}\right),...,\tilde{\phi}_n\left(H_{n\cdot}\right)\right]\right)$, where $vec\left(\cdot\right)$ is the vectorization operator, and $diag\left(x\right)$ denotes a diagonal matrix with the vector x on the main diagonal. Since $\Pi=\Lambda H$, the implied prior on the columns of Π is given by

$$vec\left(\Pi\right)|H,\Sigma\sim N\left(0,H'\tilde{\Phi}_H H\otimes\Sigma\right). \tag{3.7}$$

Being conjugate, this prior can be easily implemented using Theil mixed estimation, i.e. by adding a set of n artificial (or dummy) observations to the original sample. Each of these n dummy observations consists of a value of the variables on the left- and right-hand side of (3.1), at an artificial time t_i^*. In particular, the implementation of the prior in (3.7), with the parameterization of $\tilde{\phi}_i$ in (3.5), requires the following set of artificial observations:

$$y_{t_i^*}=y_{t_i^*-1}=...=y_{t_i^*-p}=\frac{H_{i\cdot}\bar{y}_0}{\phi_i}\left[H^{-1}\right]_{\cdot i}, \qquad i=1,...,n, \tag{3.8}$$

[3] For example, given the invariance property of maximum likelihood, an Empirical Bayes approach based on the maximization of the marginal likelihood would lead to identical inference regardless of whether one uses this specific parameterization or not.

where the corresponding observation multiplying the constant term is set to zero, and $\left[H^{-1}\right]_{\cdot i}$ denotes the i-th column of H^{-1}. We prove this result in appendix B, where we also derive the posterior distribution of the model's unknown coefficients.

To provide yet another interpretation of our prior, it is useful to substitute the dummy observations (3.8) into the level-difference representation of the model (3.2), obtaining

$$0 = \underbrace{\Pi \left[H^{-1}\right]_{\cdot i}}_{\Lambda_{\cdot i}} \left(H_{i \cdot} \bar{y}_0\right) + \phi_i \varepsilon_{t_i^*}, \qquad i = 1, ..., n. \tag{3.9}$$

This expression suggests that the prior is effectively limiting the extent to which the linear combinations $H_{i\cdot} y$ help forecasting Δy at the beginning of the sample. This feature reduces the importance of the error correction mechanisms of the model, which are responsible for the complex dynamics and excessive explanatory power of the deterministic component that we have analyzed in section 2. However, given that the value of $H_{i\cdot}\bar{y}_0$ is typically lower (in absolute value) for mean-reverting combinations of y, our prior reduces more gently the mechanisms that correct the deviations from equilibrium of likely stationary combinations of the variables, consistent with the idea of cointegration.

The representation of the prior in terms of dummy observations also provides some useful insights for the elicitation of a hyperprior. The value of $\phi_i = 1$ corresponds to using a single artificial observation in which the linear combination of variables on the right- and left-hand side is equal to its initial condition, with an error variance of this observation similar to that in the actual sample. Therefore, 1 seems a sensible reference value for ϕ_i, and we use it to center its hyperprior (we also choose 1 as a standard deviation for this hyperprior, see appendix C for details).

3.3. **Simple bivariate example.** Before turning to a more comprehensive comparison with some of the existing literature, it is useful to contrast our prior for the long run to the more standard sum-of-coefficients (SOC) prior, first proposed by Doan et al. (1984), and routinely used for the estimation of BVARs (Sims and Zha, 1998). The SOC prior also disciplines the sum of coefficients on the lags of each equation of the VAR, but corresponds to mechanically setting H equal to the identity matrix, even when there might be some linear combinations of the variables in the system that should be stationary.

For the sake of concreteness, consider the simple example of a bivariate VAR(1) with log-output (x_t) and log-investment (i_t). The SOC prior corresponds to

$$vec\,(\Pi)\,|\Sigma \sim N\left(0, \begin{bmatrix} \frac{\mu^2}{x_0^2} & 0 \\ 0 & \frac{\mu^2}{i_0^2} \end{bmatrix} \otimes \Sigma\right),$$

where μ is an hyperparameter controlling its overall tightness. Economic theory, however, does suggest that output and investment are likely to share a common trend ($x_t + i_i$), while the log-investment-to-output ratio ($i_t - x_t$) is expected to be stationary. Based on this insight, we can form the matrix H, whose rows correspond to the coefficients of these two different linear combinations of the variables:

$$H = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}. \tag{3.10}$$

One can now ask what is the prior implied by SOC on the coefficients capturing the effect of these two linear combinations on Δx_t and Δi_t (i.e. on the error correction coefficients, using the cointegration terminology). To answer this question, recall that $\Lambda = \Pi H^{-1}$, which implies that $vec\,(\Lambda) = \left(\left(H^{-1}\right)' \otimes I_n\right) vec\,(\Pi)$, and thus

$$vec\,(\Lambda)\,|H, \Sigma \sim N\left(0, \frac{1}{4}\begin{bmatrix} \frac{\mu^2}{x_0^2} + \frac{\mu^2}{i_0^2} & \frac{\mu^2}{i_0^2} - \frac{\mu^2}{x_0^2} \\ \frac{\mu^2}{i_0^2} - \frac{\mu^2}{x_0^2} & \frac{\mu^2}{x_0^2} + \frac{\mu^2}{i_0^2} \end{bmatrix} \otimes \Sigma\right).$$

Notice that the prior on the loadings of the common trend is as tight as that on the loadings of the investment ratio, which is in contrast with the predictions of most theoretical models and with the main insights of cointegration.

On the contrary, our prior for the long run with the choice of H in (3.10) corresponds to the following prior density on the error correction coefficients:

$$vec\,(\Lambda)\,|H, \Sigma \sim N\left(0, \begin{bmatrix} \frac{\phi_1^2}{(x_0+i_0)^2} & 0 \\ 0 & \frac{\phi_2^2}{(i_0-x_0)^2} \end{bmatrix} \otimes \Sigma\right).$$

Clearly, even if $\phi_1 \approx \phi_2$, and given that $(i_0 - x_0)^2$ is a much smaller number than $(x_0 + i_0)^2$, this prior performs much less shrinkage on the coefficients that correct the deviations of the investment ratio from its equilibrium value.

4. Relationship with the Literature

Before turning to the empirical application, it is useful to relate our approach more precisely to the literature on cointegration (Engle and Granger, 1987) and error-correction models (for a comprehensive review, see Watson, 1986). For the purpose of making this comparison as concrete as possible, suppose that the specific model at hand entails a natural choice of H with the following two blocks of rows:

$$H = \begin{bmatrix} \underset{(n-r)\times n}{\beta_{\perp}'} \\ \underset{r\times n}{\beta'} \end{bmatrix}, \tag{4.1}$$

where the columns of $\beta_{\perp}$ are $(n-r)$ linear combinations of y that are likely to exhibit a stochastic trend, while the columns of β are r linear combinations of y that are more likely to represent stationary deviations from long-run equilibria, i.e. that are likely to correspond to cointegrating vectors. Using this notation, we can rewrite (3.3) as

$$\Delta y_t = c + \Lambda_1 \left(\beta_{\perp}' y_{t-1}\right) + \Lambda_2 \left(\beta' y_{t-1}\right) + \Gamma_1 \Delta y_{t-1} ... + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{4.2}$$

where Λ_1 are the first $n-r$ columns of Λ, and Λ_2 are the remaining r columns.

As described in the previous section, our approach consists of placing priors on the columns of Λ_1 and Λ_2. These priors are centered around zero, and are tighter for the elements of Λ_1 than for those of Λ_2. The error-correction representation corresponds to an extreme case of our general model, obtained by enforcing a dogmatic prior belief that $\Lambda_1 = 0$. As a result, Π would equal $\Lambda_2\beta'$, and would be rank deficient. If, in addition, the prior belief that $\Lambda_2 = 0$ is also dogmatically imposed, the VAR admits a representation in first differences.

For what concerns the cointegrating vectors β, the literature has proceeded by either fixing or estimating them. The approach that is closer to ours selects the cointegrating vectors β a priori, mostly based on economic theory. This strategy is appealing since the theoretical cointegrating relations are typically quite simple and robust across a wide class of economic models. Conditional on a specific choice of β, one popular approach is to include all the theoretical cointegrating vectors in the error-correction representation, and conduct likelihood-based inference (i.e. OLS), as in King et al. (1991) or Altig et al. (2011). In our model, this is equivalent to placing a flat prior on Λ_2. An alternative strategy, however, has been to conduct some pre-testing and to include in the error correction only those deviations from equilibria for which the adjustment coefficients are statistically significantly different from zero, as in Horvath and Watson (1995). Conditional on the pre-testing results, this approach is equivalent to setting a dogmatic prior that certain columns of Λ_2 are equal to zero, and a flat prior on the remaining elements.

The other strand of the literature is more agnostic about both the cointegrating rank (r) and the cointegrating vectors (β). In these cases, classical inference is typically conducted using a multi-step methodology. The first step of these procedures requires testing for the cointegrating rank. Conditional on the results of these tests, the second step consists of the estimation of the cointegrating vectors, which are then treated as known in step three for the estimation of the remaining model parameters (Engle and Granger, 1987). Alternatively, the second and third steps can be combined to jointly estimate β and the other parameters with likelihood-based methods, as in Johansen (1995).

The Bayesian approach to cointegration is similar in spirit to the likelihood-based inference (for recent surveys, see Koop et al., 2006, Del Negro and Schorfheide, 2011, and Karlsson, 2013). This literature has also concentrated on conducting inference on the cointegrating rank and the cointegrating space. For example, the number of cointegrating relationships is typically selected using the marginal likelihood, or related Bayesian model comparison methods (Chao and Phillips, 1999, Kleibergen and Paap, 2002, Corander and Villani, 2004, Villani, 2005). In practical applications, this methodology ends up being similar to pre-testing because the uncertainty on the cointegrating rank is seldom formally incorporated into the analysis, despite the fact that the Bayesian approach would allow it (for an exception, see Villani, 2001).

Conditional on the rank, the early Bayesian cointegration literature was concerned with formulating priors on the cointegrating vectors, and with deriving and simulating their posterior (Bauwens and Lubrano, 1996, Geweke, 1996). Standard priors, however, have been shown to be problematic, in light of the pervasive local and global identification issues of error-correction models (Kleibergen and van Dijk, 1994, Strachan and van Dijk, 2005). To avoid these problems, a better strategy is to place a prior on the cointegrating space, which is the only object the data are informative about (Villani, 2000). Such priors are studied in Strachan and Inder (2004) and Villani (2005), who also develop methods for inference and posterior simulations. In particular, Villani (2005) proposes a diffuse prior on the cointegrating space, trying to provide a Bayesian interpretation to some popular likelihood-based procedures. In general, little attention has been given to the elicitation of informative priors on the adjustment coefficients, which is instead the main focus of our paper.

It is well known that maximum-likelihood (or flat-prior) inference in the context of error-correction models can be tricky (Stock, 2010). This is not only because of the practice to condition on initial conditions, as we have stressed earlier, but also because inference is extremely sensitive to the value of non-estimable nuisance parameters characterizing small deviations from non-stationarity of some variables (Elliott, 1998, Mueller and Watson, 2008). Pretesting is clearly plagued by the same problems. The selection of models based on pre-testing or Bayesian model comparison can be thought as limiting cases of our approach, in which the support of the distributions of the hyperparameters controlling the tightness of the prior on specific adjustment coefficients can only take values equal to zero or infinity. One advantage of our flexible modeling approach, instead, is that it removes such an extreme sparsity of the model space, as generally recommended by Gelman et al. (2004) and Sims (2003).

Finally, our paper is also related to the methodology of Del Negro and Schorfheide (2004) and Del Negro et al. (2007), who also use a theoretical DSGE model to set up a prior for the VAR coefficients. Their work, however, differs from ours in two important ways. First of all, the prior of Del Negro et al. (2007) is centered on the error-correction representation of the VAR, given that such a prior pushes towards a DSGE model featuring a balanced growth path. On the contrary, for the reasons highlighted in section 2, our PLR shrinks the VAR towards the representation in first differences, albeit it does so more gently for the linear combinations of the variables that are supposed to be stationary according to theory. In addition, the approach of Del Negro and Schorfheide (2004) requires the complete specification of a DSGE model, including its short-run dynamics. Instead, we use only the long-run predictions of a wide class of theoretical models to guide the setup of our PLR. Among other things, this strategy allows us to work with a conjugate prior and simplify inference.

5. Empirical Results

In this section we use our prior to conduct inference in VARs with standard macroeconomic variables, whose joint long-run dynamics is sharply pinned down by economic theory. In particular, we perform two related, but distinct exercises. We begin by re-estimating the 7-variable VAR of section 2, to show that our PLR serves the purpose of reducing the excessive explanatory power of the deterministic components implied by the model with flat or Minnesota priors. Second, we evaluate the forecasting performance of 3-, 5- and 7-variable VARs, and demonstrate that our prior yields substantial gains over more standard BVARs, especially when forecasting at long horizons. Before turning to the detailed illustration of these results, we begin by describing more precisely the 3-, 5- and 7-variable VARs and the priors that we adopt.

The 3-variable VAR includes data on log-real output (Y_t), log-real consumption (C_t) and log-real investment (I_t) for the US economy, and is similar to the VAR estimated by King et al. (1991) in their influential analysis of the sources of business cycles.[4] This model is appealing because of its simplicity and because, in this context, a prior for the long run can be easily elicited based on standard economic theory, which has robust implications about the long-run behavior of these three time series. Specifically, a wide class of macroeconomic models predicts the existence of a balanced growth path, along which output, consumption and investment share a common trend, while the great ratios (the consumption- and investment-to-output ratios) should be stationary.

The 5- and 7-variable VARs augment the small-scale model with labor-market variables—log-real wages (W_t) and log-hours worked (H_t)—and nominal variables—inflation (π_t) and the federal funds rate (R_t). These are the same time series used to estimate the DSGE model of Smets and Wouters (2007), which builds on the Real Business Cycle (RBC) literature by adding a number of real and nominal frictions. This DSGE is representative of modern medium-to-large-scale macroeconomic models, and can thus be used as a guide to set up our prior in this context. Similar to the RBC framework, this class of models typically predicts that real output, consumption, investment, and real wages share a common stochastic trend, while the great ratios (the labor share, and the consumption- and investment-to-output ratios) and hours worked should be stationary.

[4] The data used in the empirical applications are described in detail in appendix D.

In addition, some New-Keynesian models (e.g. Ireland, 2007) also include a stochastic nominal trend, common to the interest and inflation rates. While the existence of such a stochastic nominal trend is not a robust feature of this class of models, most of them do imply that the low-frequency behavior of inflation and interest rates are tightly related. This is exactly the type of situation in which it might be beneficial to formulate a prior that is centered on the existence of a common nominal trend, without imposing it dogmatically.

A compact way of summarizing the variables included in each model and the linear combinations used to set up our PLR is to illustrate the details of the choice of $\tilde{y}_t$ and H for the larger, 7-variable model:

$$\tilde{y}_t = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}}_{H} \underbrace{\begin{pmatrix} Y_t \\ C_t \\ I_t \\ W_t \\ H_t \\ \pi_t \\ R_t \end{pmatrix}}_{y_t} \begin{array}{l} \rightarrow \text{real trend} \\ \rightarrow \text{consumption-to-GDP ratio} \\ \rightarrow \text{investment-to-GDP ratio} \\ \rightarrow \text{labor share} \\ \rightarrow \text{hours} \\ \rightarrow \text{nominal trend} \\ \rightarrow \text{real interest rate.} \end{array}$$

The 5-variable model only includes the first 5 variables of y_t and the 5×5 upper-left block of H. The 3-variable VAR only includes the first 3 variables of y_t and the 3×3 upper-left block of H.

We now turn to the description of the specific exercises that we conduct and the empirical results.

5.1. **Deterministic trends.** In section 2, we have argued that a serious pathology of flat-prior VARs is that they imply rather complex dynamics and excessive explanatory power of the deterministic component of economic time series (Sims, 1996, 2000). In addition, the use of the standard Minnesota prior (with conventional hyperparameter values) does very little, if nothing at all, to solve the problem (figure 2.1). In this subsection, we analyze the extent to which our PLR eliminates or reduces this pathology.

To this end, we re-estimate the 7-variable VAR of section 2 using our prior for the long run (PLR-BVAR), and compare the deterministic trends implied by this model to those of section 2, obtained using a BVAR with flat or Minnesota priors (flat-BVAR and MN-BVAR respectively).[5] In this experiment, for simplicity, we simply set the value of the hyperparameters $\{\phi_i\}_{i=1}^{n}$ equal to one, which provides a good reference value (it corresponds to adding one dummy observation, see appendix C).

The results of this experiment are depicted in figure 2.1. In the case of GDP, the difference between the deterministic component of the PLR-BVAR and the flat or MN-BVAR is limited. For the other variables, notice that the shape of the deterministic component implied by the PLR-BVAR is simpler, and explains much less of the low frequency variation of the time series. For example, in the case of investment, the deterministic trend implied by the PLR-BVAR resembles a straight line, implying that the long-run growth rate of investment in the next decades is expected to be similar to the past. Similarly, in the case of inflation and the interest rate, the deterministic trend of the PLR-BVAR does not have the unpleasant property that somebody with the knowledge of the VAR coefficients would have perfectly predicted the hump shape of these two variables already in 1955.

So, overall, our PLR is quite successful in correcting the pathology that we have illustrated in section 2. In the next section, we will demonstrate that this is not simply a theoretical curiosity, but that it is extremely important for the forecasting performance of the model.

5.2. **Forecasting performance.** In this subsection, we compare the forecasting performance of our BVAR to a number of benchmark BVARs. More specifically, we consider the following models:

- MN-BVAR: BVAR with the Minnesota prior
- SZ-BVAR: BVAR with the Minnesota prior and the sum-of-coefficients (also known as no-cointegration) prior, as in the work of Doan et al. (1984) and Sims and Zha (1998). The latter corresponds to our prior for the long run with a mechanic choice of H equal to the identity matrix, and the same hyperparameter for each dummy observation. It has the effect of pushing the VAR parameter estimates towards the existence of a separate stochastic trend for each variable.
- Naive: BVAR with an infinitely tight Minnesota prior, which results in all the variables following independent random walks with drifts.

[5] We conduct this experiment with the 7-variable VAR, as opposed to the 3- or 5-variable VARs, because the problem is more severe in this case.

- PLR-BVAR: BVAR with the Minnesota prior and our prior for the long run.

This comparison of forecasting accuracy is interesting because the first three models are considered valid benchmarks in the literature. For example, it is well known that MN-BVARs yield substantial forecasting improvements over classical or flat-prior VARs (Litterman, 1979) and that further improvements can be achieved by adding the sum-of-coefficients prior of Doan et al. (1984). In fact, Giannone et al. (2015) show that the predictive ability of the model of Sims and Zha (1998) is comparable to that of factor models. Finally, a number of papers have demonstrated that the naive random walk model forecasts quite well, especially after the overall decline in predictability of macroeconomic time series in 1985 (Atkeson and Ohanian., 2001, Stock and Watson, 2007, D'Agostino et al., 2007, Rossi and Sekhposyan, 2010).

In what follows, our measure of forecasting accuracy is the out-of-sample mean squared forecast error (MSFE). In particular, for each of the four models, we produce the 1- to 40-quarters-ahead forecasts, starting with the estimation sample that ranges from 1955Q1 to 1974Q4. We then iterate the same procedure updating the estimation sample, one quarter at a time, until the end of the sample in 2013Q1. At each iteration, we select the tightness of the priors by maximizing the posterior of the models' hyperparameters, using the procedure proposed by Giannone et al. (2015) and summarized in appendix C. Conditional on a specific value of these hyperparameters, we then produce the out-of-sample forecasts by setting the VAR coefficients to their posterior mode. All BVARs are estimated using 5 lags. For all the forecast horizons, the evaluation sample for the computation of the MSFEs ranges from 1985Q1 to 2013Q1.

5.2.1. *3-variable VARs.* We start by focusing on the small-scale model with three variables. The upper panel in figure 5.1 reports the MSFEs of the level of each variable at horizons ranging from one to 40 quarters ahead. Christoffersen and Diebold (1998) point out that, in presence of long-run relationships across the variables, accuracy measures should adequately value the ability of the different models to preserve such long-run relationships. Hence, to assess the different models under analysis also on such grounds, in the lower panel of figure 5.1 we report the out-of-sample accuracy measures for the common trend and, more importantly, for the so-called great ratios.

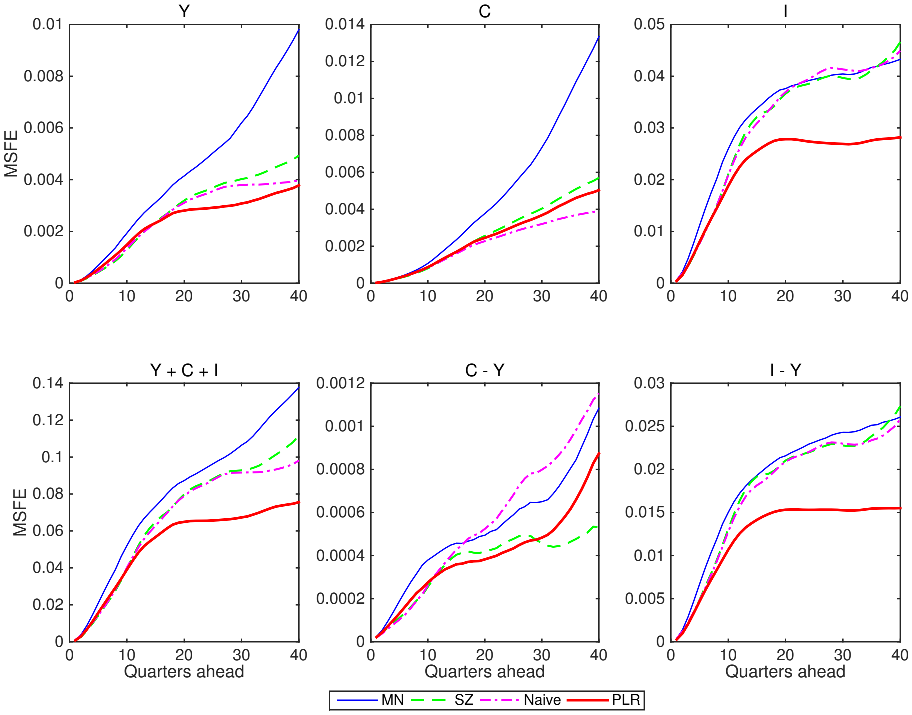

FIGURE 5.1. Mean squared forecast errors in models with three variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficient priors; Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run.

Notice that the PLR-BVAR improves uniformly over the MN-BVAR, especially at long horizons, reflecting the fact that the Minnesota prior alone is not enough to reduce the spurious explanatory power of the deterministic component typical of flat-prior VARs. According to the existing literature, one way to reduce this pathology is to augment the MN-BVAR with a sum-of-coefficients prior. The resulting SZ-BVAR does outperform the MN-BVAR, but is still substantially less accurate than the PLR-BVAR for predicting investment and the investment-to-GDP ratio. Finally, observe that the forecasting performance of the naive model is very similar to that of the SZ-BVAR, which suggests that the sum-of-coefficients prior strongly shrinks the VAR coefficients toward values consistent with the existence of independent random walks for all three variables.

The key question for us is understanding why the PLR-BVAR outperforms the SZ-BVAR and the naive model. We address this question in figure 5.2, which plots the realized value of the consumption- and investment-to-GDP ratios, and the 5-year-ahead forecasts obtained

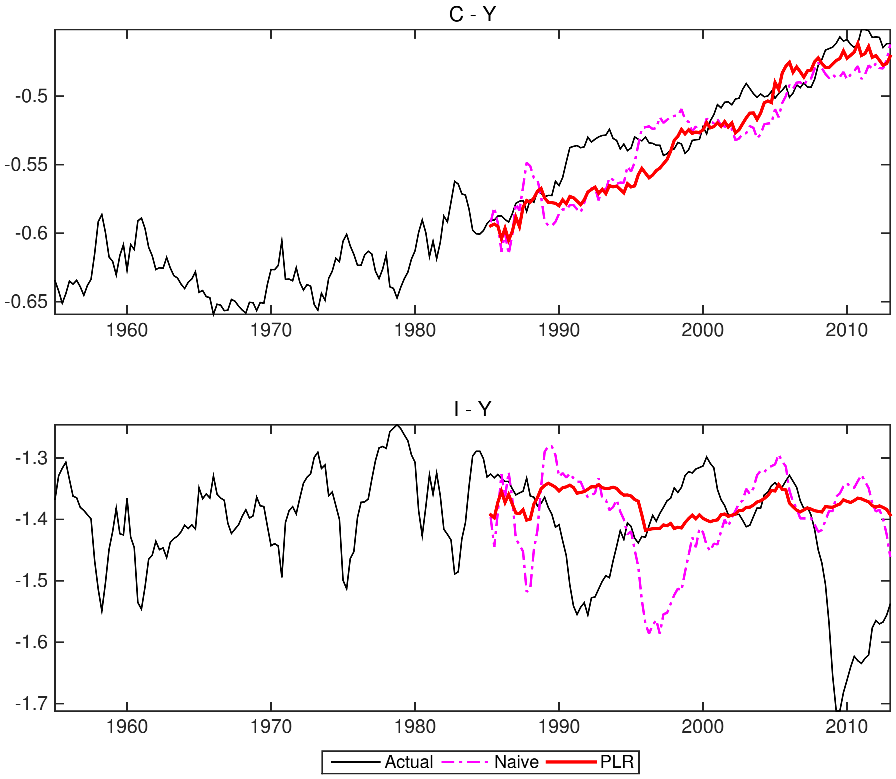

FIGURE 5.2. Forecasts of the great ratios 5-years ahead in models with three variables. Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run.

with the PLR-BVAR and the naive model (the SZ-BVAR forecasts are very close to those of the naive model, so we do not report them to avoid clogging the figure). Since all the variables follow separate random walks in the naive model, the difference between log-consumption and log-output, and the difference between log-investment and log-output are also random walks. As the first panel of figure 5.2 makes clear, a random walk is a pretty good predictor of the consumption-to-GDP ratio because this variable displays a (close to) nonstationary behavior in the data. The no-change forecasts of a random walk, however, are poor predictors of the investment-to-GDP ratio at long horizons, because this series looks mean reverting (second panel of figure 5.2).

The strength of the PLR-BVAR is the ability to push the common trend towards a unit root approximately as intensely as the SZ-BVAR or the naive model, while performing substantially less shrinkage on the consumption- and investment-to-GDP ratios. Therefore, this more sophisticated prior does not outweigh the likelihood information about the mean reversion of the investment ratio, while being consistent with the trending behavior of the consumption ratio. Finally, notice that the PLR-BVAR would also outperform the theory-based predictions of constant ratios, which is particularly at odds with the observed pattern of consumption relative to GDP.

Before turning to the VAR with five variables, we wish to briefly mention another popular prior—the so called dummy-initial-observation (or single-unit-root) prior—used in the existing literature. This elegant prior was designed to remove the bias of the sum-of-coefficients prior against cointegration, while still addressing the issue regarding overfitting of the deterministic component (Sims and Zha, 1998). For completeness, we have experimented with this prior as well, but its marginal impact on the posterior relative to the Minnesota and sum-of-coefficients priors is negligible, as we show in appendix E. Therefore, to save space, we have decided to exclude the dummy-initial-observation prior from the forecasts comparison in the main text.

5.2.2. *5-variable VARs.* We now move to the VARs with five variables. Figures 5.3 and 5.4 plot the MSFEs for various forecasting horizons for the level of all the variables included in the VAR and for the linear combinations of the variables obtained by multiplying the matrix H by the vector y (i.e. the common real trend, the great ratios and hours). Notice that the prediction accuracy of the SZ-BVAR deteriorates for GDP, consumption and investment, relative to the 3-variable case. The PLR-BVAR, instead, continues to forecast well, outperforming the MN-BVAR, SZ-BVAR and the naive model uniformly over variables and horizons. The only exceptions are consumption and the labor share, for which the forecasting accuracy of the PLR-BVAR is comparable to the naive model and the MN-BVAR, respectively.

5.2.3. *7-variable VARs.* Turning to the 7-variable case, figures 5.5 and 5.6 plot the MSFEs for the level of the variables in the VAR and for the linear combinations obtained by multiplying the matrix H by the vector y (i.e. the common trends, the great ratios and the real rate). Although there are cases in which all the BVARs perform similarly, the PLR-BVAR generally improves over the MN-BVAR and SZ-BVAR. The most substantial gains are evident for the nominal block and consumption (and the linear combinations involving these variables).

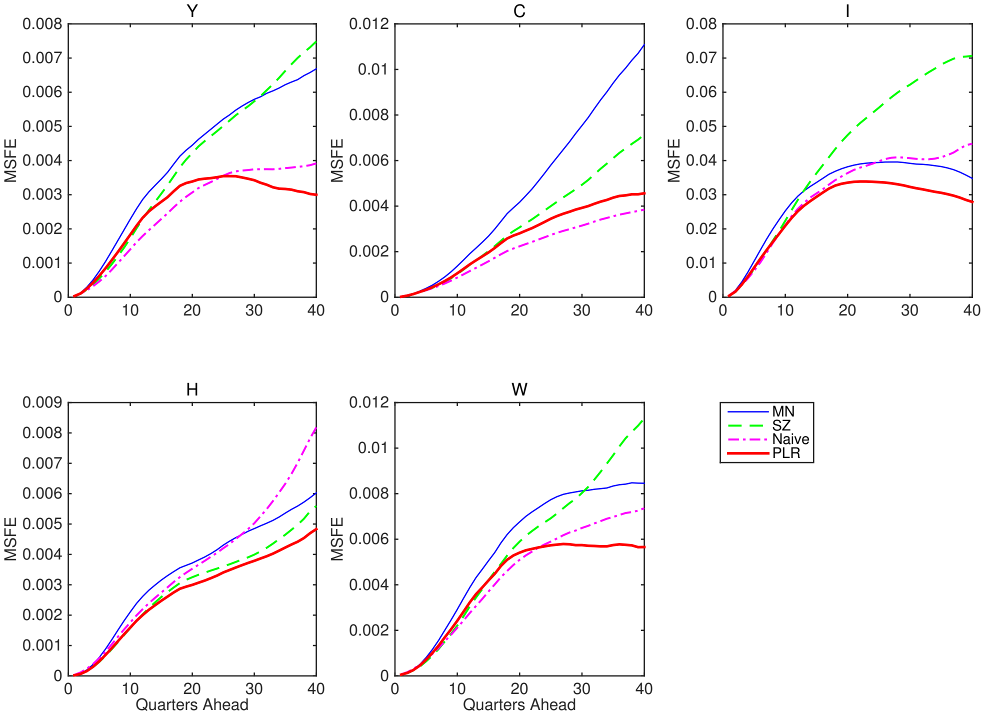

FIGURE 5.3. Mean squared forecast errors in models with five variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run.

The forecasting accuracy of the PLR-BVAR is also generally good relative to the naive model. What is interesting about the 7-variable case, however, is that the performance of *all* the BVARs deteriorates relative to the naive model for output, consumption and wages. Closer inspection reveals that this deterioration is entirely due to the inaccuracy of the BVARs' long-term forecasts produced in the late 1970s. Given the record-high level of inflation, and the historical negative correlation between inflation and real activity, all the VARs estimated in the late 1970s tend to predict a very severe and long-lasting drop in output. In reality, instead, the recession of the early 1980s ended relatively quickly, suggesting the presence of stronger long-term "nominal neutrality" than predicted by the models.[6]

[6] Observe that we have been able to uncover this interesting misbehavior of VARs estimated with nominal variables in the 1970s because of our focus on long-term predictions, which are instead typically neglected by the literature on forecast evaluation.

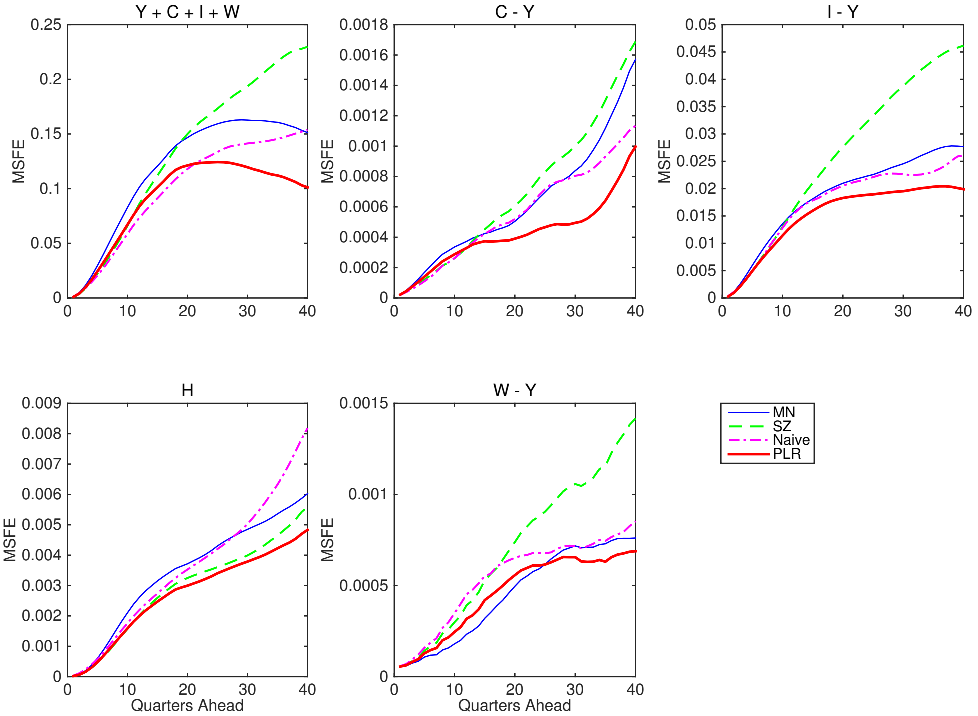

Figure 5.4. Mean squared forecast errors in models with five variables (linear combinations). MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run.

To confirm this view, the dotted lines in figures 5.5 and 5.6 represent the MSFEs produced by a PLR-BVAR with a long-term nominal neutrality restriction. Such a model corresponds to dogmatically setting to zero the hyperparameter ϕ_i controlling the variance of the prior on the column of Λ that captures the effects of the nominal trend. Relative to its unrestricted version, this model generates better MSFEs for the real variables, getting close to MSFEs of the naive model. However, the figures also show a worsening of the forecasting performance for inflation and the nominal trend. The reason of this deterioration is the symmetry of our prior, which does not allow different degrees of shrinkage on different elements of a column of Λ.[7] Therefore, imposing a long-term nominal neutrality restriction on the system comes at the cost of also impairing any effect of the nominal trend on inflation. Our findings suggest

[7] This is a feature of all conjugate priors with a Kronecker structure, including the Minnesota or sum-of-coefficients priors.

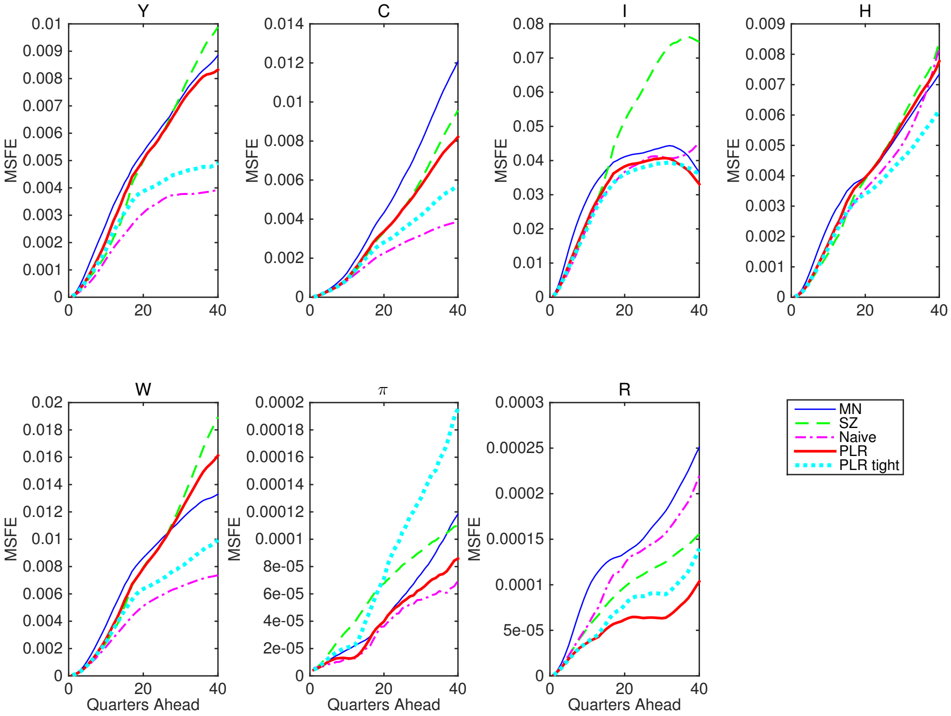

Figure 5.5. Mean squared forecast errors in models with seven variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run; PLR tight: BVAR with the Minnesota prior and the prior for the long run with maximum tightness on the dynamic effect of the common nominal trend.

that breaking this symmetry would be beneficial, although we leave the development of this more involved type of priors for future research.

6. Invariance to Rotations and Other Challenges

In the previous sections, we have discussed the motivation for our PLR, its most attractive features and success in applications. We now also want to highlight the potential limitations of our methodology, and consider extensions that might address some of them.

6.1. **Invariance to rotations.** In this subsection, we discuss the fact that our prior requires the selection of a specific matrix H. We have argued that the rows of H should be

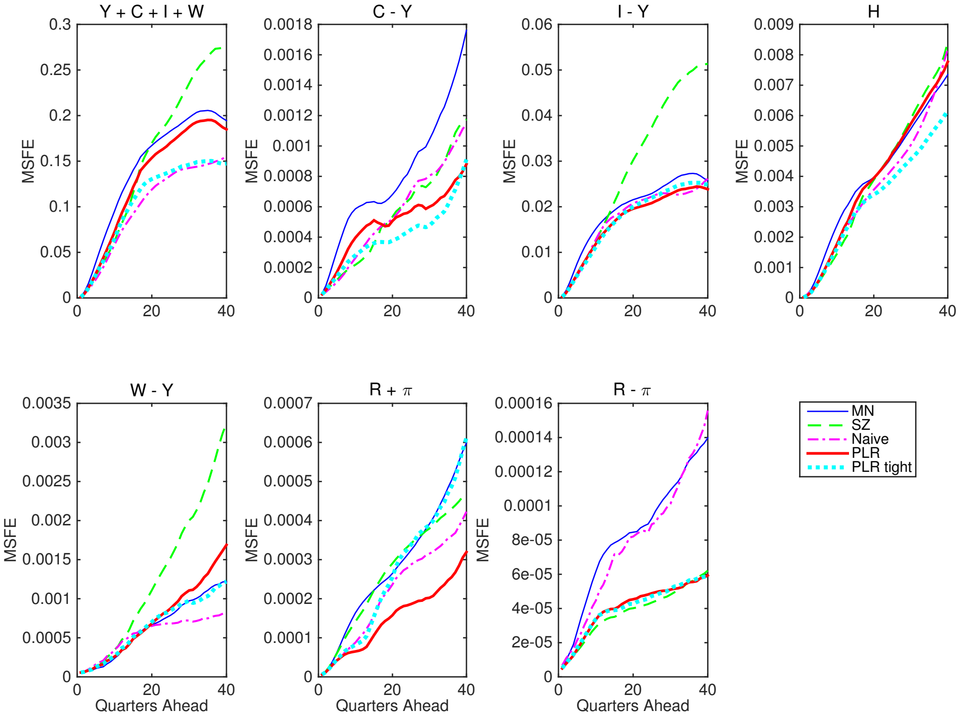

Figure 5.6. Mean squared forecast errors in models with seven variables (linear combinations). MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; Naive: random walk with drift for each variable; PLR: BVAR with the Minnesota prior and the prior for the long run; PLR tight: BVAR with the Minnesota prior and the prior for the long run with maximum tightness on the dynamic effect of the common nominal trend.

chosen to represent linear combinations of y that are likely to exhibit a stochastic trend—denote the coefficients of these combinations by $\beta'_{\perp}$—and stationary deviations from long-run equilibria—call them β'. Notice that economic theory is useful, but not sufficient to uniquely pin down a specific H. The reason is that macroeconomic models are typically informative about $\beta_{\perp}$ and the *space* spanned by β (the cointegrating space), but not about β itself.

For example, in the case of our three variable VAR, theory suggests that GDP, consumption and investment should share a common trend, and that all the linear combinations orthogonal to this trend should be stationary. We have implemented our prior selecting the consumption- and the investment-to-GDP ratios as possibly stationary linear combinations. While this choice might seem natural, it would have been equally valid to pick for instance the consumption-to-investment instead of the investment-to-GDP ratio. The baseline PLR presented in section 3 is not invariant to these rotations of β that, according to theory, are equally likely to generate stationary linear combinations of the variables.

From a theoretical perspective, this lack of invariance might seem unappealing, but it should not be considered as a serious concern, in practice. In fact, most of the gains of our prior derive from separating the common trends from the space of likely stationary combinations, and hence from shrinking more gently the strength of the error-correction mechanisms of the latter. Within this "stationary space," the specific combinations that one selects to implement the prior matter much less. Nevertheless, to fully tackle the issue of invariance, in this section we develop a version of our prior that only depends on the *space* of stationary combinations implied by economic theory.

Without loss of generality, suppose that the first $n-r$ rows of H represent the coefficients of the linear combinations of y that are likely to be nonstationary, while the remaining r rows generate likely stationary combinations of the variables. A modified version of our baseline prior can be implemented using $n-r+1$ dummy observations. The first $n-r$ dummies—the ones used to discipline the dynamic impact of the initial level of the trending combinations of y—are identical to those used to implement our baseline prior:

$$y_{t_i^*} = y_{t_i^*-1} = ... = y_{t_i^*-p} = \frac{H_{i\cdot}\bar{y}_0}{\phi_i}\left[H^{-1}\right]_{\cdot i}, \qquad i = 1, ..., n-r.$$

The last artificial observation takes instead the form

$$y_{t_i^*} = y_{t_i^*-1} = ... = y_{t_i^*-p} = \left[H^{-1}\right]_{\cdot(n-r+1:n)} \frac{H_{(n-r+1:n)\cdot}\bar{y}_0}{\phi_i}, \qquad i = n-r+1, \tag{6.1}$$

where $H_{(n-r+1:n)\cdot}$ denotes the last r rows of H, and $\left[H^{-1}\right]_{\cdot(n-r+1:n)}$ are the last r columns of H^{-1}. In appendix F, we prove that the prior implemented through this set of dummy observations is invariant to rotations of the last r rows of H.

The easiest way to appreciate the differences between the invariant and the baseline prior is to substitute the dummy observation (6.1) into the level-difference representation of the model (3.2), obtaining

$$0 = \underbrace{\Pi\left[H^{-1}\right]_{\cdot(n-r+1:n)}}_{\Lambda_{\cdot(n-r+1:n)}} H_{(n-r+1:n)\cdot}\bar{y}_0 + \phi_{n-r+1}\varepsilon_{t^*_{n-r+1}}$$

or, equivalently,

$$0 = \sum_{j=n-r+1}^{n} \Lambda_{\cdot j} H_{j\cdot} \bar{y}_0 + \phi_{n-r+1} \varepsilon_{t^*_{n-r+1}}. \tag{6.2}$$

This expression makes clear that the prior is effectively limiting the extent to which the *sum* of the linear combinations $H_{j\cdot}y$ helps forecasting Δy at the beginning of the sample. This is different from the baseline PLR, which disciplines the impact of these linear combinations one-at-a-time—see equation (3.9).

In addition to implying a prior that is invariant to certain rotations, the dummy observation in (6.1) can also be combined with a non-zero artificial observation for the VAR exogenous variable. This variable is what implicitly multiplies the constant term in (3.1), although we have omitted it for simplicity so far, since it is equal to 1. If we denote this exogenous variable by z_t, its value for the artificial time period t^*_{n-r+1} can be set to

$$z_{t^*_{n-r+1}} = \frac{1}{\phi_{n-r+1}}, \tag{6.3}$$

in which case the implied prior becomes more elegant because it also disciplines the constant. In fact, by using (6.3) instead of $z_{t^*_{n-r+1}} = 0$, the constant term would appear additively on the right-hand side of (6.2). Therefore, loosely speaking, one can think of the implied prior as shrinking the VAR parameters in one of these two directions: either (i) towards a limited strength of the error correction mechanisms and a small constant term, or (ii) towards stronger error correction mechanisms, but unconditional means of the likely stationary linear combinations of the variables not too distant from their initial observations. Notice that shrinking in *either one* of these directions should reduce the excessive explanatory power of the deterministic component, as explained in section 2.

Observe that the use of a non-zero dummy value for the exogenous variable relates our invariant PLR to the so-called dummy-initial-observation (or single-unit-root) prior of Sims and Zha (1998). The latter, however, "mixes" a-priori trending and stationary linear combinations of the variables, and ends up having a small effect on the estimates when its tightness is selected based on the marginal likelihood, as we show in appendix E. Similarly, notice that it would not be prudent to use this non-zero value of the VAR exogenous variable for the n dummy observations needed to implement the baseline version of the PLR, as they would convey conflicting views about the value of the constant term.[8]

Figures 6.1, 6.2 and 6.3 present the MSFE results produced by the 3-, 5- and 7-variable VARs estimated using this invariant version of the PLR. Compared to the baseline (solid line), the forecasting performance generally worsens when the prior is set up to be invariant with respect to all rotations orthogonal to the common real trend (dotted line). However, this deterioration is entirely due to the treatment of the consumption-to-GDP ratio, which is predicted to be stationary by conventional macroeconomic models, but is clearly trending in the data after 1980 (see figure 5.2). Therefore, the results are negatively affected by the requirement that the prior is invariant to rotations spanned also by this variable. To confirm this view, the dashed lines in figures 6.1, 6.2 and 6.3 present the MSFE when the consumption-to-GDP ratio is excluded from the invariant part of the prior, and treated as a trending variable instead. Observe that, in this case, the MSFEs improve over the baseline uniformly, across models, variables and horizons.

To sum up, the invariant version of the PLR is a joint prior on the autoregressive coefficients and the constant term. It has the potential to deliver substantial gains in forecasting accuracy, but only when the theoretical separation between the trending and stationary spaces of linear combinations is roughly in line with the empirical evidence. Making sure that this is indeed the case might require some "preliminary look at the data," which in practice makes the methodology more akin to an empirical Bayes procedure.

6.2. **(In)variance to the level of the variables.** A second issue to discuss is the fact that our prior may not be invariant to the level of the variables entering the VAR, because its tightness depends on $H\bar{y}_0$. Consider, for instance, the row of H capturing the common trend shared by the real variables of the system. The prior variance implied by this linear combination of $\bar{y}_0$ will depend, for example, on whether GDP and the other real variables are expressed in millions or billions of dollars, or in 2005 or 2009 dollars.

This being said, there are at least two reasons why this problem should have no major practical consequence. First of all, $H\bar{y}_0$ is used to provide only an approximate scaling of the prior tightness, with the hyperparameters ϕ_i's in (3.5) offering additional flexibility. In

[8] An alternative approach to place a prior jointly on the autoregressive coefficients and the constant term is proposed by Villani (2009) or Jarocinski and Marcet (2013), although these methods do not preserve conjugacy and require stationarity or an error-correction representation.

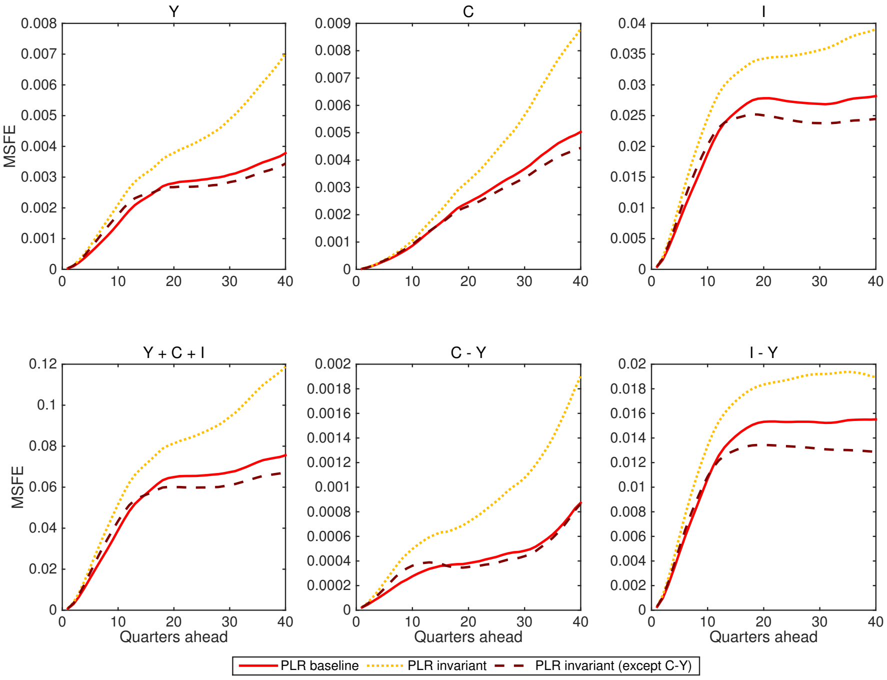

Figure 6.1. Mean squared forecast errors in models with three variables. PLR baseline: BVAR with the Minnesota prior and the baseline prior for the long run; PLR invariant: BVAR with the Minnesota prior and the invariant version of the prior for the long run; PLR invariant (except C-Y): BVAR with the Minnesota prior and the invariant version of the prior for the long run, with the consumption-to-GDP ratio treated as a trending variable.

fact, if we followed an empirical-Bayes methodology to set these hyperparameters without constraints, the lack of invariance problem would entirely disappear. Our fully Bayesian approach, however, involves the use of hyperpriors. While reasonably disperse, these hyperpriors might in practice constrain the allowed range of variation of the hyperparameters.

Second, the lack of invariance problem is not present when variables are expressed in dimensionless units—such as rates—or when a linear combination $H_{i\cdot}\bar{y}_0$ represents the logarithm of a ratio between variables expressed in the same unit—such as the labor share or the consumption- and investment-to-GDP ratios. This is the case for most of the linear combinations that we consider to set up the PLR in our macroeconomic applications. This last point constitutes a substantial advantage relative to the sum-of-coefficients prior, in which the level of the variables always affects the prior variance due to the mechanical choice of H equal to the identity matrix.

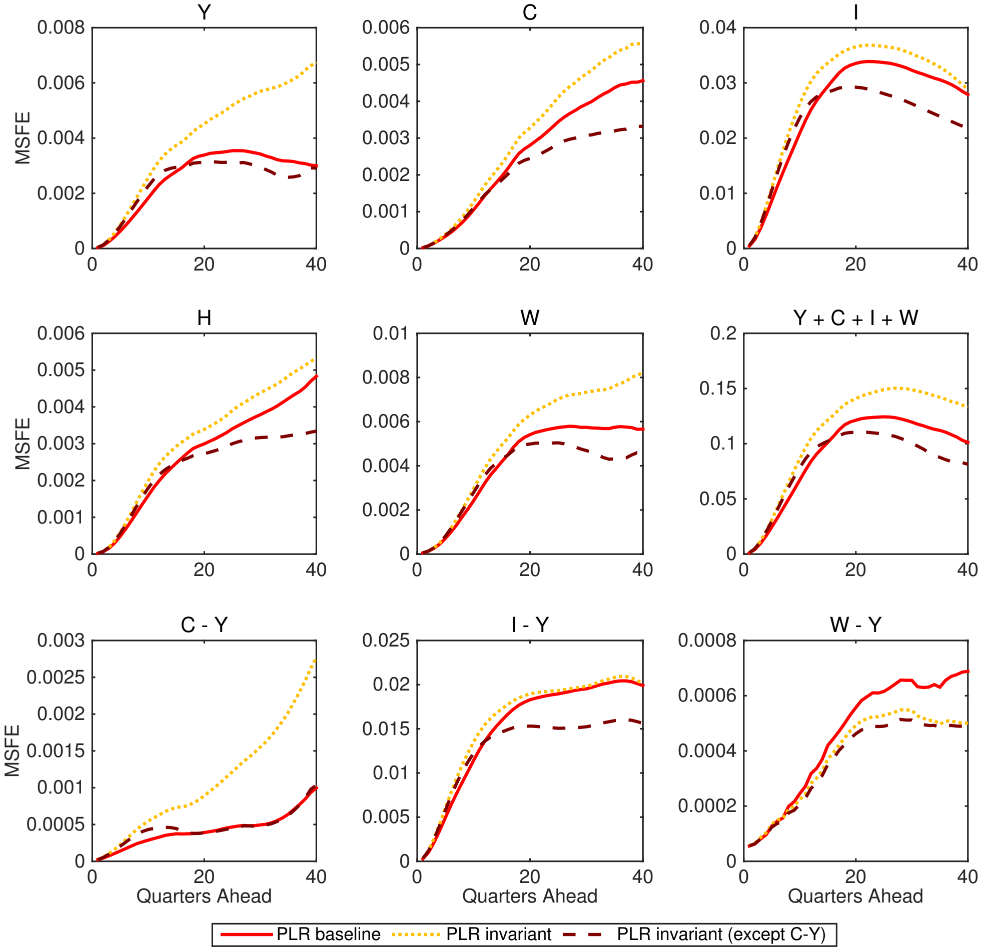

FIGURE 6.2. Mean squared forecast errors in models with five variables. PLR baseline: BVAR with the Minnesota prior and the baseline prior for the long run; PLR invariant: BVAR with the Minnesota prior and the invariant version of the prior for the long run; PLR invariant (except C-Y): BVAR with the Minnesota prior and the invariant version of the prior for the long run, with the consumption-to-GDP ratio treated as a trending variable.

6.3. **Truly predictable trends.** A final issue we wish to mention concerns the possible presence of true deterministic trends in the data. As we have discussed at length, the main purpose of our prior is to reduce the importance of the spurious deterministic components implied by VARs estimated with flat priors. Clearly, if these low-frequency, deterministic trends are a true feature of the data-generating process, a BVAR with our PLR will have a tendency to attribute them to the stochastic component of the model, at least in part.

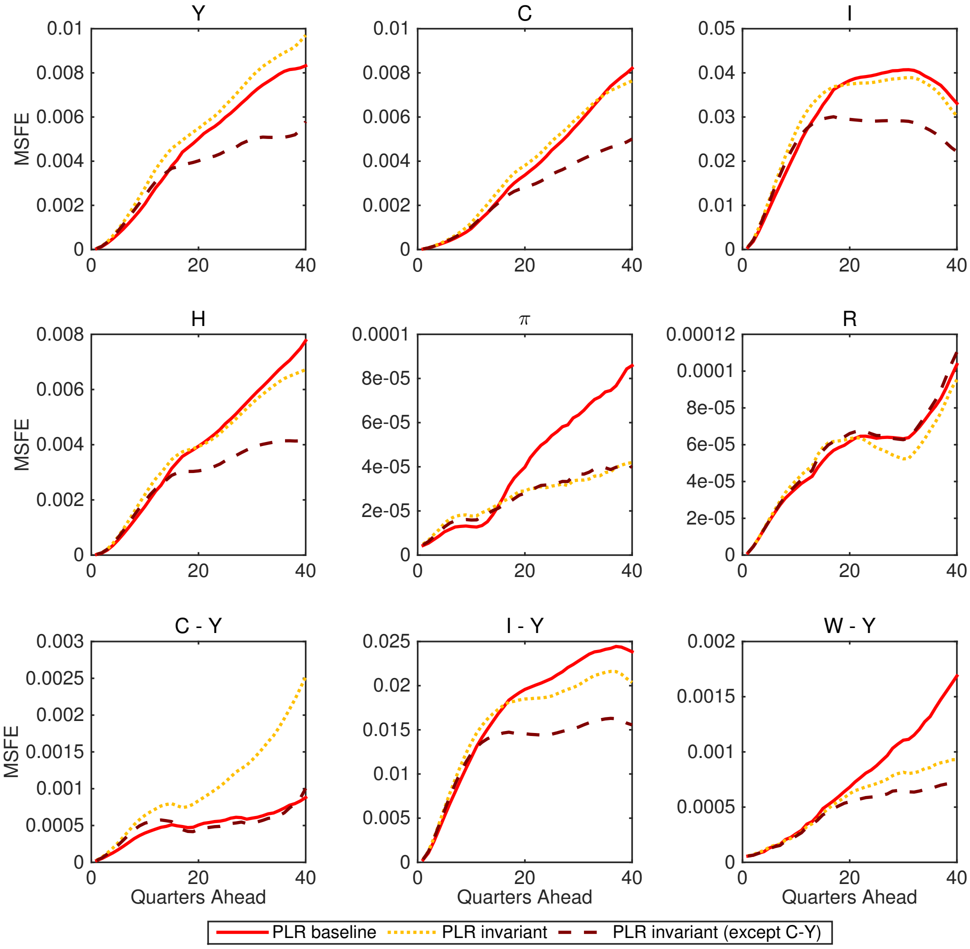

FIGURE 6.3. Mean squared forecast errors in models with seven variables. PLR baseline: BVAR with the Minnesota prior and the baseline prior for the long run; PLR invariant: BVAR with the Minnesota prior and the invariant version of the prior for the long run; PLR invariant (except C-Y): BVAR with the Minnesota prior and the invariant version of the prior for the long run, with the consumption-to-GDP ratio treated as a trending variable. To save space, the figure presents the MSFEs for only a subset of the variables and linear combinations.

Even in this case, however, a flat-prior VAR does not necessarily constitute a valid alternative. In fact, a flat prior would allow more flexibility in the choice of parameter values, to possibly fit smooth, predictable trends using the model-implied deterministic component. However, the attempt to do so would distort the stochastic properties of the system.

7. Concluding Remarks

In this paper, we have introduced a new class of prior distributions for VARs that (i) impose discipline on the long-run behavior of the model; (ii) are based on robust lessons of theoretical macroeconomic models; (iii) can be thought of as a full probabilistic approach to cointegration, and include the error-correction representation of a VAR as a special case; and (iv) perform well in forecasting, especially at long horizons.

While our priors for the long run present a number of appealing features, one potential challenge is that the procedure to set them up is not automated. It requires thinking about the economics behind the determination of the variables included in the model, even if such a model is a reduced-form one that normally does not entail any economic theory. As a consequence, setting up a PLR in large-scale models with dozens of variables may prove difficult. To deal with these situation, it would be interesting to extend the applicability of our priors to cases in which the econometrician is a-priori confident about the trending/stationarity properties of only a subset of the system, or only some combinations of the variables, while remaining agnostic about the rest.

Appendix A. Asymptotic Behavior of the Deterministic Component

In this appendix, we study the case in which the true data-generating process (DGP) is a driftless random walk, and prove that (i) the deterministic component implied by an estimated autoregressive process explains a random fraction of the sample variation of the data, even if the estimation sample is infinitely large; (ii) the long-term forecasts implied by an estimated autoregressive process diverge from the optimal forecast at rate $\sqrt{T}$, inducing an erratic behavior of forecast accuracy measures. We also argue that similar results hold in the case in which the true DGP is a local-to-unity process.

Suppose that the data are generated by the following random walk,

$$y_t = y_{t-1} + \varepsilon_t, \tag{A.1}$$

where ε_t is a martingale difference sequence with variance σ^2 and bounded fourth moment. The deterministic component implied by an AR(1) process estimated using data from time 1 to T is given by

$$\hat{d}_{t,T} \equiv \frac{\hat{c}_T}{1-\hat{\rho}_T} + \hat{\rho}_T^{t-1}\left(y_1 - \frac{\hat{c}_T}{1-\hat{\rho}_T}\right),$$

where $\hat{c}_T$ and $\hat{\rho}_T$ denote the OLS estimates of the constant and autocorrelation coefficient of the AR(1), which depend on the sample size. Observe that this expression of the deterministic component requires $\hat{\rho}_T \neq 1$, which holds with probability one.

We are interested in characterizing the asymptotic behavior of

$$\hat{F}_T = \frac{\sum_{t=1}^T \left(\hat{d}_{t,T} - y_1\right)^2}{\sum_{t=1}^T \left(y_t - y_1\right)^2}, \tag{A.2}$$

which represents the fraction of the total sample variation of y attributed to the estimated deterministic component. Notice that F_T is 0 under the true parameter values, since the deterministic component associated to the true DGP (A.1) is flat and equal to y_1. In what follows, instead, we will show that $\hat{F}_T$ converges to a random variable and not to 0.

Consider first the numerator of (A.2), which can be written as

$$\begin{aligned}
\sum_{t=1}^T \left(\hat{d}_{t,T} - y_1\right)^2 &= \sum_{t=1}^T \left[\left(\frac{\hat{c}_T}{1-\hat{\rho}_T} - y_1\right)\left(1-\hat{\rho}_T^{t-1}\right)\right]^2 \\
&= \left(\frac{\hat{c}_T}{1-\hat{\rho}_T} - y_1\right)^2 \sum_{t=1}^T \left(1 - 2\hat{\rho}_T^{t-1} + \hat{\rho}_T^{2(t-1)}\right) \\
&= \left(\frac{\hat{c}_T}{1-\hat{\rho}_T} - y_1\right)^2 \left(T - 2\frac{\left(1-\hat{\rho}_T^T\right)}{\left(1-\hat{\rho}_T\right)} + \frac{\left(1-\hat{\rho}_T^{2T}\right)}{\left(1-\hat{\rho}_T^2\right)}\right).
\end{aligned}$$

Substitute now the last expression back into (A.2), and divide numerator and denominator by T^2, obtaining

$$\hat{F}_T = \frac{\left(\frac{\sqrt{T}\hat{c}_T}{T(1-\hat{\rho}_T)} - \frac{y_1}{\sqrt{T}}\right)^2 \left(1 - 2\frac{\left(1-\hat{\rho}_T^T\right)}{T(1-\hat{\rho}_T)} + \frac{\left(1-\hat{\rho}_T^{2T}\right)}{T\left(1-\hat{\rho}_T^2\right)}\right)}{\frac{1}{T^2}\sum_{t=1}^T \left(y_t - y_1\right)^2}.$$

Notice that

$$\sqrt{T}\left(\hat{c}_T - y_1\left(1-\hat{\rho}_T\right)\right) \Rightarrow V_1 = \sigma^2 \frac{W(1)\int_0^1 W^2(r)\,dr - \frac{1}{2}\left[W^2(1) - 1\right]\int_0^1 W(r)\,dr}{\int_0^1 W^2(r)\,dr - \left[\int_0^1 W(r)\,dr\right]^2}$$

$$T\left(1-\hat{\rho}_T\right) \Rightarrow V_2 = \frac{\frac{1}{2}\left[W^2(1) - 1\right] - W(1)\int_0^1 W(r)\,dr}{\int_0^1 W^2(r)\,dr - \left[\int_0^1 W(r)\,dr\right]^2}$$

$$T\left(1-\hat{\rho}_T^2\right) \Rightarrow 2V_2$$

$$\hat{\rho}_T^T \Rightarrow e^{-V_2}$$

$$\hat{\rho}_T^{2T} \Rightarrow e^{-2V_2}$$

$$\frac{1}{T^2}\sum_{t=1}^{T}\left(y_t - y_1\right)^2 \Rightarrow V_3 = \sigma^2 \int_0^1 W^2\left(r\right)dr,$$

where the symbol "⇒" denotes convergence in distribution and $W\left(r\right)$ is a Wiener process. These convergence results are standard and can be found, for example, in Hamilton (1994).[9] By the continuous mapping theorem, it follows that

$$\hat{F}_T \Rightarrow \frac{\left(\frac{V_1}{V_2}\right)^2\left(1 - 2\frac{1-e^{-V_2}}{V_2} + \frac{1-e^{-2V_2}}{2V_2}\right)}{V_3}, \tag{A.3}$$

which proves that the share of sample variation explained by the deterministic component does not converge to zero, but to a random quantity. In other words, if the true data-generating process exhibits a very high degree of autocorrelation, estimated AR and VAR models imply a spurious and excessive explanatory power of the deterministic component, even if estimated using an arbitrarily large sample. For example, a Monte Carlo simulation of (A.3) suggests that $\Pr\left(\hat{F}_T > 0.5\right)$ converges to approximately $\frac{2}{3}$ as T goes to infinity.

The logic behind this problematic behavior of the deterministic component also helps to understand the fragility of long-term forecasts in highly persistent processes—an issue also studied by Stock (1996) and Rossi (2005), among others. These forecasts, in fact, not only do not converge to the optimal forecast, but actually diverge at rate $\sqrt{T}$, implying an erratic behavior of long-term forecast accuracy measures. To see this point, suppose once again that the data are generated by (A.1), and that the researcher estimates an AR(1) process by OLS, to construct an h-step-ahead out-of-sample forecast. The deviation between such a forecast, $\hat{y}_{T+h|T}$, and the optimal forecast obtained using the true data-generating process, y_T, is given by

$$\hat{y}_{T+h|T} - y_T = \left(\frac{\hat{c}_T}{1-\hat{\rho}_T} - y_T\right)\left(1 - \hat{\rho}_T^h\right).$$

As in Stock (1996), define as "long-term" a forecast with an horizon that is a sizable fraction of the sample size, i.e. $h = [\lambda T]$, where $[\cdot]$ denotes the the largest smaller integer function. Observe that the previous expression can be rewritten as

$$\frac{\hat{y}_{T+h|T} - y_T}{\sqrt{T}} = \left(\frac{\sqrt{T}\hat{c}_T}{T\left(1-\hat{\rho}_T\right)} - \frac{y_T}{\sqrt{T}}\right)\left(1 - \hat{\rho}_T^{[\lambda T]}\right).$$

[9] Observe that $\left(\hat{c}_T - y_1\left(1-\hat{\rho}_T\right)\right)$ is nothing else but the estimate of the constant term obtained using y_t in deviation from y_1.

Given the convergence results established above, and since $\hat{\rho}_T^{[\lambda T]} \Rightarrow e^{-\lambda V_2}$, it is easy to show that $\frac{\hat{y}_{T+h|T} - y_T}{\sqrt{T}} \Rightarrow \left(\frac{V_1}{V_2}\right)\left(1 - e^{-\lambda V_2}\right)$ and thus that $\hat{y}_{T+h|T}$ diverges at rate $\sqrt{T}$ from the optimal forecast y_T. This result is in sharp contrast with the stationary case, in which long-term forecasts converge to optimal forecasts—the unconditional mean of the process—at rate $\sqrt{T}$.

Finally, since all the rates of convergence above are the same if the DGP is local-to-unity (with a possibly $O\left(T^{1/2}\right)$ constant, see Stock and Watson, 1996 or Rossi, 2005), it follows that $\hat{F}_T$ and $\frac{\hat{y}_{T+h|T} - y_T}{\sqrt{T}}$ converge to random quantities also in that case.

Appendix B. Posterior Distributions

In this appendix we describe the posterior distribution of the VAR coefficients under the various prior densities that we have used in the paper. Except for the derivations related to the PLR, the other results are standard, and we report them only to make the paper self-contained.

Consider the VAR model of section 3

$$y_t = c + B_1 y_{t-1} + .. + B_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim \text{i.i.d. } N\left(0, \Sigma\right),$$

and rewrite it as

$$Y = X\beta + \epsilon$$

$$\epsilon \sim N\left(0, \Sigma \otimes I_{T-p}\right),$$

where $y \equiv [y_{p+1}, ..., y_T]'$, $Y \equiv vec\left(y\right)$, $x_t \equiv \left[1, y_{t-1}', ..., y_{t-p}'\right]'$, $X_t \equiv I_n \otimes x_t'$, $x \equiv [x_{p+1}, ..., x_T]'$, $X \equiv I_n \otimes x$, $\varepsilon \equiv [\varepsilon_{p+1}, ..., \varepsilon_T]'$, $\epsilon \equiv vec\left(\varepsilon\right)$, $B \equiv [C, B_1, ..., B_p]'$ and $\beta \equiv vec(B)$. Finally, define the number of regressors for each equation by $k \equiv np + 1$.

B.1. **Flat prior.**

With a flat prior, the posterior of (β, Σ) belongs to the usual Normal-Inverse-Wishart family:

$$\Sigma | Y \sim IW\left(\hat{\varepsilon}_{ols}' \hat{\varepsilon}_{ols}, T - p - n - k - 1\right)$$

$$\beta | \Sigma, Y \sim N\left(\hat{\beta}_{ols}, \Sigma \otimes \left(x'x\right)^{-1}\right),$$

where $\hat{B}_{ols} \equiv \left(x'x\right)^{-1}\left(x'y\right)$, $\hat{\beta}_{ols} \equiv vec\left(\hat{B}_{ols}\right)$, $\hat{\varepsilon}_{ols} \equiv y - x\hat{B}_{ols}$.

B.2. **Minnesota prior.** The so-called Minnesota prior, first introduced in Litterman (1979), is centered on the assumption that each variable follows a random walk process, possibly with drift. More precisely, this prior is characterized by the following first and second moments:

$$E\left[\left(B_s\right)_{ij}|\Sigma\right] = \begin{cases} 1 & \text{if } i=j \text{ and } s=1 \\ 0 & \text{otherwise} \end{cases}$$

$$cov\left(\left(B_s\right)_{ij},\left(B_r\right)_{hm}|\Sigma\right) = \begin{cases} \lambda^2 \frac{1}{s^2}\frac{\Sigma_{ih}}{\hat{\sigma}_j^2/(d-n-1)} & \text{if } m=j\text{and } r=s \\ 0 & \text{otherwise} \end{cases},$$

where the hyperparameter λ controls the overall tightness of this prior and, as customary, the $\hat{\sigma}_j^2$'s are set equal to the residual variance of an AR(1) estimated using the available data for variable j. To obtain a proper prior, we also specify a standard Inverse-Wishart prior on Σ, as in Kadiyala and Karlsson (1997):

$$\Sigma|Y \sim IW\left(diag\left(\left[\hat{\sigma}_1^2,...,\hat{\sigma}_n^2\right]\right), n+2\right).$$

Through an appropriate choice of Ψ, d, b and Ω, such a prior can be easily cast into the Normal-Inverse-Wishart form

$$\Sigma \sim IW\left(\Psi; d\right)$$

$$\beta|\Sigma, \sim N\left(b, \Sigma\otimes\Omega\right),$$

and leads to the following posterior distribution for the VAR coefficients

$$\Sigma|Y \sim IW\left(\Psi+\hat{\varepsilon}'\hat{\varepsilon}+\left(\hat{B}-\flat\right)'\Omega^{-1}\left(\hat{B}-\flat\right), T-p+d\right)$$

$$\beta|\Sigma, Y, \sim N\left(\hat{\beta}, \Sigma\otimes\left(x'x+\Omega^{-1}\right)^{-1}\right),$$

where $\hat{B}\equiv\left(x'x+\Omega^{-1}\right)^{-1}\left(x'y+\Omega^{-1}\flat\right)$, $\hat{\beta}\equiv vec\left(\hat{B}\right)$, $\hat{\varepsilon}\equiv y-x\hat{B}$, and $\flat$ is a $k\times n$ matrix obtained by reshaping the vector b in such a way that each column corresponds to the prior mean of the coefficients of each equation (i.e. $b\equiv vec\left(\flat\right)$).

B.3. **Prior for the long run.** In the main text, we have stated that the implementation of the PLR requires the following set of dummy observations for the artificial times $t_1^*,...,t_n^*$:

$$y_{t_i^*}=y_{t_i^*-1}=...=y_{t_i^*-p}=\frac{H_{i\cdot}\bar{y}_0}{\phi_i}\left[H^{-1}\right]_{\cdot i}, \qquad i=1,...,n. \tag{B.1}$$

The left- and right-hand side of these artificial observations can be collected into the matrices

$$\underset{n\times n}{y^{+}} = diag\left(\left[\frac{H_{1\cdot}\bar{y}_0}{\phi_1}, ..., \frac{H_{n\cdot}\bar{y}_0}{\phi_n}\right]\right)\left[H^{-1}\right]'$$

$$\underset{n\times(1+np)}{x^{+}} = \left[\underset{n\times 1}{0}, y^{+}, ..., y^{+}\right],$$

which can then be added on top of the data matrices y and x to conduct inference as if they were part of the actual sample.

To prove that these dummy observations imply the density in (3.6), substitute the observations in (B.1) into the level-difference representation of the VAR (3.2), obtaining

$$\Pi\frac{H_{i\cdot}\bar{y}_0}{\phi_i}\left[H^{-1}\right]_{\cdot i} = -\varepsilon_{t_i^*}, \qquad i = 1, ..., n.$$

Grouping the columns on the left- and right-hand side of this expression for each i, we obtain

$$\left[\Pi\frac{H_{1\cdot}\bar{y}_0}{\phi_1}\left[H^{-1}\right]_{\cdot 1}, ..., \Pi\frac{H_{n\cdot}\bar{y}_0}{\phi_n}\left[H^{-1}\right]_{\cdot n}\right] = -\left[\varepsilon_{t_1^*}, ..., \varepsilon_{t_n^*}\right],$$

which can be rewritten as

$$\Pi H^{-1} diag\left(\left[\frac{H_{1\cdot}\bar{y}_0}{\phi_1}, ..., \frac{H_{n\cdot}\bar{y}_0}{\phi_n}\right]\right) = -\left[\varepsilon_{t_1^*}, ..., \varepsilon_{t_n^*}\right].$$

Post-multiplying both sides by $diag\left(\left[\frac{\phi_1}{H_{1\cdot}\bar{y}_0}, ..., \frac{\phi_n}{H_{n\cdot}\bar{y}_0}\right]\right)$, recalling that $\Lambda = \Pi H^{-1}$, and applying the vec operator to both sides, we obtain

$$vec\left(\Lambda\right)|H, \Sigma \sim N\left(0, diag\left(\left[\frac{\phi_1^2}{\left(H_{1\cdot}\bar{y}_0\right)^2}, ..., \frac{\phi_n^2}{\left(H_{n\cdot}\bar{y}_0\right)^2}\right]\right)\otimes\Sigma\right),$$

which corresponds to the expression in (3.6).

B.4. **Sum-of-coefficients prior.** The SOC prior of Doan et al. (1984) and Sims and Zha (1998) corresponds to a special case of the PLR, with a mechanical choice of $H = I_n$, and hyperparameters $\phi_1 = ... = \phi_n = \mu$.

Appendix C. Setting of the Hyperparameters

In this appendix we briefly describe the setting of the hyperparameters to generate our empirical results.

We have performed the exercise of section 2 and 5.1 about the shape of the deterministic component using some reference hyperparameter values. In particular, the hyperparameter

λ controlling the tightness of the Minnesota prior has been set to 0.2, which is standard. For the hyperparameters $\{\phi_i\}_{i=1}^n$ of the PLR, we have chosen a value equal to 1, which corresponds to using a single set of dummy observations, with error variance approximately similar to error variance in the actual sample.

For the forecasting exercise of section 5.2, we have adopted a hierarchical interpretation of the model as in Giannone et al. (2015), and set the hyperparameters by maximizing their posterior. The posterior of the hyperparameters is given by the product of the marginal likelihood and the hyperpriors (the prior density on the hyperparameters). Given that our priors are conjugate, the marginal likelihood is available in closed form (see Giannone et al., 2015, and the derivations in their appendix). As priors for the hyperparameters λ and μ, we have chosen Gamma densities with mode equal to 0.2 and 1—the values recommended by Sims and Zha (1998)—and standard deviations equal to 0.4 and 1 respectively, as in Giannone et al. (2015). For the hyperparameters of the PLR, $\{\phi_i\}_{i=1}^n$, we have also used Gamma densities with mode and standard deviation equal to 1. As argued by Giannone et al. (2015), an appealing feature of non-flat hyperpriors is that they help stabilize inference when the marginal likelihood happens to have little curvature with respect to some hyperparameters.

Appendix D. Data

This appendix describes the data series used for the estimation of the 3-, 5- and 7-variable VARs. The source of most of our data is the FRED dataset, available on the website of the Federal Reserve Bank of St. Louis. The sample ranges from 1955Q1 to 2013Q1. The variables entering the 3-variable VARs correspond to the following definitions (series acronym in parenthesis):

- log-real GDP per capita:

$$Y = log\left[\frac{\text{Gross Domestic Product (GDP)}}{\text{Population} \cdot \text{GDP Implicit Price Deflator (GDPDEF)}}\right]$$

- log-real consumption per capita:

$$C = log\left[\frac{\text{Personal Consumption Expenditure: Nondurable Goods (PCND)} + \text{Services (PCESV)}}{\text{Population} \cdot \text{GDP Implicit Price Deflator (GDPDEF)}}\right]$$

- log-real investment per capita:

$$I = log\left[\frac{\text{Gross Private Domestic Investment (GPDI)} + \text{Personal Consumption Expenditures: Durable Goods (PCDG)}}{\text{Population} \cdot \text{GDP Implicit Price Deflator (GDPDEF)}}\right],$$

where the population series used to compute the quantities per capita is the Hodrick-Prescott trend (estimated with smoothing parameter equal to 1600) of the logarithm of the Civilian Noninstitutional Population (CNP16OV) series. The reason to use this smooth population series is to avoid the spikes in the original series that correspond to the census years. The series of GDP, PCND, PCESV, GDPDI and PCDG are in current dollars, while GDPDEF is a chain-type price index that is equal to 100 in 2009.

The 5-variable VARs also includes:

- log-hours per capita:

$$H = log\left[\frac{\text{Total Economy: Hours of All Persons}}{\text{Population } \cdot \text{ 2080}}\right]$$

- log-real wages:

$$W = log\left[\frac{\text{Total Economy: Compensation of Employees (W209RC1Q027SBEA)}}{\text{Population} \cdot \text{GDP Implicit Price Deflator (GDPDEF)}}\right],$$

where the series of hours worked comes from the Total U.S. Economy Hours & Employment data file, available on the Bureau of Labor Statistics website at www.bls.gov/lpc/special_requests/us_total_hrs and 2080 is a scale factor representing a reference number of hours worked by a person in a year (obtained by multiplying the 52 number of weeks by 40).

The 7-variable VARs also includes:

- inflation:

$$\pi = \Delta log\left[\text{GDP Implicit Price Deflator (GDPDEF)}\right]$$

- short-term nominal interest rate:

$$R = \frac{\text{Effective Federal Funds Rate (FEDFUNDS)}}{400}.$$

Appendix E. Comparison with the Dummy-Initial-Observation Prior

In this appendix, we evaluate the accuracy of the forecasts obtained when we include a dummy-initial-observation prior in the 3-, 5- and 7-variable VARs. This prior was designed to avoid the bias against cointegration of the sum-of-coefficients prior, while still reducing the explanatory power of the deterministic component of the model (see Sims and Zha, 1998 for the details of its implementation). In the existing literature, it is often combined with the Minnesota and sum-of-coefficients priors (see, for example, Sims and Zha, 1998 or Giannone et al., 2015).

Figure E.1, E.2 and E.3 compare the forecasting performance (in terms of MSFEs) of the 3-, 5- and 7-variable MN- and SZ-BVARs *without* (as in the main text of the paper) and *with* the dummy-initial-observation prior. As usual, all hyperparameters are selected by maximizing their posterior. These figures make clear that the marginal contribution of the dummy-initial-observation prior is negligible, and the forecasting results of the MN+DIO- and SZ+DIO-BVARs are nearly identical to those of the MN- and SZ-BVARs reported in the main text of the paper.

Appendix F. Proof of the Invariance Result

The purpose of this appendix is to prove that the prior implied by the dummy observation (6.1) in section 6 is invariant to rotations of the linear combinations of y that should be stationary according to economic theory.

Without loss of generality, suppose that the first $n-r$ rows of H represent the coefficients of the likely nonstationary linear combinations of y, while the remaining r rows contain the coefficients of the likely stationary combinations of the variables. To prove the invariance property, we need to show that the dummy observation (6.1) only depends on the space spanned by the last r rows of H. In other words, we need to demonstrate that (6.1) is invariant to pre-multiplications of H by a block diagonal matrix of the form

$$R = \begin{bmatrix} I_{n-r} & 0 \\ 0 & Q \end{bmatrix},$$

where Q is a generic $r \times r$ invertible matrix.

To this end, define $\tilde{H} = RH$, whose last r rows are linear combinations of the last r rows of H, i.e. $\tilde{H}_{(n-r+1:n)\cdot} = QH_{(n-r+1:n)\cdot}$ Notice that $\tilde{H}^{-1} = H^{-1}R^{-1}$, implying

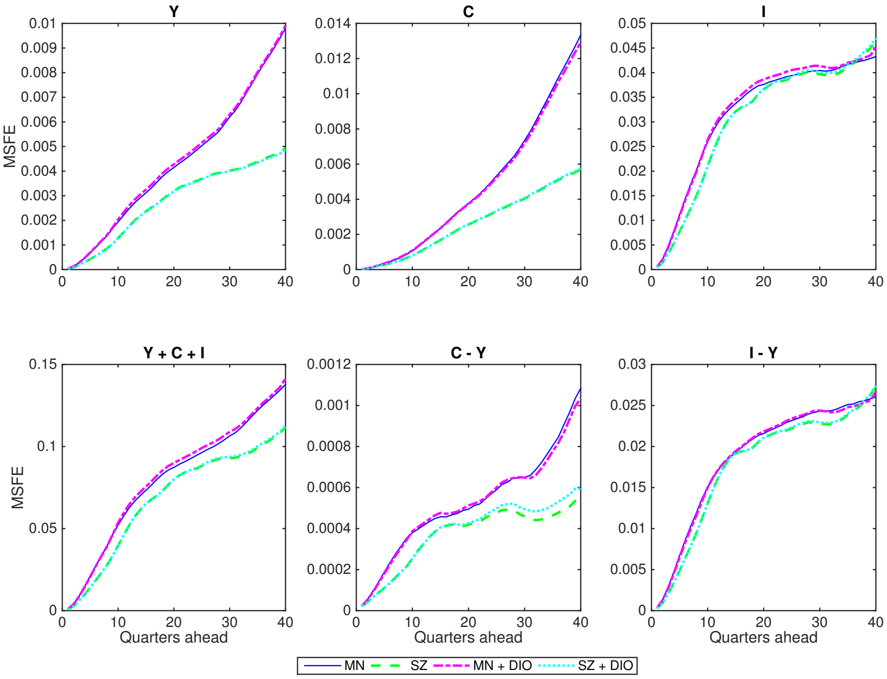

Figure E.1. Mean squared forecast errors in models with three variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; MN+DIO: BVAR with Minnesota and dummy-initial-observation priors; SZ+DIO: BVAR with the Minnesota, sum-of-coefficients and dummy-initial-observation priors.

that the last r columns of $\tilde{H}^{-1}$ are linear combinations of the last r columns of H^{-1}, i.e. $\left[\tilde{H}^{-1}\right]_{\cdot(n-r+1:n)} = \left[H^{-1}\right]_{\cdot(n-r+1:n)} Q^{-1}$. Using $\tilde{H}$ instead of H in (6.1) yields

$$\left[\tilde{H}^{-1}\right]_{\cdot(n-r+1:n)} \frac{\tilde{H}_{(n-r+1:n)\cdot}\bar{y}_0}{\phi_i} = \left[H^{-1}\right]_{\cdot(n-r+1:n)} \underbrace{Q^{-1}Q}_{I_r} \frac{H_{(n-r+1:n)\cdot}\bar{y}_0}{\phi_i},$$

which does not depend on Q, proving that the prior only depends on the space spanned by the last r rows of H.

References

Altig, D., L. Christiano, M. Eichenbaum, and J. Linde (2011): "Firm-Specific Capital, Nominal Rigidities and the Business Cycle," *Review of Economic Dynamics*, 14, 225–247.

Atkeson, A. and L. E. Ohanian. (2001): "Are Phillips curves useful for forecasting inflation?" *Quarterly Review*, 2–11.

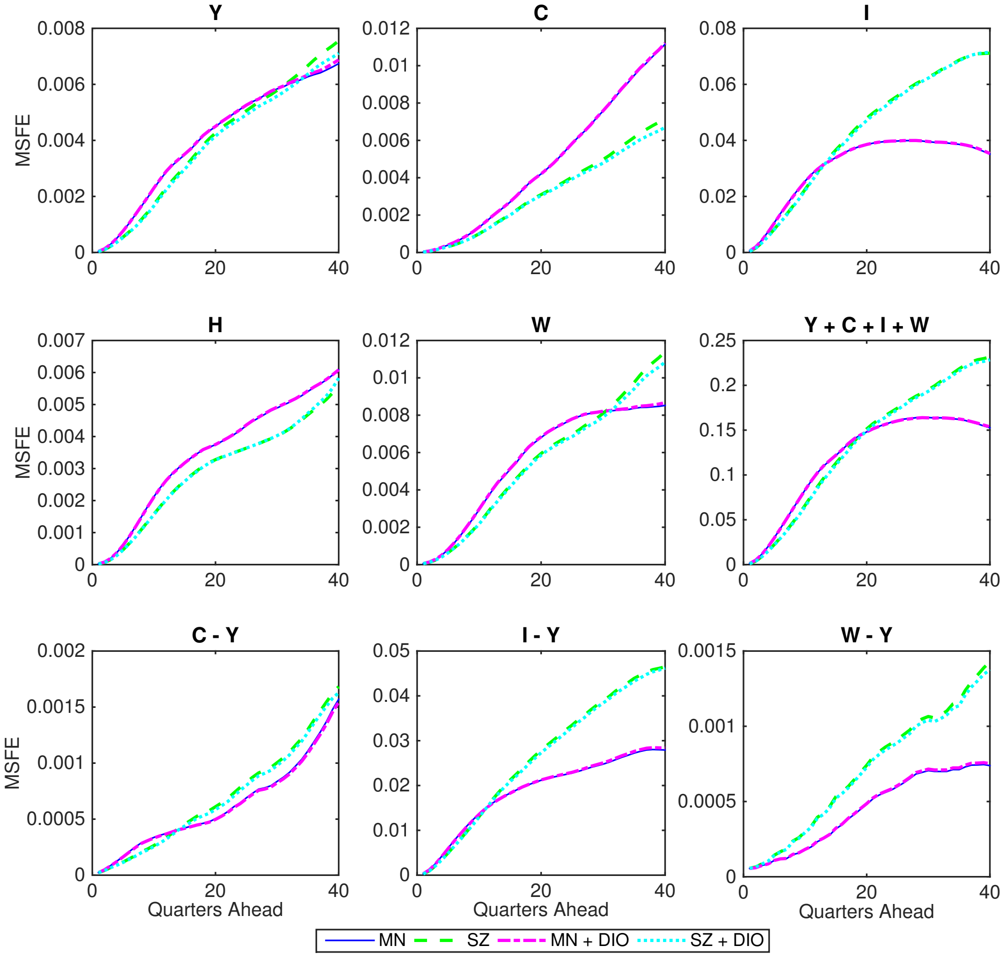

FIGURE E.2. Mean squared forecast errors in models with five variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; MN+DIO: BVAR with Minnesota and dummy-initial-observation priors; SZ+DIO: BVAR with the Minnesota, sum-of-coefficients and dummy-initial-observation priors.

BAUWENS, L. AND M. LUBRANO (1996): "Identification restrictions and posterior densities in cointegrated Gaussian VAR systems," in *Advances in Econometrics*, ed. by T. Fomby, JAI Press, vol. 11B, 3–28.

CHAO, J. C. AND P. C. B. PHILLIPS (1999): "Model selection in partially nonstationary vector autoregressive processes with reduced rank structure," *Journal of Econometrics*, 91, 227–271.

CHRISTOFFERSEN, P. F. AND F. X. DIEBOLD (1998): "Cointegration and Long-Horizon Forecasting," *Journal of Business & Economic Statistics*, 16, 450–58.

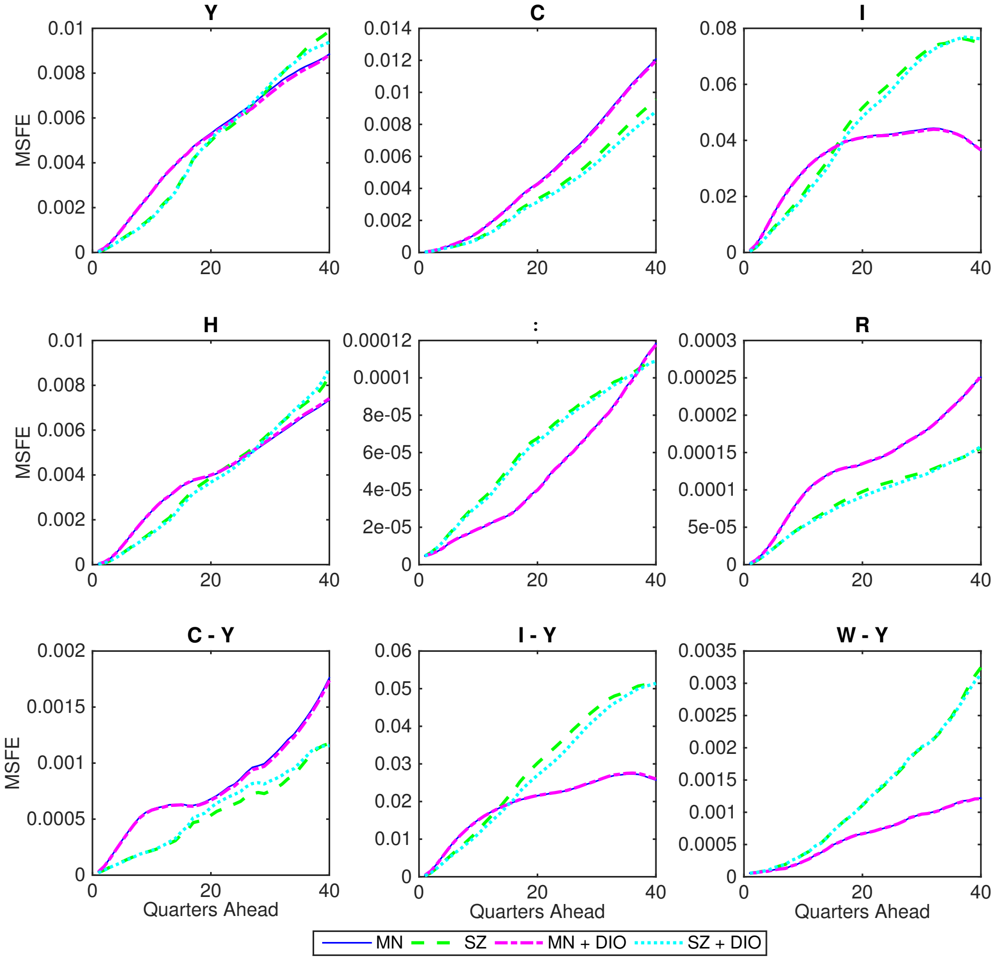

FIGURE E.3. Mean squared forecast errors in models with seven variables. MN: BVAR with the Minnesota prior; SZ: BVAR with the Minnesota and sum-of-coefficients priors; MN+DIO: BVAR with Minnesota and dummy-initial-observation priors; SZ+DIO: BVAR with the Minnesota, sum-of-coefficients and dummy-initial-observation priors. To save space, the figure presents the MSFEs for only a subset of the variables and linear combinations.

CORANDER, J. AND M. VILLANI (2004): "Bayesian assessment of dimensionality in reduced rank regression," *Statistica Neerlandica*, 58, 255–270.

D'AGOSTINO, A., D. GIANNONE, AND P. SURICO (2007): "(Un)Predictability and Macroeconomic Stability," CEPR Discussion Papers 6594, C.E.P.R. Discussion Papers.

DEL NEGRO, M. AND F. SCHORFHEIDE (2004): "Priors from General Equilibrium Models for VARS," *International Economic Review*, 45, 643–673.

——— (2011): "Bayesian Macroeconometrics," in *The Oxford Handbook of Bayesian Econometrics*, ed. by G. K. J. Geweke and H. van Dijk, Oxford University Press, vol. 1, chap. 7, 293–389.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007): "On the Fit of New Keynesian Models," *Journal of Business & Economic Statistics*, 25, 123–143.

Doan, T., R. Litterman, and C. A. Sims (1984): "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100.

Elliott, G. (1998): "On the Robustness of Cointegration Methods when Regressors Almost Have Unit Roots," *Econometrica*, 66, 149–158.

Engle, R. F. and C. W. J. Granger (1987): "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251–76.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004): *Bayesian Data Analysis: Second Edition*, Boca Raton: Chapman and Hall CRC.

Geweke, J. (1996): "Bayesian reduced rank regression in econometrics," *Journal of Econometrics*, 75, 121–146.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior Selection for Vector Autoregressions," *The Review of Economics and Statistics*, 97, 436–451.

Hamilton, J. D. (1994): *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

Horvath, M. T. and M. W. Watson (1995): "Testing for Cointegration When Some of the Cointegrating Vectors are Prespecified," *Econometric Theory*, 11, 984–1014.

Ireland, P. N. (2007): "Changes in the Federal Reserve's Inflation Target: Causes and Consequences," *Journal of Money, Credit and Banking*, 39, 1851–1882.

Jarocinski, M. and A. Marcet (2011): "Autoregressions in Small Samples, Priors about Observables and Initial Conditions," CEP Discussion Papers 1061, Centre for Economic Performance, LSE.

——— (2013): "Priors about Observables in Vector Autoregressions," UFAE and IAE Working Papers 929.13, Unitat de Fonaments de l'Analisi Economica (UAB) and Institut d'Analisi Economica (CSIC).

——— (2015): "Contrasting Bayesian and Frequentist Approaches to Autoregressions: the Role of the Initial Condition," Working Papers 776, Barcelona Graduate School of Economics.

Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press.

Justiniano, A., G. E. Primiceri, and A. Tambalotti (2010): "Investment Shocks and Business Cycles," *Journal of Monetary Economics*, 57, 132–145.

KADIYALA, K. R. AND S. KARLSSON (1997): "Numerical Methods for Estimation and Inference in Bayesian VAR-Models," *Journal of Applied Econometrics*, 12, 99–132.

KARLSSON, S. (2013): "Forecasting with Bayesian Vector Autoregressions," in *Handbook of Economic Forecasting*, ed. by G. Ellit and A. Timmermann, Elsevier, vol. 2, chap. 15, 791–897.

KING, R. G., C. I. PLOSSER, J. H. STOCK, AND M. W. WATSON (1991): "Stochastic Trends and Economic Fluctuations," *American Economic Review*, 81, 819–40.

KLEIBERGEN, F. AND R. PAAP (2002): "Priors, posteriors and bayes factors for a Bayesian analysis of cointegration," *Journal of Econometrics*, 111, 223–249.

KLEIBERGEN, F. AND H. K. VAN DIJK (1994): "On the Shape of the Likelihood/Posterior in Cointegration Models," *Econometric Theory*, 10, 514–551.

KOOP, G., R. STRACHAN, H. VAN DIJK, AND M. VILLANI (2006): "Bayesian approaches to cointegration," in *The Palgrave Handbook of Theoretical Econometrics*, ed. by T. C. Mills and K. Patterson, Palgrave McMillan, vol. 1, chap. 25, 871–898.

LITTERMAN, R. B. (1979): "Techniques of forecasting using vector autoregressions," Working Papers 115, Federal Reserve Bank of Minneapolis.

MUELLER, U. K. AND G. ELLIOTT (2003): "Tests for Unit Roots and the Initial Condition," *Econometrica*, 71, 1269–1286.

MUELLER, U. K. AND M. W. WATSON (2008): "Testing Models of Low-Frequency Variability," *Econometrica*, 76, 979–1016.

PHILLIPS, P. C. B. (1991a): "Bayesian Routes and Unit Roots: De Rebus Prioribus Semper Est Disputandum," *Journal of Applied Econometrics*, 6, 435–73.

——— (1991b): "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics*, 6, 333–64.

ROSSI, B. (2005): "Testing Long-Horizon Predictive Ability With High Persistence, And The Meese-Rogoff Puzzle," *International Economic Review*, 46, 61–92.

ROSSI, B. AND T. SEKHPOSYAN (2010): "Have economic models' forecasting performance for US output growth and inflation changed over time, and when?" *International Journal of Forecasting*, 26, 808–835.

SIMS, C. A. (1996): "Inference For Multivariate Time Series Models With Trend," Princeton University, mimeo.

——— (2000): "Using a likelihood perspective to sharpen econometric discourse: Three examples," *Journal of Econometrics*, 95, 443–462.

——— (2003): "Probability Models for Monetary Policy Decisions," Princeton University, mimeo.

Sims, C. A. and T. Zha (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39, 949–68.

Smets, F. and R. Wouters (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

Stock, J. (2010): "Cointegration in Theory and Practice: A Tribute to Clive Granger," ASSA Meetings Presentation Slides.

Stock, J. H. (1996): "VAR, Error Correction and Pretest Forecasts at Long Horizons," *Oxford Bulletin of Economics and Statistics*, 58, 685–701.

Stock, J. H. and M. W. Watson (1996): "Confidence sets in regressions with highly serially correlated regressors," Harvard University, mimeo.

——— (2007): "Why Has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking*, 39, 3–33.

Strachan, R. W. and B. Inder (2004): "Bayesian analysis of the error correction model," *Journal of Econometrics*, 123, 307–325.

Strachan, R. W. and H. K. van Dijk (2005): "Valuing Structure, Model Uncertainty and Model Averaging in Vector Autoregressive Process," Money Macro and Finance (MMF) Research Group Conference 2005 30, Money Macro and Finance Research Group.

Uhlig, H. (1994a): "On Jeffreys Prior when Using the Exact Likelihood Function," *Econometric Theory*, 10, 633–644.

——— (1994b): "What Macroeconomists Should Know about Unit Roots: A Bayesian Perspective," *Econometric Theory*, 10, 645–671.

Villani, M. (2000): "Aspects of Bayesian cointegration," Ph.D. thesis, Stockholm University.

——— (2001): "Bayesian prediction with cointegrated vector autoregressions," *International Journal of Forecasting*, 17, 585–605.

——— (2005): "Bayesian Reference Analysis Of Cointegration," *Econometric Theory*, 21, 326–357.

——— (2009): "Steady-state priors for vector autoregressions," *Journal of Applied Econometrics*, 24, 630–650.

Watson, M. W. (1986): "Vector Autoregressions and Cointegration," in *Handbook of Econometrics*, ed. by R. F. Engle and D. McFadden, Elsevier, vol. 4, chap. 47, 2843–2915.

Federal Reserve Bank of New York and CEPR

E-mail address: dgiannon2@gmail.com

European Central Bank and ECARES

E-mail address: michele.lenza@ecb.int

Northwestern University, CEPR and NBER

E-mail address: g-primiceri@northwestern.edu