Robust open Bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models

Antonio Pacifico

To cite this article: Antonio Pacifico (2021) Robust open Bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models, *Econometric Reviews*, 40:2, 148-176, DOI: [10.1080/07474938.2020.1770996](https://doi.org/10.1080/07474938.2020.1770996)

To link to this article: <https://doi.org/10.1080/07474938.2020.1770996>

Published online: 05 Jun 2020.

Submit your article to this journal [↗](#)

Article views: 119

View related articles [↗](#)

View Crossmark data [↗](#)

Robust open Bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models

Antonio Pacifico

Postdoc in Applied Statistics & Econometrics, LUISS Guido Carli University, CEFOP-LUISS, Rome, Italy

ABSTRACT

The paper develops a computational method to deal with some open issues related to Bayesian model averaging for multiple linear models: overfitting, model uncertainty, endogeneity issues, and misspecified dynamics. The methodology takes the name of Robust Open Bayesian procedure. It is *robust* because the Bayesian inference is performed with a set of priors rather than a single prior and *open* because the model class is not fully known in advance, but rather is defined iteratively by MCMC algorithm. Conjugate informative priors are used to compute exact posterior probabilities. Empirical and simulated examples describe the functioning and performance of the procedure. Discussions with related works are also accounted for.

KEYWORDS

Bayesian model averaging; endogeneity issues; MCMC integrations; misspecified dynamics; multiple linear regression

JEL CLASSIFICATION

A1; C01; E02; H3; N01; O4

1. Introduction

Given a set of all possible candidate covariates $X_1, \dots, X_m$ and a dependent variable Y , standard statistical procedures to select subsets of covariates when building multiple linear regression models consist in conditioning on a single *best* model and then to make inferences as if the selected model has been the true model (see, e.g., Miller (1984), Breiman (1992, 1995), and Breiman and Spector (1992)). Hereafter, *best* stands for the model providing the most accurate predictive performance overall candidate models. For instance, consider the following model:

$$Y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i \quad (1)$$

where, say $\beta_3 = 0$. Model selection should exclude x_3 because it will not improve prediction. A similar argument goes when $\beta_3 \simeq 0$ is very small. However, if x_1 and x_2 are sufficiently correlated, estimating both β_1 and β_2 in one model is unlikely to represent the best strategy since it could result in larger estimation variance and lower prediction accuracy. More precisely, this approach would lead to underestimation of uncertainty in model selection and thus to over-confident inferences and decisions about quantities of interest.

A practical Bayesian solution to these problems involves estimating two separate single-regressor models, the first based on x_1 and the next based on x_2 , and averaging over them instead. Even though each model is biased (omitted variables), it also has smaller variance can be further reduced by averaging. Each member of this sequence is called a submodel and the number of variables in the equation is the size of the submodel. The *best* submodel of the generated sequence is generally selected in order to give higher prediction accuracy. This is known as Bayesian Model Averaging (BMA) and is widely used today as the basic methodology of accounting for model uncertainty (see, e.g., Madigan and Raftery (1994), Madigan et al. (1995), Raftery

et al. (1995, 1997), Clyde (1999), Clyde and George (2004), and Hansen (2007)). To better understand that procedure and thus the implementation proposed in this study, given the model (1), the posterior distribution for the quantity of interest Y is:

$$pr(Y|D) = \sum_{k=1}^m pr(Y|M_k, D) \cdot pr(M_k|D) \quad (2)$$

where D denotes the data.

Equation (2) is the average of the posterior distributions under each of $M_1, \dots, M_m$ models considered, weighted by their posterior model probability. Nevertheless, in many applications, this averaging becomes unpractical, especially when the number of terms in model (1) tends to increase, rendering exhaustive summation unfeasible. Several empirical evidence and MCMC methods—discussed in the paper—have been proposed in order to select promising subsets in multiple linear regression models and estimate their associated model-averaged posterior distributions dealing with the problem of accounting for model uncertainty.

The contribution of this article is to propose and develop a computational approach for implementing Bayesian model selection in order to deal with some open research questions related to BMA and MCMC methods. They include: (i) model uncertainty in hierarchical normal mixture models (Section 2.1); (ii) overfitting and thus overestimation of effect sizes since more complex models will always provide a somewhat better fit to the data than simpler models, where the “complexity” stands (for example) for the number of unknown parameters (Section 2.2); (iii) the choice and specification of prior distributions over multiple competing models (Sections 2.3 and 2.4); (iv) endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems, when studying and modeling economic interactions (Section 3); and (v) variable selection problems in implementing BMA methodology for multiple models classes when the number of parameters and thus the number of possible models is high dimensional (Section 5).

The methodology proposed in this article takes the name of Robust Open Bayesian (ROB) procedure. It is *robust* because the Bayesian inference is performed with a set of priors rather than a single prior (see, for instance, Berger (1990) and Berger et al. (1994)) and *open* since the model class is not fully known in advance, but rather is determined and defined iteratively by MCMC routine (see, for instance, Madigan et al. (1995) and George and McCulloch (1993, 1997)). It consists of two stages. First, a fully enumerated Markov Chain Monte Carlo (MC^F) integration is used to move through the model space and the parameter space at the same time in order to obtain a reduced set containing *best* (possible) model solutions (or *best* combination of predictors) that mainly explain and thus fit the data. Second, a further shrinkage is conducted in order to obtain a smallest final subset of *top best* submodels containing the only *significant* solutions. Hereafter, *top best* stands for the model providing the most accurate predictive performance over all candidate submodels obtained in the first stage and *significant* stands for models having statistically significant predictive capability. Finally, the regression model with higher Bayes Factor (BF) will be the final solution containing a potential subset of predictors with higher *significant* overall F value and sufficiently *strong* adjusted- R^2 ($\bar{R}^2$) measure. Hereafter, *strong* refers to $\bar{R}^2$ value equal to or bigger than 30%.

My approach and empirical application aim build on Raftery et al. (1997), who develop two alternative approaches to account for model uncertainty in linear regression models: Occam’s window algorithm, that involves constructing a small set of models over which a model average can be computed, and MCMC model composition (MC^3) methodology of Madigan et al. (1995), that directly approximates the exact solution. The former consists of two basic principles, likewise the ROB procedure. First, if a model predicts the data far less well than the model that provides the *best* combination of predictors, then it has effectively been discredited and should no longer be considered. Second, models receiving less support from the data than any of their simpler

submodels should be excluded. However, the methodology proposed in this paper differs from their analyses in three main issues: (i) both model uncertainty and overfitting are accounted for; (ii) a hierarchical structural framework is provided to assign multiple class of prior distributions; and (iii) MCMC-based Posterior Model Probabilities (PMPs) derive from the number of iteration counts for all possible regressor combinations, without eliminating any submodels from consideration and thus acting as a strong Occam's razor.

I also build on the Stochastic Search Variable Selection (SSVS) method of George and McCulloch (1993). It defines a Markov chain that moves through model space and parameter space at the same time. The SSVS method never actually removes a predictor from the full model, but only sets it closes to zero with high probability. My procedure avoids this by integrating analytically over parameter space.

The main thrust to the potential of ROB procedure is the use of Conjugate Informative Proper (CIP) priors for each promising model in a hierarchical and structural framework. This is an important issue because, in multiple model class, common parameters can change meaning from one model to another, so that prior distributions must change in a corresponding fashion and weight more according to model size. Thus, each possible model solution being considered is likely to be exactly true. In this way, when the true model is not in the model class and there is no relationship between potential predictors, there is no need to introduce some penalty term (e.g., when allowing for approximate Bayesian model selection) or restrictions on data-supported models (e.g., in Occam's window when there is no signal in the data¹). Thus, the hierarchical conjugate formulation can be used to jointly address the problem of outlier detection and variable selection.

When studying and modeling economic interactions and policy implications in a context of large (if not big) dynamic data, endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems, can occur. More precisely, since ROB procedure is strictly related to BMA approach, it is constrained to model associations between covariates and the dependent variable, ignoring issues of endogeneity and uncertainty surrounding instrument specifications. Moreover, recent studies suggest that there are large potential gains from jointly modeling and quantifying multicountry data using the information contained in a large set of endogenous variables instead of single univariate time-series models (see, for instance, Canova and Ciccarelli (2009), Canova et al. (2007), Ciccarelli et al. (2018), and Pacifico (2019a, 2019b)). Thus, I extend and implement the ROB procedure in order to deal with endogeneity issues and functional form misspecification, when accounting for dynamics of the economy in either time-invariant moderate data ($m \leq 15$) or time-varying high dimensional multicountry data ($m \geq 20$), with m denoting the number of predictors (X). The former is obtained by building on and extending the framework of Lenkoski et al. (2014), who combine the BMA approach to regression variable uncertainty and the Two-Stage Least Squares (TSLS) procedure. More precisely, I derive CIP priors for all the endogenous predictors (X) in order to model uncertainty and investigate endogeneity issues at both the instrument and covariate levels in the ROB procedure. Then, the validity of instruments in the TSLS-ROB procedure is addressed by constructing a Bayesian test of the identification restrictions based on model averaged posterior predictive p-values. In a context of dynamic multicountry setups, I build on Pacifico (2019b) and estimate a Structural Panel Bayesian Vector Autoregressive (SPBVAR) model by defining hierarchical prior specification strategy and MC^F implementations in order to extend variable selection procedure to a wide array of candidate models and thus candidate density forecast combinations. Bayesian methods are used to reduce the dimensionality of the model, structure the time variations, and evaluate issues of endogeneity. In this way, policies designed to

¹The Occam's window usually reports the null model or the simplest model solution including the null model as the only one to be selected. See, for instance, Raftery et al. (1997).

protect the economy against the worst-case consequences of misspecified dynamics are less aggressive and good approximations of the estimated rule. Moreover, some drawbacks of robust policies, a connection between the degree of structure an authority assigns to the uncertainty it faces, and the likelihood of policy attenuation in response to that uncertainty can be assessed.

The outline of this paper is as follows. [Section 2](#) discusses the econometric methodology describing Bayesian framework and then the procedure, and illustrating as conjugate hierarchical setups and MCMC algorithms are constructed. [Section 3](#) extends the methodology to investigate endogeneity issues in time-invariant moderate data. [Section 4](#) illustrates empirical and simulated examples by comparing the predictive performance of ROB procedure with related works. [Section 5](#) extends the methodology to jointly deal with endogeneity issues and misspecified dynamics in time-varying multicountry setups. Finally, [Section 6](#) contains some concluding remarks.

2 Econometric methodology

2.1. Bayesian framework

The aim in building multiple regression models is to predict an outcome Y_i for each individual i , with $i = 1, 2, \dots, n$, based on the behavior of a few or large number of continuous and/or discrete covariates X_{ik} , with $k = 1, 2, \dots, m$. Given a vector of regression parameters $\theta = \{\theta_k\}$ of size m , the response Y_i is modeled as a linear combination of the explanatory variables X_{ik} :

$$Y_i = \alpha + \sum_{k=1}^m \theta_k X_{ik} + \epsilon_i \quad \text{or} \quad f(Y|\theta, \sigma) = N(X\theta, \sigma^2 I_N) \quad (3)$$

where Y_i is a $n \cdot 1$ vector, $X = X_1, X_2, \dots, X_m$ is a $[n \cdot (m+1)]$ matrix, $\theta = (\theta_1, \theta_2, \dots, \theta_m)'$ is a $k \cdot 1$ vector of unknown regression coefficients, α is the intercept, and $\epsilon_i \sim N(0, \sigma^2)$ is a $n \cdot 1$ vector of disturbances, with σ to be an unknown positive scalar.

Stacking the model (3) for n observations:

$$Y = \alpha + \sum_{k=1}^m \theta_k X_k + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma^2) \quad (4)$$

Let $\{M_k, k \in \mathcal{K}, M_k \in \mathcal{M}\}$ be a countable collection of candidate models, where M_k contains the vector of the unknown parameters θ , $\{\Theta_k, \theta_k \in \Theta_k, \Theta_k \in \Theta\}$ be the set of all possible values for the parameters of model M_k , and $f(M_k)$ be the prior probability of model M_k , then the PMP is given:

$$f(M_k|Y) = \frac{f(M_k) \cdot f(Y|M_k)}{\sum_{M_k \in \mathcal{M}} f(M_k) \cdot f(Y|M_k)} \quad \text{with} \quad M_k \in \mathcal{M} \quad (5)$$

where $f(Y|M_k)$ is the marginal likelihood corresponding to $f(Y|M_k) = \int f(Y|M_k, \theta_k) \cdot f(\theta_k|M_k) d\theta_k$ and $f(\theta_k|M_k)$ is the conditional prior distribution of θ_k . The natural parameter space and model space for (M_k, θ_k) are, respectively:

$$\Theta = \bigcup_{M_k \in \mathcal{M}} \{M_k\} \times \Theta_k \quad (6)$$

$$\mathcal{M} = \bigcup_{k \in \mathcal{K}} \{k\} \times M_k \quad (7)$$

When the size of the set of possible model solutions $\mathcal{M}$ is high dimensional, the calculation of the integral $f(Y|M_k)$ becomes unfeasible. Thus, a MCMC method is required in order to generate observations from the joint posterior distribution $f(M_k, \theta_k|Y)$ of (M_k, θ_k) for estimating $f(M_k|Y)$ and $f(\theta_k|M_k, Y)$. In this context, no perfect collinearity is assumed to hold for each candidate model class (M_k) .

The main aim of BMA is to average over all possible solution models $M_k \in \mathcal{M}$ in order to choose the *best* model providing the most accurate predictive performance of the form:

$$Y = \delta + \sum_{l=1}^{\varpi} \theta_l X_l + \eta \quad \text{with} \quad \eta \sim N(0, \sigma^2) \quad (8)$$

where $X_1, X_2, \dots, X_{\varpi}$ is a subset of $X_1, X_2, \dots, X_m$, $l = 1, 2, \dots, \varpi$ is a subparameter index with $l \ll k$. The error terms in Eqs. (4) and (8) are $(n \cdot 1)$ vectors independent and identically distributed and uncorrelated between them. Thus, $(\epsilon, \eta) \sim i.i.d.$ and $Cov(\epsilon, \eta) = E(\epsilon' \cdot \eta) = 0$.

In hierarchical models, many problems involve multiple parameters which can be regarded as related in some way by the structure of the problem. A joint probability model for those parameters should reflect their mutual dependence. Typically, the dependence can be summarized by viewing these parameters as a sample from a common population distribution. Thus, the problem can be modeled hierarchically, with observable outcomes (Y_i) created conditionally on certain parameters (θ_k), which themselves are assigned a distribution in terms of further (possibly common) parameters, hyperparameters. In addition, common parameters would change meaning from one model to another, so that prior distributions must change in a corresponding fashion. This hierarchical thinking may help solve the trade-off between inaccurate fit and overfitting, and it plays an important role in developing computational strategies.

2.2. Robust open Bayesian procedure

The variable selection problem arises when there is some unknown subset of predictors X_{ik} with regression coefficients so small that it would be preferable to ignore them. Thus, the variable selection procedure can be seen as one of deciding which of the regression parameters, the θ_k 's, are equal to zero. Throughout this paper, I use two auxiliary indicator variables. One corresponds to a vector $\chi = \{\chi_k\}$, with $\chi = (\chi_1, \chi_2, \dots, \chi_m)'$, containing each of the possible 2^m subset choices, where $\chi_k = 0$ if θ_k is small (absence of k -th covariate in the model) and $\chi_k = 1$ if θ_k is sufficiently large (presence of k -th covariate in the model). The other auxiliary variable used to implement variable selection is a vector $\beta = \{\beta_k\}$, with $\beta = (\beta_1, \beta_2, \dots, \beta_m)'$, corresponding to the regression parameter θ_k when $I_k = 1$ (e.g., by defining $\theta_k = I_k \beta_k$). When $I_k = 0$, the variable β_k would be sufficiently small so that the predictor X_k should be ruled out from the procedure.

Given the model in Eq. (4), the full model class set is:

$$\mathcal{F} = \{M_k : M_k \subset \mathcal{F}, M_k \in \mathcal{M}, k \in \mathcal{K}, \alpha + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_m X_{nm} + \epsilon\} \quad (9)$$

where $\mathcal{M} = [\{k\} \cdot M_k]$ is the natural model space and X_k denotes the regressors contained in the data for each i .

The ROB procedure consists in two stages. First, a fully enumerated Markov Chain Monte Carlo (MCC^F) integration² is used to move through both model space and parameter space at the same time in order to obtain a reduced set containing *best* model solutions (or *best* combination of predictors) that mainly explain and thus fit the data. Second, a further shrinkage is conducted in order to obtain a smallest final subset of *top best* submodels containing the only *significant* solutions. Finally, the regression model with higher Bayes Factor (BF) will be the final solution containing a potential subset of predictors with higher *significant* overall F value and sufficiently strong $\bar{R}^2$ measure.

²See Section 2.4 for more details.

2.2.1. First stage

The first stage of the ROB procedure entails a joint reduction of model space and parameter space, by approximating the Eq. (4) via MC^F integration. This means that the posterior $f(M_k|Y)$ would put more weight on the more promising subsets of predictors. This would correspond to a mixed strategy, where the model selection is based on a bridge between an objective and a subjective proposal to move from the full model (M_k) to the promising submodels (M_l) and thus followed by acceptance or rejection of it. Here, the model sampler corresponds to the total number of models to be sampled, starting from 0. In this way, all regression coefficients will be averaged over all candidate models, including the models wherein the variable has not been contained (that is, the regression coefficient is zero). In addition, since the model space and the parameter space are fully enumerated at the same time, the procedure is able to iterate all possible combination of predictors with respect to the full model (M_k). It is similar in spirit to the Stochastic Search Variable Selection method of George and McCulloch (1993), with a main difference. The SSVS method never removes a predictor from the full model, but only sets it close to zero with high probability. The ROB procedure avoids this by excluding all regressor combinations that predict the data far less well than the submodels that provide the *best* combination of predictors. The same problem has been avoided from Raftery et al. (1997), who adopt the MCMC model composition (MC^3) methodology of Madigan et al. (1995), but it directly approximates the exact solution and thus neglects the problem of accounting for the trade-off between inaccurate fit and overfitting in high dimensional linear regression models.

Let the reduced parameter space and model space be, respectively:

$$\dot{\Theta} = \bigcup_{M_l \in \dot{\mathcal{M}}} \{M_l\} \times \Theta_l \quad (10)$$

$$\dot{\mathcal{M}} = \bigcup_{l \in \mathcal{K}} \{l\} \times M_l \quad (11)$$

with $\Theta_l \subset \Theta$ and $M_l \subset \mathcal{M}$, the subset containing the *best* model solutions will be:

$$\mathcal{S} = \{M_l : M_l \subset \mathcal{S}, \mathcal{S} \in \mathcal{F}, \Theta_l \subset \Theta_k, \dot{\mathcal{M}} \subset \mathcal{M}, \sum_{l=1}^{\varpi} \pi(M_l|Y_n = y_n, \chi) \geq \tau\} \quad (12)$$

where M_l denotes the submodels of the multiple linear regression $Y_n = \delta + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_{\varpi} X_{n\varpi} + \eta$, with $M_l < M_k$, $l \ll k$, $\{1 \leq l < k\}$, X_l denotes the potential candidate covariates contained in the data, with $X_l < X_k$, and τ is a threshold chosen by the data analyst. The subset in Eq. (12) holds if and only if $\exists \beta_l \neq 0$. In the empirical example,³ I use $\tau = 1\%$ for an enough posterior consistency⁴ with predictors ≤ 15 . Thus, the predictors X_l that seem to be not good predictors (that is, their regression parameters β_l lie below the threshold in one of the selected models) will be excluded from the procedure, with $\chi_k = 0$; otherwise, X_l will be included in the subset $\mathcal{S}$, with $\chi_k = 1$ and $\beta_l = \theta_l$.

Finally, the conditional regression parameters and the PIPs, corresponding to the sum of the PMPs, will be, respectively:

$$E(\beta|Y_n = y_n, \chi) = \sum_{l=1}^{\varpi} \pi(\beta_l|y_n, \chi, M_l) \cdot \pi(M_l|y_n, \chi) \quad (13)$$

$$\sum_{l=1}^{\varpi} \pi(M_l|Y_n = y_n, \chi) = \frac{\pi(y_n|M_l, \chi) \cdot \pi(M_l)}{\pi(y_n|\chi)} = \frac{\pi(y_n|M_l, \chi) \cdot \pi(M_l)}{\sum_{k=1}^K \pi(y_n|M_k, \chi) \cdot \pi(M_k)} \quad (14)$$

³See Section 4.

⁴In Bayesian analysis, this term is usually taken to mean that the posterior probability concentrates on the true model.

2.2.2. Second stage

The second stage of the ROB procedure entails shrinking further the subset $\mathcal{S}$ in order to obtain a smallest final subset of *top best* submodels (M_ξ) containing the only *significant* solutions. The shrinkage does not account for choosing the model with higher PMP (e.g., in standard Bayesian model selection), but matching a (possible) pool of submodels in order to jointly deal with over-estimation of effect sizes (or individual contributions) and model uncertainty (implicit in the procedure).

Here, two specifications are in order: (i) I define a new auxiliary variable $\dot{\chi}$ containing the only *best* model solutions in the subset $\mathcal{S}$ and (ii) I define a new threshold $\hat{\tau}$ according to the probability of the submodels in $\mathcal{S}$ to perform the data (PMPs). In the empirical example,⁵ I use $\hat{\tau} = 0.5\%$ for a sufficient prediction accuracy in explaining the data. Thus, the new and final subset will be:

$$\mathcal{E} = \{M_\xi : M_\xi \subset \mathcal{E}, \mathcal{E} \in \mathcal{S}, \pi(M_l|Y_n = y_n, \dot{\chi}) \geq \hat{\tau}\} \quad (15)$$

where $M_\xi \ll M_l$ and $\pi(M_l|Y_n = y_n, \dot{\chi})$ denotes the PMPs. Finally, the exact and final solution will correspond to one of the submodels M_ξ with higher log natural Bayes Factor (IBF):

$$IBF_{\xi,l} = \log \left\{ \frac{\pi(M_\xi|Y_n = y_n)}{\pi(M_l|Y_n = y_n)} \right\} \quad (16)$$

where $\xi \leq l$. In this procedure, the IBF would also be called the log Weighted Likelihood Ratio (IWLR) factor of M_ξ to M_l with the priors being the weighting functions. The scale of evidence⁶ for interpreting the IWLR factor in Eq. (16) is defined as:

$$\begin{cases} 0 < IB_{\xi,l} < 2 & \text{no evidence for submodel } M_\xi \\ 2 < IB_{\xi,l} < 6 & \text{moderate evidence for submodel } M_\xi \\ 6 < IB_{\xi,l} < 10 & \text{strong evidence for submodel } M_\xi \\ IB_{\xi,l} > 10 & \text{very strong evidence for submodel } M_\xi \end{cases} \quad (17)$$

The prediction of the outcome in Eq. (8) will be:

$$\hat{\pi}(Y) = \sum_{\xi} \hat{\pi}(Y) \cdot \pi(M_\xi|y_n) \quad (18)$$

2.3. Conjugate prior setups and posterior distributions

The variable selection procedure entails estimating χ_k and β_k . Thus, the probability that a variable is *in* the model (e.g., the Posterior Probability) can be simply calculated as the mean value of the indicator χ_k . Here, the size of the χ -th subset is denoted as $m_\chi = \chi'1$. Since the appropriate value of χ is unknown, one could model the uncertainty underlying variable selection by a mixture prior:

$$\pi(\beta, \sigma, \chi) = \pi(\beta|\sigma, \chi) \cdot \pi(\sigma|\chi) \cdot \pi(\chi) \quad (19)$$

The χ -th subset model is described by modeling β as a realization from a multivariate normal prior:

$$\pi(\beta|\sigma, \chi) \equiv \underbrace{\pi(\beta|\chi)}_{\text{conjugate}} = N_k(\mu_\chi, \Sigma_{(\sigma, \chi)}) \quad (20)$$

where μ_χ is a hyperparameter for the auxiliary regression coefficients β_k , according to the non-0 components of χ , restricted to a benchmark prior⁷ $\max(N, |\chi|)$, with $|\chi|$ denoting the model size and $\Sigma_{(\sigma, \chi)} = \Gamma_\chi \cdot V_k$ denoting the $[(k+1) \cdot (k+1)]$ covariance matrix.

⁵See Section 4.

⁶See, for instance, Kass and Raftery (1995).

⁷It would be a generalization of the standard objective prior of Zellner (1986).

The hierarchical mixture prior in Eq. (20) corresponds to the conjugate prior since the conditional distribution of β and σ , given χ , is conjugate for the Eq. (3). In this way, β and σ can be eliminated by routine integration from the full posterior $\pi(\beta, \sigma, \chi|Y)$. This feature yields attractive computational methods for posterior evaluation and exploration.

In the ROB procedure, the covariance matrix is obtained from two components⁸: $V_k = (\sigma^2 \cdot I_n)$, where σ^2 denotes the unknown residual variance for the χ -th model, and the diagonal matrix $\Gamma_\chi = \text{diag}(I, a_{\chi_1}^2 \Sigma'_{X_1 X_1}, a_{\chi_2}^2 \Sigma'_{X_2 X_2}, \dots, a_{\chi_m}^2 \Sigma'_{X_m X_m})$, where $a_{\chi_k}^2$ is a hyperparameter equal to 0 whether $\chi_k = 0$ and equal to 1 otherwise, and $\Sigma_{X_k X_k} = \frac{1}{2} (X_k' X_k)^{-1}$.

The k -th diagonal element of Γ_χ is appropriately set to be small or large according to whether $\chi_k = 0$ or $\chi_k = 1$, respectively. It is defined as:

$$(\Gamma_\chi)_{kk} = \begin{cases} \gamma_{0, \chi_k} & \text{when } \chi_k = 0 \\ \gamma_{1, \chi_k} & \text{when } \chi_k = 1 \end{cases} \quad (21)$$

The residual variance σ^2 is still modeled as a realization from an Inverse Gamma (IG) prior:

$$\pi(\sigma^2|\chi) = IG\left(\frac{\nu}{2}, \frac{\nu \cdot \lambda_\chi}{2}\right) \quad \text{equivalent to} \quad \frac{\nu \cdot \lambda_\chi}{\sigma^2} \sim \chi_\nu^2 \quad (22)$$

where ν and λ_χ are hyperparameters to be chosen in order to decrease with the size of the selected subset m_χ . Here, λ_χ and ν would correspond to the prior estimate of σ^2 and the prior sample size associated with this estimate, respectively. Because there is only a partial information on the distribution of the covariates, the variable selection procedure will choose $\lambda_\chi = s_{LS}^2$ and $\nu = s_Y^2$ so that $\pi(\sigma^2|\chi)$ would assign substantial probability to the interval (s_{LS}^2, s_Y^2) , where s_{LS}^2 and s_Y^2 are the classical least squares estimate of σ^2 based on the saturated model and the sample variance of Y , respectively.

Finally, under the prior defined in Eq. (20), each component of β is still modeled as coming from a scale mixture of normals:

$$\pi(\beta_k|\sigma, \chi) = (1 - \chi_k) \cdot N(\mu_\chi, \sigma^2 \gamma_{0, \chi_k}) + \chi_k \cdot N(\mu_\chi, \sigma^2 \gamma_{1, \chi_k}) \quad (23)$$

As with the nonconjugate prior (see, e.g., George and McCulloch (1993, 1997)), the idea is that γ_{0, χ_k} and γ_{1, χ_k} are sets small and large respectively, so that when the data supports $\chi_k = 0$ over $\chi_k = 1$, then β_k is probably small enough so that χ will not be needed in the model. However, the way in which γ_{0, χ_k} and γ_{1, χ_k} are selected is affected by the unknown value of σ . Generally, when a reasonable estimate s^2 of σ^2 is available, the nonconjugate and conjugate priors would be simple parametrizations of each other. Nevertheless, the priors are different between them. Indeed, in the conjugate case, the marginal distribution of β_k , given χ , is:

$$\pi(\beta_k|\chi) = (1 - \chi_k) \cdot T(\varphi, 0, \lambda_\chi \gamma_{0, \chi_k}) + \chi_k \cdot T(\varphi, 0, \lambda_\chi \gamma_{1, \chi_k}) \quad (24)$$

where $T(\varphi, 0, \lambda_\chi \gamma_{p, \chi_k})$, with $p = (0, 1)$, is the t distribution with φ degrees of freedom and scale parameter $\lambda_\chi \gamma_{p, \chi_k}$.

Given the hierarchical setup, the marginal posterior distribution $\pi(\chi|Y)$ contains the relevant information for variable selection. Based on the data Y , the posterior $\pi(\chi|Y)$ updates the prior probabilities on each of the 2^m possible values of χ . Identifying each χ with a submodel via $\chi_k = 1$ if and only if χ_k is included, the χ 's with higher posterior probability $\pi(\chi|Y)$ will identify the more promising submodels and thus supported most by the data and the prior distributions.

⁸Similar frameworks, with appropriate Bayesian empirical specifications, have been used to make inference and obtain posterior distributions among time-varying macroeconomic-financial variables in multicountry panel data (see, e.g., Canova and Ciccarelli (2004, 2009), Canova et al. (2007, 2012), and Pacifico (2019a, 2019b)).

Thus, a reasonable choice might have the χ_k 's independent with marginal distribution:

$$\pi(\chi_k) = w_{|\chi|} \cdot \binom{k}{|\chi|}^{-1} \quad (25)$$

where $w_{|\chi|}$ denotes the model prior choice related to the Prior Inclusion Probability (PIP) with respect to the model size $|\chi|$, through which the β_k 's will require a non-0 estimate or the χ_k 's should be included in the model. In this way, one would weight more according to model size and, by setting $w_{|\chi|}$ large for smaller $|\chi|$, assign more weight to parsimonious models. Such priors would work well in both situations when k is moderate (e.g., equal to or less than 15) and large (e.g., more than 15 but less than 20), yielding sensible results.

2.4. MCMC integration

The factorization in β and σ , given $|\chi|$, allows us to easily construct MCMC algorithms for simulating a Markov chain:

$$\chi^{(0)}, \chi^{(1)}, \chi^{(2)}, \dots, \chi^{(m)} | \chi_k, Y \xrightarrow{d} \pi(\chi | Y) \quad (26)$$

where $\chi^{(0)}$ is automatically assigned to the model selection procedure in absence of any relationship between potential predictors. The sequence in Eq. (26) is converging in distribution to $\pi(\chi | Y)$ and exactly contains the information relevant to variable selection. Thus, an ergodic Markov chain in which it is embedded is:

$$\beta^{(0)}, \sigma^{(0)}, \chi^{(0)}, \beta^{(1)}, \sigma^{(1)}, \chi^{(1)}, \beta^{(2)}, \sigma^{(2)}, \chi^{(2)}, \dots, \xrightarrow{d} \pi(\beta, \sigma, \chi | Y) \quad (27)$$

where $\beta^{(0)}, \sigma^{(0)}, \chi^{(0)}$ are automatically assigned to the model selection procedure in absence of any relationship between potential predictors, with $\sigma^{(0)}$ corresponding to the full variance of Y . The sequence in Eq. (27) converges in distribution to the full posterior $\pi(\beta, \sigma, \chi | Y)$ and would correspond to an auxiliary Gibbs sequence.

In general variable selection problems, where the number of potential predictors k is moderate (e.g., equal to or less than 15), the sequence in Eq. (26) can be used to evaluate the full posterior $\pi(\chi | Y)$. In large problems (e.g., when k is more than 15 but less than 20), it will still provide useful and faster information, performing more with respect to model size and being more effective than a *per-iteration* basis for learning about $\pi(\chi | Y)$.

The main advantage of using the conjugate hierarchical prior is that it enables analytical marginalizing out of β and σ from $\pi(\beta, \sigma, \chi | Y) = f(Y | \beta, \sigma) \cdot \pi(\beta | \sigma, \chi) \cdot \pi(\sigma) \cdot \pi(\chi)$. Combining the likelihood from (3) with the priors (20) and (22), it yields to the joint posterior:

$$\pi(\beta, \sigma, \chi | Y) \propto \sigma^{-(n+m_\chi+\nu+1)} \cdot |\Sigma_{(\sigma, \chi)}|^{-1/2} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \cdot |\bar{Y} - \bar{X}_\chi \beta|^2 \right\} \cdot \exp \left\{ -\frac{\nu \lambda_\chi}{2\sigma^2} \right\} \cdot \pi(\chi) \quad (28)$$

where $\bar{Y} = [Y \ 0]'$ and $\bar{X}_\chi = [\mathcal{X}_\chi \ (\Sigma_{\sigma, \chi})^{-1/2}]'$ are $(2 \cdot 1)$ vectors, with $\mathcal{X}_\chi$ being a $(n \cdot m_\chi)$ matrix whose columns correspond to the components of β .

Integrating out β and σ yields:

$$\pi(\chi | Y) \propto \pi(\chi) \equiv |\bar{\mathcal{X}}'_\chi \bar{\mathcal{X}}_\chi|^{-1/2} \cdot |\Sigma_{(\sigma, \chi)}|^{-1/2} \cdot (\nu \lambda_\chi + S_\chi^2)^{-(n+\nu)/2} \cdot \pi(\chi) \quad (29)$$

where $\pi(\chi)$ denotes the marginal distribution of the χ_k 's and S_χ^2 is the decomposition of the variance in the procedure selection and defined as:

$$S_\chi^2 = \bar{Y}' \bar{Y} - \bar{Y}' \bar{\mathcal{X}}_\chi (\bar{\mathcal{X}}'_\chi \bar{\mathcal{X}}_\chi)^{-1} \bar{\mathcal{X}}'_\chi \bar{Y} = Y' Y - Y' \mathcal{X}_\chi (\mathcal{X}'_\chi \mathcal{X}_\chi + (\Sigma_{(\sigma, \chi)})^{-1})^{-1} \mathcal{X}'_\chi Y \quad (30)$$

The generation of the components in Eq. (26) in conjunction with $\pi(\chi)$ in Eq. (29) can be obtained trivially as simulations of Bernoulli draws. The required sequence of Bernoulli probabilities can be computed fast and efficiently by exploiting the appropriate updating scheme for $\pi(\chi)$:

$$\frac{\pi(\chi_k = 1, \chi_{(k)}|Y)}{\pi(\chi_k = 0, \chi_{(k)}|Y)} = \frac{\pi(\chi_k = 1, \chi_{(k)})}{\pi(\chi_k = 0, \chi_{(k)})} \quad (31)$$

At each step of the iterative simulation from (26), one of the values of $\pi(\chi)$ in Eq. (31) will be available from the previous component simulation.

The attractive feature of the conjugate prior is in the availability of the exact $\pi(\chi)$ values, providing useful information about $\pi(\chi|Y)$. Firstly, the exact relative probability of two values χ_0 and χ_1 is obtained as $[\pi(\chi_0)/\pi(\chi_1)]$. This allows for more accurate identification of the high probability models among those selected. Then, only minimal additional effort is required to obtain these relative probabilities since $\pi(\chi)$ must be calculated for each of the visited χ values in the execution of the MCMC algorithm. In addition, the availability of $\pi(\chi)$ also makes it possible to estimate a normalizing constant C as:

$$\pi(\chi|Y) = C \cdot \pi(\chi) \quad (32)$$

To be more precise, let $\mathcal{A}$ be a preselected subset of χ values and let $\pi(\mathcal{A}) = \sum_{\chi \in \mathcal{A}} \pi(\chi)$ so that $\pi(\mathcal{A}|Y) = C \cdot \pi(\mathcal{A})$, a consistent estimate⁹ of C for a simulation of $\mathcal{A}$ is obtained by:

$$\hat{C} = \frac{1}{\pi(\mathcal{A}) \cdot T} \cdot \sum_{i=1}^n I_{\mathcal{A}}(\chi_k^i) \quad (33)$$

where $I_{\mathcal{A}}()$ is the indicator of the set $\mathcal{A}$. Because the selection of $\mathcal{A}$ cannot depend on the simulation sequence used to evaluate the Eq. (33), $\mathcal{A}$ is chosen to be a set of χ values visited by a preliminary simulation of $\mathcal{A}$.

Inserting $\hat{C}$ into the Eq. (32), it yields improved estimates of the probability of individual χ values:

$$\hat{\pi}(\chi|Y) = \hat{C} \cdot \pi(\chi) \quad (34)$$

and thus an estimate of the total visited probability:

$$\hat{\pi}(\mathcal{B}|Y) = \hat{C} \cdot \pi(\mathcal{B}) \quad (35)$$

where $\mathcal{B}$ is the set of visited χ values.

Finally, since $\hat{\pi}(\chi|Y)/\pi(\chi|Y) = \hat{C}/C$, it is easy to obtain an useful quantity to measure the uniform prediction accuracy of the probability estimates:

$$|(\hat{C}/C) - 1| \quad (36)$$

The Eq. (36) also provides a measure of the total probability discrepancy since:

$$\sum_{\chi} |\hat{\pi}(\chi|Y) - \pi(\chi|Y)| = |\hat{C} - C| \cdot \sum_{\chi} \pi(\chi) = |(\hat{C}/C) - 1| \quad (37)$$

3. Investigating endogeneity issues: A TSLS-ROB procedure

When studying and modeling economic interactions, endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems, can occur. More precisely, since the ROB procedure is strictly related to BMA approach, is

⁹Here, a consistent estimator or asymptotically consistent estimator is an estimator having the property that, as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to C .

constrained to modeling associations between covariates and the dependent variable, ignoring issues of endogeneity and uncertainty surrounding instrument specifications. Thus, to investigate endogeneity issues, I implement the methodology in order to model uncertainty at both the instrument and covariate level in ROB procedure. To do it, I build on Lenkoski et al. (2014) by combining the BMA approach to regression variable uncertainty and the Two-Stage Least Squares (TSLS) procedure and extend their framework by using CIP priors for the endogenous variables in order to address model uncertainty in regression models with the classic technique of TSLS. Finally, the validity of instruments in the TSLS-ROB procedure is addressed by constructing a Bayesian test of the identification restrictions based on model averaged posterior predictive p-values.

The main thrust of TSLS-ROB procedure would be to choose the *best* model providing the most accurate predictive performance of the form (8), where the predictors belonging to the subset of $X_1, X_2, \dots, X_m, \sum_{l \ll k} X_l$, are going to possibly be instrumentalized.

According to Eq. (3), stacking for n observations and dropping the intercept for convenience, I consider the system:

$$Y = \sum_{k=1}^m \theta_k (W_k X_k) + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma_\epsilon^2) \quad (38)$$

$$W = \sum_{\tilde{k}=1}^{\tilde{m}} \tilde{\theta}_{\tilde{k}} (Z_{\tilde{k}} X_{\tilde{k}}) + \vartheta \quad \text{with} \quad \vartheta \sim N(0, \sigma_\vartheta^2) \quad (39)$$

where Y_i is the $n \cdot 1$ vector denoting the variable of interest, X is the $[n \cdot m]$ matrix of covariates, W is the $n \cdot 1$ vector denoting the set of endogenous variables, θ_k and $\tilde{\theta}_{\tilde{k}}$ are the vectors of unknown regression coefficients, with $\tilde{k} \neq k$, Z is the $[n \cdot \tilde{m}]$ matrix of instruments, and ϑ_i is the $n \cdot 1$ vector of disturbances independent and identically distributed with respect to ϵ_i , with σ_ϑ to be an unknown positive scalar.

Following the framework of Kleibergen and Zivot (2003), let $U = (WX)$ and $D = (ZX)$, I rewrite the Eqs. (38) and (39) as:

$$Y = \tilde{U}(\tilde{\theta}_{\tilde{k}})\theta_k + \varphi \quad \text{with} \quad \varphi \sim N(0, \sigma_\epsilon^2) \quad (40)$$

$$W = D\tilde{\theta}_{\tilde{k}} + \vartheta \quad \text{with} \quad \vartheta \sim N(0, \sigma_\vartheta^2) \quad (41)$$

where $\tilde{U}(\tilde{\theta}_{\tilde{k}}) = (D\tilde{\theta}_{\tilde{k}}X)$ is the replacement of W in U with its fitted value, and the Ordinary Least Squares (OLS) estimate regressing W on D is given by $\hat{U} = \tilde{U}(\hat{\theta}_{\tilde{k}})$, with $\hat{\theta}_{\tilde{k}} = (D'D)^{-1}D'W$. Finally, since in the presence of endogeneity the correlation between φ and ϑ leads to inconsistent estimates of the coefficient vector θ_k under standard OLS, I use the TSLS estimator in order to solve this inconsistency problem when Z instruments occur. The latter would be independent of Y , given the set of endogenous variables (W) and covariates (X). It is defined as:

$$\hat{\theta}^{TSLS} = (\hat{U}'\hat{U})^{-1}\hat{U}'Y \quad (42)$$

According to Eqs. (40) and (41), the TSLS conditional likelihoods are given by:

$$L(\theta|Y, \hat{\theta}, \sigma_\varphi^2) = \sigma_\varphi^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \frac{(Y - \hat{U}\theta)' \cdot (Y - \hat{U}\theta)}{\sigma_\varphi^2}\right) \quad (43)$$

$$L(\tilde{\theta}|W, D, \sigma_\vartheta^2) = \sigma_\vartheta^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \frac{(W - D\tilde{\theta})' \cdot (W - D\tilde{\theta})}{\sigma_\vartheta^2}\right) \quad (44)$$

Before estimating Eqs. (40) and (41), informative priors need to be constructed. Let the CIP prior in Eq. (19), I suggest the form:

$$\pi(\beta, \tilde{\beta}, \sigma_\phi, \chi) = \pi(\beta|\tilde{\beta}, \sigma_\phi, \chi) \cdot \pi(\sigma_\phi|\chi) \cdot \tilde{\pi}(\chi) \quad (45)$$

where $\tilde{\beta}$ is an additional auxiliary variable corresponding to the regression parameters in implementing variable selection when instruments occur, with $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_m)$, $\sigma_\phi = \Omega^{1/2}$, with $\Omega = \text{Var}\left(\frac{\varphi}{\vartheta}\right)$, and $\tilde{\pi}(\chi) = w_{|\chi|} \cdot \left(\frac{\tilde{k}}{|\chi|}\right)^{-1}$ as defined in Eq. (25), but adjusted for the auxiliary regression in (41). The right-hand side priors of Eq. (45) can be defined as:

$$\pi(\beta|\tilde{\beta}, \sigma_\phi, \chi) = \underbrace{\pi(\beta|\tilde{\beta}, \chi)}_{\text{conjugate}} = N_k(\mu_\chi, \Sigma_{\sigma_\phi, \chi}) \quad (46)$$

$$\pi(\sigma_\phi|\chi) = IG\left(\frac{\nu}{2}, \frac{\nu \cdot \tilde{\lambda}_\chi}{2}\right) \quad (47)$$

Here, the underlying logic of TSLS-ROB procedure is the same as for ROB procedure (Section 2.2), but Eqs. (46) and (47) differ from Eqs. (20) and (22) for three reasons: (i) in the second stage (Section 2.2.2), I am averaging by using a variable selection of possible set of instruments $Z_{\tilde{k}}$ given the set of covariates excluded in the first stage ($X_k \in M_k$); (ii) $\tilde{\lambda}_\chi = s_{TSLS}^2$; and (iii) the IG prior is used to model the residual variance σ_ϕ^2 , implied in (40).

Combining the likelihoods in (43) and (44) with the priors (45) and (47), it yields to the joint posterior adjusted for the TSLS-ROB procedure:

$$\tilde{\pi}(\chi|Y) \propto \tilde{\pi}(\chi) \equiv |\hat{\tilde{X}}_\chi' \hat{\tilde{X}}_\chi|^{-1/2} \cdot |\Sigma_{(\sigma_\phi, \chi)}|^{-1/2} \cdot (\nu \tilde{\lambda}_\chi + \tilde{S}_\chi^2)^{-(n+\nu)/2} \cdot \tilde{\pi}(\chi) \quad (48)$$

where $\tilde{S}_\chi^2$ is the decomposition of the variance defined in Eq. (30), but adjusted for the TSLS-ROB procedure, and $\hat{\tilde{X}}_\chi = [\tilde{X}_\chi \quad (\Sigma_{(\sigma_\phi, \chi)})^{-1/2}]'$, with $\tilde{X}_\chi$ being a $(n \cdot m_\chi)$ matrix whose columns correspond to the components of $\tilde{\beta}$. Here, MCMC algorithms for simulating a Markov chains are easily constructed through the factorization in β , $\tilde{\beta}$, and σ_ϕ , given the model size $|\chi|$.

To construct a Bayesian test of identification restrictions, posterior predictive p-values are used since they are useful tools in applied Bayesian statistics to verify model assumptions and whether each submodel—which is used to form TSLS-ROB procedure—is appropriately identified. In the case of TSLS estimation in the presence of model uncertainty, the BMA framework must be extended to account for model uncertainty at both stages. Thus, let $\tilde{\pi} = \pi(M_l|Y_n = y_n, \hat{\chi})$ denoting the PMPs in TSLS-ROB procedure, the TSLS estimator in (42) becomes:

$$\hat{\theta}^{TSLS} = \sum_l \tilde{\pi} \cdot \hat{\theta}_{\tilde{k}} \quad (49)$$

The corresponding p-value in the TSLS-BMA approach and thus the appropriate Bayesian test of identification restrictions can be defined as:

$$\tilde{\omega} = \sum_l \hat{\pi}(Y) \cdot \tilde{\pi} \cdot \tilde{\omega}_l \quad (50)$$

where $\tilde{\omega}_l = \text{Pr}(nR_l^2 > \chi_{p_X + p_Z - 1}^2 | W)$, with R_l^2 denoting the R^2 associated with the auxiliary regression in (41) on all X and Z variables, given the data Y and the total number of X (p_X) and Z (p_Z) included in the TSLS-ROB procedure, and χ^2 being the usual chi-squared sampling distribution.

The performance and functioning of the TSLS-ROB procedure are assessed in Section (4.1.1).

Table 1. Crime data.

Idx.	Predictor	Label	PIP (%)
1	Percentage of males aged 14–24 in total state population	(<i>pmale</i>)	0.00
2	Indicator variable for a southern state	(<i>south</i>)	0.00
3	Mean years of schooling of the population aged ≥ 25	(<i>edu</i>)	1.26
4	Per capita expenditure on police protection in 1960	(<i>po60</i>)	1.41
5	Per capita expenditure on police protection in 1959	(<i>po59</i>)	1.36
6	Labor force participation rate of males aged 14–24	(<i>labor</i>)	0.62
7	Number of males per 1,000 females	(<i>nmale</i>)	1.06
8	State population in 1960 in hundred thousands	(<i>pop</i>)	1.03
9	Percentage of nonwhites in the population	(<i>nnwhite</i>)	1.26
10	Unemployment rate of urban males 14–24	(<i>unem1</i>)	2.06
11	Unemployment rate of urban males 35–39	(<i>unem2</i>)	3.53
12	Wealth (median value)	(<i>wealth</i>)	8.54
13	Income inequality (%)	(<i>ineq</i>)	1.16
14	Probability of imprisonment	(<i>prob</i>)	0.59
15	Average time in months in state prisons	(<i>time</i>)	0.73
–	Crime rate per 100,000 population in 1960	(<i>crime</i>)	–

The Table is so split: the first column denotes the predictor number; the second and the third column describe the predictors and the corresponding labels; and the fourth column displays the PIPs (in %) in Eq. (14) for each predictor. The last row refers to the dependent variable.

4. Applications

4.1. *Us crime data*

In order to display the performance of the ROB procedure in Section (2.2), I replicate the example in Ehrlich's analysis about the criminal behavior in 47 U.S. states in 1960.¹⁰ He argues that the costs of crime are related to the probability of imprisonment and the average time served in prison, affected by police expenditures. The benefits of crime are related to both the aggregate wealth and income inequality in the enclosing community. The expected net payoff from legitimate activities are expected to be lower for nonwhites and for young males than for others in 1960 and thus states with high proportions of these are expected to have higher crime rates.

The dataset collects 15 candidate predictors of crime rate that denotes the variable of interest (Table 1). As in the original analysis, all data have been transformed in logarithms and there are $2^{15} = 32,768$ possible model solutions. Ehrlich's analysis focuses on the relationship between the probability of imprisonment (predictor 14) and the average time served in state prisons (predictor 15), by running two regression models: model 1 consisting of the predictors (9, 12, 13, 14, 15) and model 2 consisting of the predictors (1, 6, 9, 10, 12, 13, 14, 15).

A MC^F approach is used in the first stage of the ROB procedure. More precisely, I sample 100,000 iteration draws, starting from 2^{m-1} , and I fix the benchmark prior to $\max(47, 32768)$ in order to include all possible combination of predictors counted in the natural model space $\mathcal{M}$. Here, some considerations are addressed: (i) the probability of imprisonment (predictor 14) and the average time served in state prisons (predictor 15) are not potential predictors of crime rate as indicated in Ehrlich's analysis; (ii) the aggregate wealth (predictor 12) and the income inequality (predictor 13), accounting for the benefits of crime, should be interpreted with care; (iii) the decision to engage in criminal activity is related to some legitimate activities, such as the educational level (predictor 3) and the availability of employment measured by the unemployment rates (predictors 10 and 11)¹¹; (iv) the percentage of nonwhites (predictor 9) would be a potential predictor of crime rate, but it is not the same for the percentage of young males (predictor 1) as in

¹⁰See, e.g., Ehrlich (1973).

¹¹The result follows the studies of Becker (1968) and Stigler (1970), who argue that the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to some legitimate opportunities.

Table 2. Best potential combination of predictors for crime data.

Predictor	Idx.	PIP (%)	PSD	CPS
edu	3	22.09	1.77	0.00
po60	4	26.28	0.65	0.00
po59	5	17.72	0.54	0.04
nmale	7	8.15	2.30	0.00
pop	8	5.34	0.06	0.01
nnwhite	9	98.70	0.14	1.00
unem1	10	3.78	0.17	0.00
unem2	11	2.64	0.09	0.63
wealth	12	35.42	1.20	0.00
ineq	13	15.47	0.87	1.00

The Table is so split: the first column refers to the predictors according to the labels defined in Table 1, the second column displays the predictor number as defined in Table 1, and the last three columns display the PIPs (in %), the PSDs, and the CPS, respectively.

Ehrlich's theory; and (v) the predictors not belonging to the subset $\mathcal{S}$ and thus excluded from the procedure are (1, 2, 6, 14, 15).

The inclusion¹² of an indicator variable for southern states (predictor 2), the sex ratio (predictor 7), and the state population in 1960 (predictor 8) has not found empirical validation.

By looking into which predictors are included with higher frequency in the all (possible) model solutions, 10 best potential combination of predictors for crime rate are selected (Table 2), with PMPs $\geq \tau$ in Eq. (12),¹³ obtaining $2^{10} = 1,024$ best model solutions. In Table 2, I display the revised PIPs (e.g., the sum of PMPs for all best models wherein a covariate X_i has been included with the auxiliary variable $\chi = 1$), the coefficients' Posterior Standard Deviations (PSDs), and the Conditional Posterior Sign (CPS) for the sign certainty (e.g., the posterior probability of a positive coefficient expected value conditional on inclusion). Here, the CPS indicates a positive effect on crime rate whether it is close to 1 and a negative effect whether it is close to 0.

The predictors (3, 4, 9, 12) would look like the top best combination of covariates with higher PIPs. They negatively affect the crime rate, except for the percentage of nonwhite (predictor 9). Thus, the expected net payoff to engage in criminal activity is determined from alternative legitimate activities related to: educational level (predictor 3); per capita expenditure on police protection in 1960 (predictor 4); and an indicator variable for origin (predictor 9).

Finally, looking into which models included perform better the data (Table 3), I find 24 top best model solutions (M_{ξ}), with PMPs $\geq \hat{\tau}$ in Eq. (15).¹⁴ Each of them shows strong evidence with respect to the submodels M_l according to the log Bayes Factors in Eq. (16). Accounting for the benefits of crime, related to the aggregate wealth (predictor 12) and income inequality (predictor 13), also if they are accurately specified from a theoretical point of view (fourth column in Table 2), they are excluded in the final subset $\mathcal{E}$ described in Eq. (15).

The final solution (e.g., M_{fp}), in bold type in Table 3, corresponds to the model consisting of predictors (3, 4, 9) with higher LBF. To be more precise, it consists of the following regression model:

$$crime_i = \varrho + \theta_1 edu_{i1} + \theta_2 po60_{i2} + \theta_3 nnwhite_{i3} + \vartheta \quad (51)$$

where ϱ denotes the intercept and ϑ is a $(n \cdot 1)$ vector containing the disturbances. The latter is independent and identically distributed (*i.i.d.*) with respect to the errors ϵ and η defined in Eqs. (3) and (8).

In Table 4, I summarize the ROB procedure step by step. To be more precise, I focus on: (i) the selected predictors according to the criteria used for each stage; (ii) the corresponding size of the model space and the parameter space; and (iii) the corresponding Predictive Coverage (PC),

¹²See, e.g., Vandaele (1987).

¹³I use $\tau = 1\%$ for an enough posterior consistency with predictors ≤ 15 .

¹⁴I use $\hat{\tau} = 0.5\%$ for a sufficient prediction accuracy in explaining the data with predictors ≤ 15 .

Table 3. *Top Best* model solutions with PMPs of 0.5% or larger.

Model	PMP (%)	$IBF_{\mathcal{E},I}$
9 12	25.55	9.09
4 9	16.60	9.18
5 9	10.67	9.16
3 9	8.54	9.02
9 13	6.94	8.71
3 4 9	2.83	9.21
3 9 12	2.29	9.09
3 5 9	1.96	9.20
3 9 13	1.64	9.01
4 7 9	1.59	9.18
7 9 12	1.36	9.07
7 8 9 10	1.14	8.85
5 7 9	1.03	9.17
4 9 12	0.99	9.17
3 8 9	0.92	8.95
9 10 12	0.85	9.06
5 9 12	0.84	9.17
7 9 13	0.81	8.91
9 12 13	0.80	9.06
4 9 13	0.75	9.20
4 9 10	0.65	9.15
8 9 12	0.65	9.03
9 11 12	0.57	9.06
5 9 13	0.52	9.18

The Table is so split: the first column displays the *top best* model solutions ($M_{\mathcal{E}}^*$); the second column displays the PMPs of 0.5% or Larger for each regression model in subset $\mathcal{E} \subset \mathcal{S}$; and the third column shows the log Bayes Factor.

Eq. (35), measured by using the percentage of the proportion of observations in the performance set that fall in the 95% prediction interval.

In Table 5, I compare the results with some related works: (i) two popular alternative variable selection techniques, such as Occam’s window (see, e.g., Raftery et al. (1997)) and MC^3 methodology (see, e.g., Madigan et al. (1995)); (ii) the two regression models tested in the original Ehrlich’s analysis¹⁵; (iii) a Bayesian penalization technique¹⁶ in which the prior distribution performs a function similar to that of the penalty term in classical penalization, such as horseshoe procedure (see, e.g., Bhattacharya et al. (2016)); and (iv) an alternative Bayesian variable selection procedure with spike-and-slab regression,¹⁷ which combines variable selection with estimation of the regression parameters by using mixture priors on regression coefficients with density functions that are peaked at zero (the “spike”) and have large probability mass for a wide range of non-zero values (the “slab”). Regarding the last two procedures, some considerations are in order. The horseshoe priors are computed by assuming the same number of draws used to run MC^F integrations and the *top best* model solutions correspond to the only statistically significant Bayes estimates.¹⁸ The exact and final solution in spike-and-slab regression is obtained by using a cross-validation variable selection procedure with a stability¹⁹ $> 80\%$.

¹⁵Own computations.

¹⁶When accounting for BMA procedure in linear regression models with many predictors, penalized regression techniques are often used to deal with overfitting and select promising variables for predicting a variable of interest.

¹⁷Spike-and-slab regression is a Bayesian variable selection technique particularly useful in high dimensional datasets when the number of possible predictors is larger than the number of observations. See, e.g., Mitchell and Beauchamp (1988), Raftery and Madigan (1994), and George and McCulloch (1997) for further implementations.

¹⁸Here, statistical significance refers to the concept of the standard statistical hypothesis testing with a significance level setting to 5%.

¹⁹In spike-and-slab regression, the *top best* covariates in explaining the data are determined by minimizing the test-set Mean Squared Error (MSE) for its entire solution-path. The average number of times a predictor is selected in this way defines a value recorded from 0% – 100% denoting its stability.

Table 4. ROB procedure for crime data.

Stage	Predictor	Parameter Space	Model Space	Criterion		PC (%)
				PIPs	PMPs	
Full model	All	15 (θ_k)	32, 768 (M_k)	–	–	–
First stage	3 4 5 7 8 9 10 11 12 13	10 (θ_l)	1, 024 (M_l)	$\geq 1\%$	–	100
Second stage	–	–	24 (M_ε)	–	$\geq 0.5\%$	99.95
IBF $_{\xi,l}$	3 4 9	3 ($\subset \theta_l$)	$M_{tp} = 1$	–	–	–

The Table is so split: the first column displays the ROB procedure step by step; the second column describes the potential predictors observed for each step; the third and the fourth column display the parameter space and model space with the corresponding labels; the fifth column describes the criteria used for the shrinking; and the last column refers to the Predictive Coverage (in %).

Table 5. Models selected for crime data: A comparison.

Method	Predictor	SER	$\bar{R}^2$ (%)	F-Test	IBF $_{(M_{tp}, M_*)}$	Evidence
Full model	All	0.10	77.59	11.62	9.21	Very strong
Occam's window/MC ³	1 3 4 9 11 13 14	0.10	78.27	24.67	8.91	Very strong
Ehrlich model 1	9 12 13 14 15	0.12	53.25	11.48	4.99	Moderate
Ehrlich model 2	1 6 9 10 12 13 14 15	0.12	55.47	8.16	5.19	Moderate
Horseshoe	1 3 4 9 10 11 13	0.10	75.76	21.54	3.04	Weak
Spike-and-slab	1 3 4 5 9 14	0.10	67.88	17.21	5.54	Moderate
ROB	3 4 9	0.10	60.74	24.72	–	–

The Table is so split: the first column describes the method; the second column displays the predictors accounted for; the third and the fourth column display the goodness of fit for each method; the fifth column displays the significance for each method; the last two columns refer to the log Bayes Factor and the corresponding scale of evidence (Eq. (17)).

In summary, the final solution reached accounting for ROB procedure shows higher *significant* overall F value and sufficiently *strong* $\bar{R}^2$. By running the IBF between the model M_{tp} and the models related to the alternative approaches (M_*), I find very strong support for model averaging in Raftery et al. (1997), moderate support for some of Ehrlich's theory and spike-and-slab regression, and weak support for horseshoe Bayesian penalization technique.²⁰ In particular, by averaging over all models, the results indicate support for a relationship between crime rate and the predictors (3, 4, 9), but not for the inclusion of indicator variables for the offender's presumed exceptional psychological and social circumstances (predictors 1, 12, 13, 14, 15).²¹

A further support for my results is found in Fig. 1. Here, I show that the posterior distributions for the coefficients of the selected *top best* combination of predictors (3, 4, 9), based on the model averaging results of the ROB procedure, are centered away from 0.

4.1.1. *Us crime and endogeneity issues*

The empirical example displayed in Section 4.1 could be deepened by accounting for possible unobserved endogeneity problems between the crime rate (Y) and, for example, some covariates X_k dropped in the first stage of ROB procedure (Table 2). Thus, interesting endogeneity issues would arise between the crime rate and predictors accounting for labor force participation rate of males aged 14–24 (predictor 6), probability of imprisonment (predictor 14), and average time in months in state prisons (predictor 15). These variables would be the set of instruments Z and the remaining are going to be the set of covariates X (predictors 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13 in Table 1).

Running the TSLS-ROB procedure, I obtain the following model:

$$Y = W + \hat{\theta}_1 X_1 + \hat{\theta}_3 X_3 + \hat{\theta}_4 X_4 + \hat{\theta}_7 X_7 + \hat{\theta}_9 X_9 + \hat{\theta}_{10} X_{10} + \hat{\theta}_{12} X_{12} + \epsilon \quad (52)$$

²⁰Own computations.

²¹See, e.g., Becker (1968), Stigler (1970), and Brier and Fienberg (1980).

Figure 1. Posterior Distributions for the coefficients of the selected *top best* combination of predictors (3, 4, 9), based on the model averaging results of ROB procedure, are displayed. A 95% confidence bands are used to draw posteriors. The Y and X axis represent the probability density and the posterior means of the estimators $\hat{\theta}_i$, where $i = 3, 4, 9$, respectively.

$$W = \hat{\theta}_1 X_1 + \hat{\theta}_3 X_3 + \hat{\theta}_4 X_4 + \hat{\theta}_9 X_9 + \hat{\theta}_{11} X_{11} + \hat{\theta}_{13} X_{13} + \hat{\theta}_{14} Z_{14} + \vartheta \quad (53)$$

The Table 6 summarizes the results. Here, some considerations are addressed: (i) only one variable in Z (predictor 14) serves as instrument; (ii) predictors (3, 4, 9), denoting alternative legitimate activities, tend to determine the expected net payoff to engage in criminal activities with higher significance, just as in the ROB procedure (Table 2); and (iii) variables accounting for males (predictor 7), unemployment (predictor 10), and wealth (predictor 12) also strongly affect Y , as found in the ROB procedure (Table 2).

Nevertheless, four of the variables in X (predictors 1, 3, 4, and 9) have explanatory power on both Y and W and thus also reflect endogeneity issues when instruments occur. This explanatory power tends to increase in (53), when the instrument Z_{14} is accounted for. Moreover, the only one instrument considered (predictor 14) negatively affects W and thus Y . These findings highlight that the crime rate would be driven by the probability of imprisonment over time and not in an inverse way.

Finally, the posterior predictive p-values for both Eqs. (52) and (53) are close to zero and thus model assumptions are verified and each submodel of X_k , used to form TSLS-ROB procedure, is appropriately identified.

Table 6. TSLS-ROB.

Predictor	Variable	Model (52)	Model (53)	Sign/Effect
X_1	pmale	(0.062)*	(0.046)**	(-)
X_3	edu	(0.004)***	(0.002)***	(-)
X_4	po60	(0.004)***	(0.002)***	(-)
X_7	nmale	(0.042)**	-	(-)
X_9	nnwhite	(0.006)***	(0.001)***	(+)
X_{10}	unem1	(0.037)**	-	(-)
X_{11}	unem2	-	(0.002)***	(+)
X_{12}	wealth	(0.033)**	-	(-)
X_{13}	ineq	-	(0.001)***	(+)
X_{14}	prob	-	(0.000)***	(-)
		$\bar{R}_{MS2}^2 = 76.59$	$\bar{R}_{MS3}^2 = 77.94$	
		$\bar{\alpha}_{I,MS2} = 0.00$	$\bar{\alpha}_{I,MS3} = 0.00$	

The Table refers to TSLS-ROB procedure accounting for Eqs. (52) and (53). It is so split: the first column refers to predictors; the second column denotes the variables according to labels used in Table 1; the third and fourth column display the estimates in terms of p-values with their significant codes for both the two models; and the fifth column displays the effect on the crime rate. The last two rows refer to the adjusted R^2 and the posterior predictive p-values for both the models. The significant codes are: *** significance at 1%, ** significance at 5%, and * significance at 10%.

4.2. Simulated example

In this example, the performance and the usefulness of ROB procedure are investigated by following the format of George and McCulloch (1993). To be more precise, I construct two distinct sets: a training set by simulating $p = 15$ predictors and $n = 50$ observations as independent standard normal vectors and a prediction set by generating 50 additional observations in the same manner. The model used to generate the response is:

$$Y = X_3 + X_4 + X_9 + \ddot{q} \quad (54)$$

where $\ddot{q} \sim N_{50}(0, \sigma^2)$, with $\sigma = 2.5$. Least squares estimates²² for these data are given in Table 7. Here, the tuning parameters τ and $\hat{\tau}$ are put equal to 0.5% for a sufficient prediction accuracy and posterior consistency with predictors ≤ 15 (Eqs. (12) and (15)).

In this example, the predictors are highly correlated between them and there is moderate model uncertainty since the only p -values for β_1 , β_2 , β_{10} , and β_{11} are ≤ 0.10 .

Looking into which methods fit better the data, I compare the performance of the ROB procedure with MC³ methodology of Madigan et al. (1995) and two popular variable selection techniques (Table 8). They are²³: minimum Mallows C_p and maximum adjusted R^2 . The main thrust of this example is to prove that model averaging has better predictive performance than other single models selected using standard techniques. The ROB procedure, even if only slightly, would perform better the data than MC³ methodology dealing with both model uncertainty and overfitting when constructing multiple linear regression models.

5. ROB procedure in time-varying high dimensional data

Although the importance of uncertainty on decision-making has been realized by monetary policy-makers, the vast bulk of analysis has taken refuge in a certainty equivalence framework that enables policy-makers to disregard the properties of uncertainty. For example, if the countries' responses and dynamics of the economy can be adequately represented with linear equations, certainty equivalence implies it is only necessary to focus on the first moments of future outcomes, appropriately discounted, to derive optimal policy. Nevertheless, when jointly modeling and quantifying multicountry data using the information contained in a large set of endogenous

²²Own computations.

²³See, e.g., Weisberg (2005).

Table 7. Least squares estimates with $\hat{\sigma} = 2.75$.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
$\hat{\beta}$	0	0	0	1.00	1.00	0	0	0	0	1.00	0	0	0	0	0	0
$SE(\hat{\beta})$	.24	.25	-.30	-.04	.04	-.10	.06	.09	.09	.03	.30	.24	.15	.00	.18	.02
p -value	.48	.10	.04	.80	.78	.50	.69	.54	.46	.84	.02	.09	.33	.99	.20	.92

The Table displays the Least Squares Estimates for each of the simulated predictors (second row) and the corresponding Standard Errors and p -values (third and fourth row, respectively).

Table 8. Performance comparison.

Method	Predictor	PC (%)
Full model	All	59.76
MC^3	Model Averaging	98.04
C_p	1 2 6 7 8 9 13	65.32
Adj. R^2	1 2 10 11	65.97
ROB	Model Averaging	99.05

The Table is so split: the first column describes the method; the second column displays the predictors accounted for; and the third column refers to the Predictive Coverage (in %) for a 95% prediction interval, with PMPs greater than or equal to 0.5%.

macroeconomic-financial variables, it no longer suffices to focus solely on the first moments of possible outcomes entering, for example, feedback effects from policy implications and temporary or persistent long-run effects associated with systemic events. Thus, one may need to characterize all moments or, equivalently, the entire distribution of possible outcomes. Consequently, forecasters should provide density forecast combinations rather than simply point forecasts reflecting expected values, by using different prior information sets in order to deal with misspecified dynamics. From the same token, Bayesian methods should be used to reduce the dimensionality of the model, structure the time variations, and evaluate issues of endogeneity. In this way, policies designed to protect the economy against the worst-case consequences of misspecified dynamics are less aggressive and good approximations of the estimated rule. Moreover, some drawbacks of robust policies, a connection between the degree of structure an authority assigns to the uncertainty it faces, and the likelihood of policy attenuation in response to that uncertainty can be assessed.

In this section, I extend the ROB procedure to dynamic multicountry panel data in order to jointly deal with endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems. More precisely, I build on Pacifico (2019b) and estimate a simplified version of the SPBVAR by defining hierarchical prior specification strategy and MC^F implementations in order to extend variable selection procedure to a wide array of candidate models and thus candidate density forecast combinations. In this context, all possible model solutions are structured and thus they can include parametric models in which parameter values vary over time in ways that the decision maker cannot describe probabilistically. Because all parametric models are misspecified, I also consider alternative probability distribution setups with much less structure.

In this empirical example, by following the framework of Pacifico (2019b), ROB performance and forecast density combinations are highlighted according to a time-varying multicountry setup of six developed European economies and current members of the Eurozone.²⁴ In this context, I focus on the only six endogenous variables involved to describe the real and financial dimensions: three real variables (general government spending, gross fixed capital formation, GDP growth rate) and three financial variables (general government debt, current account balance, interest

²⁴They are: Italy (IT); Spain (ES); France (FR); Germany (DE); Greece (GR); and Portugal (PT).

Table 9. Data.

Variables	Description
General government spending	Financial accounts for general government spending.
Gross fixed capital formation	Investments of fixed assets at current prices weighted.
GDP growth rate	It is calculated as: $\text{Log}\left(\frac{\text{GDP}_{t,j}}{\text{GDP}_{t-1,j}}\right)$.
General government debt	Non-financial accounts for general government debt.
Current account balance	Non-financial accounts for general government net.
Interest rate	Long term interest rate.
Productivity	It corresponds to logarithm of the real GDP per capita.

Here, general government spending, gross fixed capital formation, public debt, and current account balance are weighted for the GDP.

rate). The real GDP per capita in logarithmic form (hereafter called *productivity*) is used as variable of interest. The overall series is expressed in quarters, and all data points originate from Eurostat and OECD data sources (Table 9).

The simplified version of the time-varying SPBVAR developed in Pacifico (2019b) takes the form:

$$Y_{it\ddot{m}} = A_{it,j\ddot{m}}(L)Y_{i,t-1\ddot{m}} + \ddot{e}_{it} \quad (55)$$

where $i, j = 1, 2, \dots, 6$ are country indices, $t = 1, 2, \dots, T$ denotes time, $\ddot{m} = 1, \dots, 6$ denotes the set of endogenous variables, $A_{it,j}$ is a $36 \cdot 36$ matrix of real and financial variables for each pair of countries (i, j) for a given $\ddot{m}$, $Y_{i,t-1}$ is a $36 \cdot 1$ vector of lagged variables of interest that accounts for the real and financial dimensions for each i for a given $\ddot{m}$, and $\ddot{e}_{it} \sim i.i.d.N(0, \ddot{\Sigma})$ is an $36 \cdot 1$ vector of disturbance terms. For convenience, I suppose one lag and no intercept.

The estimation sample covers the period from March 1999 to December 2013. It amounts, without restrictions, to 2,160 regression parameters. More precisely, each equation of the time-varying SPBVAR in (55) has 36 coefficients and there are 60 equations in the system. According to the BMA methodology, there are $2^{36} = 68,719,476,736$ possible model solutions. In this context, it is very costly to select subsets of covariates via MC^F simulations by fixing the benchmark prior $\max(N, |\chi|)$, with $|\chi|$ denoting the model size. In fact, the number of coefficients is increased by $N\ddot{M}$ factors, with $\ddot{M}$ denoting the overall lagged endogenous variables accounted for. Moreover, all coefficients vary over time and thus dynamic relationships, cross-unit lagged interdependencies, and dynamic feedback matter. Then, in order to be able in applying the specifications underlying the ROB procedure (Section 2.3), I need to express the time-varying SPBVAR in (55) in terms of a multivariate normal distribution.

Let $\ddot{k} = N\ddot{M}$ be the number of all matrix coefficients in each equation of the model for each pair of countries (i, j) , a $1 \cdot \ddot{k}$ vector $\ddot{X}_t = (Y_{i,t-1}^{\ddot{m}}, Y_{i,t-2}^{\ddot{m}}, \dots, Y_{i,t-l}^{\ddot{m}})'$ can be defined containing all lagged variables in the system for each i . Then, I define an $N\ddot{M}\ddot{k} \cdot 1$ vector $\ddot{\gamma}_{it,j\ddot{k}} = \text{vec}(\ddot{g}_{it,j\ddot{k}})$ containing all columns, stacked into a vector,²⁵ of the matrix $A_t(L)$ for each pair of countries (i, j) for a given $\ddot{k}$, with $\ddot{g}_{it,j\ddot{k}} = (A_{it,j}^{1'}, A_{it,j}^{2'}, \dots, A_{it,j}^{\ddot{M}'})'$, and $\ddot{\gamma}_t = (\ddot{\gamma}_{1t}', \ddot{\gamma}_{2t}', \dots, \ddot{\gamma}_{Nt}')'$ denoting the time-varying coefficient vectors, stacked for i , for each country-variable pair.

By adapting Bayesian specifications described in Section (2.1) and following the framework of Pacifico (2019b), I can able to express the time-varying SPBVAR in (55) in terms of multivariate normal distribution:

$$Y_t = (I_{N\ddot{M}} \otimes \ddot{X}_t)\ddot{\gamma}_t + \ddot{E}_t \quad (56)$$

²⁵The vec operator transforms a matrix into a vector by stacking the columns of the matrix, one underneath the other.

where $Y_t = (Y_{1t}^{\ddot{m}}, \dots, Y_{Nt}^{\ddot{m}})'$ is an $36 \cdot 1$ vector containing the observable variables of interest for each i for a given $\ddot{m}$, and $\ddot{E}_t = (\ddot{e}_{1t}', \dots, \ddot{e}_{Nt}')'$ is an $36 \cdot 1$ vector containing the random disturbances of the model. In model (56) there is no subscript i since all lagged variables in the system are stacked in $\ddot{X}_t$.

Now, because the coefficient vectors in $\ddot{\gamma}_t$ vary in different time periods for each country-variable pair and there are more coefficients than data points, it is impossible to eliminate $\ddot{\gamma}_t$. Thus, to cope with the problem of dimensionality, I adapt the framework in Pacifico (2019b) and assume $\ddot{\gamma}_t$ has the following factor structure:

$$\ddot{\gamma}_t = \sum_{f=1}^{\ddot{F}} \ddot{G}_f \cdot \beta_{ft} + \ddot{u}_t \quad \text{with} \quad \ddot{u}_t \sim N(0, \Sigma_{\ddot{u}}) \quad (57)$$

where $\ddot{F} \ll N\ddot{M}\ddot{k}$ and $\dim(\beta_{ft}) \ll \dim(\ddot{\gamma}_t)$ by construction, $\ddot{G}_f = [\ddot{G}_{1f}, \ddot{G}_{2f}, \dots, \ddot{G}_{\ddot{F}f}]$ are $N\ddot{M}\ddot{k} \cdot 1$ matrices obtained by multiplying the matrix coefficients $(\ddot{g}_{it,jk})$ stacked in the vector $\ddot{\gamma}_t$ by conformable matrices $\mathfrak{X}_f$ with elements equal to zero and one,²⁶ $\ddot{u}_t$ is an $N\ddot{M}\ddot{k} \cdot 1$ vector of unmodeled variations present in $\ddot{\gamma}_t$, and $\ddot{E}(\ddot{u}_t \ddot{u}_t') = \Sigma_{\ddot{u}} = \Sigma_{\ddot{e}} \otimes \ddot{V}$, where $\Sigma_{\ddot{e}}$ is the covariance matrix of the vector $\ddot{E}_t$ and $\ddot{V} = (\sigma^2 I_{\ddot{k}})$ as in Kadiyala and Karlsson (1997). The idea is to shrink $\ddot{\gamma}_t$ to a much smaller dimensional vector β_t , with $\beta_t = (\beta_{1t}', \beta_{2t}', \dots, \beta_{\ddot{F}t}')'$ denoting the adjusted auxiliary indicator defined in Section 2.3 ($\beta = \{\beta_k\}$), and containing all regression coefficients stacked into a vector.

Running Eqs. (56) and (57) for the model described in (55), I assume that the coefficient vectors in $\ddot{\gamma}_t$ depends on three factors:

$$\ddot{G}_f \beta_{ft} = \ddot{G}_1 \beta_{1t} + \ddot{G}_2 \beta_{2t} + \ddot{G}_3 \beta_{3t} + \ddot{u}_t \quad (58)$$

where, stacking for t , $\beta_f = (\beta_1, \beta_2, \beta_3)'$ contains all the time-varying regression parameters to be estimated and grouped in (potential) *best* submodels. Furthermore, given the structural framework, I am able to also investigate endogeneity issues among countries and sectors.

In this study, the factorization is adapted to the thrust of the ROB procedure. More precisely, the coefficient vectors in $\ddot{\gamma}_t$ represent all the model solutions counted in the natural model space $\mathcal{M}$ (Eq. (7)), and each factor would describe subsets of (potential) *best* combination of covariates to be included in the reduced model space $\hat{\mathcal{M}}$ (Eq. (11)).

Given the factorization in Eq. (58), the reduced-form SPBVAR model in Eq. (56) can be transformed into a Structural Normal Linear Regression (SNLR) model²⁷ with an error covariance matrix of an Inverse-Wishart (IW) distribution.²⁸ By Eqs. (56) and (57), the SNLR model can be written as

$$Y_t = \ddot{\Theta}_t \left(\sum_{f=1}^3 \ddot{G}_f \beta_{ft} + \ddot{u}_t \right) + \ddot{E}_t \equiv \ddot{\chi}_{ft} \beta_{ft} + \ddot{\eta}_t \quad \text{with} \quad \ddot{\Theta}_t = (I_{N\ddot{M}} \otimes \ddot{X}_t) \quad (59)$$

where $\ddot{\Theta}_t$ contains all lagged time-varying variables in the system by construction, $\ddot{\chi}_{ft} \equiv \ddot{\Theta}_{ft} \ddot{G}_{ft}$ is an $N\ddot{M} \cdot 1$ matrix that stacks all coefficients of the system, with $\ddot{\chi}_t = \text{diag}(\ddot{\chi}_{1t}', \ddot{\chi}_{2t}', \ddot{\chi}_{3t}')$, and $\ddot{\eta}_t \equiv \ddot{\Theta}_{ft} \ddot{u}_t + \ddot{E}_t \sim N(0, \sigma_t \cdot \Sigma_{\ddot{u}})$ has a particular heteroskedastic covariance matrix that needs to

²⁶Here, $\mathfrak{X}_f$ would correspond to the auxiliary indicator $\chi = (\chi_k)$ in matrix form.

²⁷See, for instance, Pacifico (2019b) for an illustration of the conformation of the time-varying multicountry SPBVAR model and the exact form of the β_{ft} 's, $\ddot{\chi}_{ft}$'s, and the $\ddot{G}_f$'s.

²⁸The Wishart distribution is a multivariate extension of χ^2 distribution and, in Bayesian statistics, corresponds to the conjugate prior of the inverse covariance-matrix of a multivariate normal random vector.

be taken into account, with $\sigma_t = (I_N + \sigma^2 \ddot{\Theta}_t' \ddot{\Theta}_t)$. If σ^2 converges to zero, the factorization in (58) becomes exact and thus $\ddot{\eta}_t \sim N(0, \ddot{\Sigma})$, with $\ddot{\Sigma} = \text{diag}(\Sigma \ddot{e}', \Sigma \ddot{e}', \dots, \Sigma \ddot{e}')$, depends on the only disturbances contained in $\ddot{E}_t$. Thus, it is uncorrelated with the regressors, and classical OLS can be used to estimate the vectors β_t and $\ddot{\gamma}_t$, and the Bayesian specifications in Section (2.1) hold.

Let $\ddot{M}_{\ddot{v}} = \ddot{M}_{\ddot{v}_1}, \ddot{M}_{\ddot{v}_2}$ be the number of variable groups, with $\ddot{M}_{\ddot{v}_1}$ and $\ddot{M}_{\ddot{v}_2}$ denoting the real and financial variables, respectively, and $\ddot{M}_{\ddot{e}}$ be the number of common factors, referring to total economy (both the real and financial variables). The specified factors in Eq. (58) I assume for the empirical analysis are: (1) $\ddot{\chi}_{1t} \beta_{1t}$ is a $N\ddot{M} \cdot \ddot{M}_{\ddot{v}_1}$ observable cross-country variable-specific indicator for Y_t that accounts for the only real variables in order to assess the *best* combination of predictors affecting the productivity among countries in real economy when formulating policies and forecasting, with $\ddot{\chi}_{1t} = \sum_i \sum_{\ddot{M}_{\ddot{v}_1}} y_{i\ddot{M}_{\ddot{v}_1}t-1}$; (2) $\ddot{\chi}_{2t} \beta_{2t}$ is a $N\ddot{M} \cdot \ddot{M}_{\ddot{v}_2}$ observable cross-country variable-specific indicator for Y_t that accounts for the only financial variables in order to evaluate the *best* combination of predictors affecting the productivity among countries in the financial dimension when formulating policies and forecasting, with $\ddot{\chi}_{2t} = \sum_i \sum_{\ddot{M}_{\ddot{v}_2}} y_{i\ddot{M}_{\ddot{v}_2}t-1}$; and (3) $\ddot{\chi}_{3t} \beta_{3t}$ is a $N\ddot{M} \cdot \ddot{M}_{\ddot{e}}$ observable common indicator for Y_t that accounts for total economy in order to investigate (possible) *best* combination of predictors when studying endogeneity issues and dynamic interdependence in the real and financial dimensions, with $\ddot{\chi}_{3t} = \sum_i \sum_{\ddot{M}_{\ddot{e}}} y_{i\ddot{M}_{\ddot{e}}t-1}$.

By construction, hierarchical Conjugate Informative Priors (Section 2.3) and MCMC simulations (Section 2.4) can be used without loss of estimation efficiency. Nevertheless, further specifications and MC^F implementations need to be done. More precisely, since the coefficient vectors β_t vary over time, I need to suppose the following state-space structure:

$$\beta_t = \beta_{t-1} + \ddot{v}_t \quad \text{with} \quad \ddot{v}_t \sim N(0, \ddot{H}_t) \quad (60)$$

where $\beta_t = (\beta_{1t}, \beta_{2t}, \dots)'$, $\ddot{H}_t = \text{diag}(\ddot{H}_{1t}, \ddot{H}_{2t}, \ddot{H}_{3t})$ is a block diagonal matrix, and $\ddot{H}_{\ddot{f}t} = (\ddot{h}_{\ddot{f}t} \cdot I)$, where $\ddot{h}_{\ddot{f}t}$ controls the tightness (stringent conditions) of the factorization ($\ddot{f}$) of the time-varying coefficient parameter (β_t) in order to make it estimable. The errors $\ddot{E}_t, \ddot{u}_t$, and $\ddot{v}_t$ are mutually independent. The random-walk assumption in (60) is very common in the time-varying VAR literature and has the advantage of focusing on permanent shifts and reducing the number of parameters in the estimation procedure.

Supposing exact factorization, mixture prior distributions in (19) need to be improved. More precisely, I need to define prior moments on $(\Sigma \ddot{e}^{-1}, \ddot{h}_{\ddot{f}0}, \beta_0)$. Thus, collecting them in a vector $\ddot{\phi}_0$, with $\ddot{\phi}_0 = (\Sigma \ddot{e}^{-1}, \ddot{h}_{\ddot{f}0}, \beta_0)$ be a mixture density of conjugate priors, the likelihood function can be derived from the sampling density $p(Y|\ddot{\phi}_0)$ by using a mixture hierarchical distribution. In other words, I use: (i) a Normal distribution for factors β ; (ii) a Wishart distribution for $\Sigma \ddot{e}^{-1}$; and (iii) an Inverse-Gamma distribution for $\ddot{h}_{\ddot{f}}$, where $\ddot{h}_{\ddot{f}} = \text{vec}(\ddot{H}_t)$. That is,

$$\beta | \Sigma \ddot{e}^{-1}, Y \sim N(\hat{\beta}, \Sigma \ddot{e}^{-1} \otimes (\ddot{\chi}' \ddot{\chi})^{-1}) \quad (61)$$

$$\Sigma \ddot{e}^{-1} | Y \sim W(\ddot{S}^{-1}, T - \ddot{k} - 1) \quad (62)$$

$$\ddot{h}_{\ddot{f}} | Y \sim IG\left\{\frac{\ddot{\omega}}{2}, \frac{\ddot{v}\ddot{S}}{2}\right\} \quad (63)$$

where $\hat{\beta} = (\sum_t \ddot{\chi}_t' \ddot{\chi}_t)^{-1} (\sum_t \ddot{\chi}_t' Y_t)$ is the OLS estimate of β , $\ddot{S} = (Y_t - \hat{\beta} \ddot{\chi}_t)' (Y_t - \hat{\beta} \ddot{\chi}_t)$ is the sum of squared errors, $\hat{\Sigma}_{\ddot{e}} = \ddot{S} / (T - \ddot{k})$ is the OLS estimate of $\Sigma_{\ddot{e}}$, and $\ddot{\omega}$ and $\ddot{v}$ are hyper-parameters to be chosen. In addition, because such prior assumptions will generally be influenced (for example) by common or subjective beliefs about marginal effects of economic variables,

Independent Normal Wishart prior is used in this analysis, since it assumes that tentative beliefs on $\ddot{\phi}_0 = (\Sigma\ddot{e}^{-1}, \ddot{h}_{\ddot{f}0}, \beta_0)$ derive from separate considerations.

The posterior distributions for $\ddot{\phi} = (\Sigma\ddot{e}^{-1}, \ddot{h}_{\ddot{f}t}, \{\beta_t\}_{t=1}^T)$ are calculated by combining the prior with the (conditional) likelihood for the initial conditions of the data. The resulting function for the full model M_k , containing all coefficient vectors in $\ddot{y}_t$ in (57), is then proportional to

$$L(Y^T | \ddot{\phi}, M_k) \propto (\Sigma\ddot{e})^{-\frac{T}{2}} \exp \left\{ -\frac{1}{2} [\Sigma_t(Y_t - (\ddot{\Theta}_t\ddot{G})\beta_t)'] \Sigma\ddot{e}^{-1} [\Sigma_t(Y_t - (\ddot{\Theta}_t\ddot{G})\beta_t)] \right\} \quad (64)$$

where $Y^T = (Y_1, \dots, Y_T)$ denotes the data, and $\ddot{\phi} = (\Sigma\ddot{e}^{-1}, \ddot{h}_{\ddot{f}t}, \{\beta_t\}_{t=1}^T)$ refers to the unknowns whose joint probability distribution needs to be found. According to Eqs. (59) and (60), dynamic analysis in ROB procedure can be conducted by MC^F implementations in order to construct exact mixed posterior distributions of the time-varying β_t . More precisely, a variant of the Gibbs sampler approach can be assessed in this framework by making use of the Kalman filter,²⁹ so it only requires knowledge of the conditional posterior of $(\beta_1, \dots, \beta_T | Y^T, \ddot{\phi}_{-\beta_t})$. The result would be a Bayesian Model Averaging. Thus, the posterior distributions for $\ddot{\phi}$ can be computed as:

$$\beta_t | Y^T, \ddot{\phi}_{-\beta_t} \sim N(\tilde{\beta}_{t|T}, \ddot{R}_{t|T}) \quad \text{or} \quad p(\beta_t | Y^T, \phi_{-\beta_t}) = N(\tilde{\beta}_{t|T}, \ddot{R}_{t|T}) \quad (65)$$

$$\Sigma\ddot{e} | Y^T, \ddot{\phi}_{-\Sigma\ddot{e}} \sim iW(\ddot{z}_1, \ddot{\beta}_1) \quad \text{or} \quad p(\Sigma\ddot{e} | Y^T, \ddot{\phi}_{-\Sigma\ddot{e}}) = iW(\ddot{z}_1, \ddot{\beta}_1) \quad (66)$$

$$\ddot{h}_{\ddot{f}t} | Y^T, \phi_{-\ddot{h}_{\ddot{f}t}} \sim IG\left\{ \frac{\ddot{\omega}_{\ddot{f}}}{2}, \frac{\ddot{S}}{2} \right\} \quad \text{or} \quad p(\ddot{h}_{\ddot{f}t} | Y^T, \phi_{-\ddot{h}_{\ddot{f}t}}) = IG\left\{ \frac{\ddot{\omega}_{\ddot{f}}}{2}, \frac{\ddot{S}}{2} \right\} \quad (67)$$

where $\tilde{\beta}_{t|T}$ and $\ddot{R}_{t|T}$ denote the smoothed one-period-ahead forecasts of β_t and of the variance-covariance matrix of the forecast error, $\ddot{z}_1 = \ddot{z}_1 + T$, $\ddot{\beta}_1 = \ddot{\beta}_1 + \sum_t (Y_t - (\ddot{\Theta}_t\ddot{G})\beta_t)$, $\ddot{\omega}_{\ddot{f}} = \ddot{k} + \ddot{\omega}_0$, and $\ddot{S} = \ddot{S}_0 + \sum_t (\beta_{t\ddot{f}} - \beta_{t-1\ddot{f}})'(\beta_{t\ddot{f}} - \beta_{t-1\ddot{f}})$, with $\beta_{t\ddot{f}}$ denoting the $\ddot{f}^{th}$ subvector of β_t and $\ddot{f}$ referring to the factors described in Eq. (57).

The choice of correlating $\ddot{E}_t$ and $\ddot{u}_t$ enables conjugation between the prior and the likelihood, dealing with structural model uncertainty under misspecified dynamics and endogeneity issues, and greatly simplifies the computation of posterior distributions. Finally, since the fit improves when $\sigma^2 \rightarrow 0$, the model in Eq. (59) presents an exact factorization of $\ddot{y}_t$.

In this empirical application, the hyperparameters are all known.³⁰ More precisely, collecting them in a vector δ , where $\delta = (\ddot{z}_1, \ddot{\omega}_0, \ddot{S}_0, \ddot{\beta}_1)$, they are treated as fixed and are either obtained from the data to tune the prior to the specific applications (this is the case for $\ddot{\omega}_0$ and $\ddot{\beta}_1$) or selected a priori to produce relatively loose priors (this is the case for $\ddot{z}_1$ and $\ddot{S}_0$). The values are: $\ddot{z}_1 = NM = 36$, $\ddot{\omega}_0 = 10^0$, $\ddot{S}_0 = 1.0$, and $\ddot{\beta}_1 = \text{diag}(\ddot{Q}_{11}, \dots, \ddot{Q}_{1N})$, where $\ddot{Q}_{1i}$ is the estimated covariance matrix for each i .

Given the above specifications, posterior distributions can be simply obtained by recursively calculating the first two moments of the posterior of the parameters. I use a set of 1,000 until 5,000 draws and find that convergence is obtained at about 3,000 draws.³¹ The total number of draws has been 2,000 + 3,000 = 5,000, which corresponds to the sum of the final number of draws to discard and save, respectively. A total of 3,000 retained replications have been used to conduct posterior inference at each t . Here, the outcomes absorb the conditional forecasts computed for a time frame of 9 quarters (2.25 years).

²⁹See, e.g., Chib and Greenberg (1995).

³⁰Own computations.

³¹The convergence has been found by amounting to about 1.5 draws per regression parameter.

Finally, once the (conditional) marginal likelihood in (64) is obtained for any model M_k , further shrinkage need to be conducted until to obtain the exact and final solution with higher log Bayes factor (Section 2.2.2). However, the sum of PMPs—according to Eqs. (65), (66), and (67)—refers to mixture density of conjugate priors and the tuning parameters τ and $\hat{\tau}$ in Eqs. (12) and (15) are dropped by focusing on the only hyperparameters collected in the vector δ . Thus, I am able to discriminate across models with different indices (M_k) by directly choosing a pool of *top best* submodels (M_{ξ}) that contain the only regression parameters with posterior means ($\hat{\beta}_{jt}$'s) different from zero. In this way, I can jointly deal with overestimation of effect sizes (or individual contributions) and structured model uncertainty (implicit in the procedure by construction) without loss of estimation efficiency. The log Bayes Factor in (16) is computed as:

$$lBF_{k,\xi} = \log \left(\frac{L(Y_T|M_k)}{L(Y_T|M_{\xi})} \right) \quad (68)$$

where Y_T denotes the data and $L(Y_T|M_{\xi})$ refers to the (conditional) likelihood function³² conducted on submodels M_{ξ} by MC^F implementations. Support for the presence of cross-country interdependence and endogeneity issues is obtained, according to the indicator $\ddot{\chi}_{3t}\hat{\beta}_{3t}$, by comparing the marginal likelihoods of the unrestricted model (M_k) and a vector of the submodels (M_{ξ}).

5.1. Empirical results

In Fig. 2, I show the cross-country indicators in standard deviation from the historical average of annual growth rates, spanning the period 1999q1 to 2016q1. They account for a pool of *top best* submodels (M_{ξ}) that contain the only regression parameters with posterior means ($\hat{\beta}_{jt}$'s) different from zero. More precisely, I obtain p-values setting to 0.026, 0.024, and 0.016 for $\hat{\beta}_{1t}$, $\hat{\beta}_{2t}$, and $\hat{\beta}_{3t}$, respectively. The higher log Bayes Factor is found for the last indicator (Fig. 2(c)), in which I consider countries' dynamics in total economy. It amounts to 5.94 for $\hat{\beta}_{3t}$, 4.17 for $\hat{\beta}_{2t}$, and 3.96 for $\hat{\beta}_{1t}$.

In Fig. 3, I draw the corresponding density forecast combinations for each of the cross-country indicators ($\ddot{\chi}_{jt}\hat{\beta}_{jt}$). They correspond to conditional projections of each variable drawn in the *top best* submodels (M_{ξ}) for output growth (change of productivity) and have been computed according to posterior distributions derived in Eqs. (65), (66), and (67). The yellow and red curves denote the 95% confidence bands, and the purple and blue curves denote the conditional³³ and unconditional³⁴ projection of output growth for each country in the period from 1999q1 to 2016q1, respectively. Here, the data used for the estimation spans the period 1999q1 to 2008q1 and the training sample for the hierarchical CIPs spans the period 2008q2 to 2013q4 in order to emphasize endogeneity issues by investigating (relevant) misspecified dynamics during, e.g., triggering events.

The results find confirmation of the preliminary results obtained in Pacifico (2019b)'s analysis. More precisely, focusing on the cross-country indicators (Fig. 2), three main considerations are in order. First, stronger cross-country heterogeneity among sectors is found in real economy because of larger economic linkages among sectors (structural uncertainty due to misspecified dynamics

³²The result function would be $L(Y_T|M_{\xi}) = \int \mathcal{E}(Y_T|\ddot{\phi}_1, M_{\xi})p(\ddot{\phi}_1|M_{\xi})d\ddot{\phi}_1$, where $p(\ddot{\phi}_1|M_{\xi})$ is the mixture density priors for submodels M_{ξ} .

³³The conditional projection for output growth is the one that the model would have obtained over the same period conditionally on the actual path of unexpected dynamics for that period.

³⁴The unconditional projection is the one that the model would obtain for output growth for that period only on the basis of historical information, and it is consistent with a model-based forecast path for the other variables.

(a) $\check{\chi}_{1t}\hat{\beta}_{1t}$ - Real Dimension (b) $\check{\chi}_{2t}\hat{\beta}_{2t}$ - Financial Dimension (c) $\check{\chi}_{3t}\hat{\beta}_{3t}$ - Total Economy

Figure 2. Cross-country variable-specific and common indicators are shown in standard deviation, by means of box plots, from the historical average of annual growth rates, spanning the period 1999q1 to 2016q1. They account for $\check{\chi}_{1t}\hat{\beta}_{1t}$ (plot a), $\check{\chi}_{2t}\hat{\beta}_{2t}$ (plot b), and $\check{\chi}_{3t}\hat{\beta}_{3t}$ (plot c) cross-country indicators, where $\hat{\beta}_{it}$'s are posterior means.

of the economy). Second, cross-country co-movements among sectors matter more in the financial dimension because of more intense policy implications (model uncertainty). Third, relevant outliers among countries matter in total economy, but with some common behaviors because of relevant cross-country interdependencies (endogeneity issues).

The above findings also hold by observing the corresponding density forecast combinations for each of the cross-country indicators (Fig. 3). Indeed, forecasts by accounting for total economy (third plot) tend to show a similar evolution to the ones according to financial economy (second plot) because of highly stringent cross-country interdependencies. On the contrary, forecasts in real economy tend to diverge (larger projections) because of stronger cross-country heterogeneity (first plot). Overall, the persistent divergence observed in the real and financial dimensions and even more in total economy highlights the relevant impact of significant endogeneity issues and misspecified dynamics among countries and sectors, and thus the need for forecasters and policy-makers to account for additional economic-institutional effects when formulating policies and forecasting in a large set of multicountry data.³⁵

³⁵See, for instance, Pacifico (2019a, 2019b), Groen et al. (2013), Koop and Korobilis (2012), and Raftery et al. (2010) for interesting implementations when studying macroeconomic-financial linkages in multicountry setups.

Figure 3. Density forecast combinations for output growth for each of the cross-country indicators ($\hat{\gamma}_{it}, \hat{\beta}_{it}$) are drawn. They correspond to conditional and unconditional projections of each variable drawn in the *top best* submodels (M_t^c) according to posterior distributions derived in Eqs. (65), (66), and (67).

6. Concluding remarks

This paper suggests and develops a computational approach, Robust Open Bayesian procedure, in order to deal with open recurring problems when evaluating BMA for multiple linear regression models. They include: (i) overfitting and thus overestimation of effect sizes since more complex models will always provide a somewhat better fit to the data than simpler models, where the “complexity” stands (for example) for the number of unknown parameters; (ii) model uncertainty in hierarchical normal mixture models; (iii) the choice and specification of prior distributions over multiple competing models; and (iv) variable selection and outlier detection when implementing BMA methodology for multiple models classes when the number of parameters and thus the number of possible models is high dimensional.

The main thrust to the potential of ROB procedure is the use of Conjugate Informative Proper priors for each promising model in a hierarchical and structural framework. This is an important

issue because, in multiple model class, common parameters can change meaning from one model to another, so that prior distributions must change in a corresponding fashion and weight more according to model size. The ROB procedure consists in two stages. First, a fully enumerated Markov Chain Monte Carlo approach is used to move through the model space and the parameter space at the same time in order to obtain a reduced set containing *best* (possible) model solutions that mainly explain and thus fit the data. Second, a further shrinkage is conducted in order to obtain a smallest final subset of *top best* submodels containing the only *significant* solutions. Finally, the regression model with higher Bayes Factor will be the final solution containing a potential subset of predictors with higher *significant* overall F value and sufficiently *strong* $\bar{R}^2$ measure. The performance and the usefulness of the procedure is debated by means of two examples: an empirical application to the criminal behavior in U.S. states in 1960, by replying the analysis of Ehrlich (1973), and a simulated experiment, by following the format of George and McCulloch (1993). Here, discussion and relationships with other works have been displayed.

In this study, I also extend and implement the methodology in order to deal with endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems, when accounting for dynamics of the economy in either time-invariant moderate data or time-varying high dimensional multicountry data. The former is obtained by building on and extending the framework of Lenkoski et al. (2014), who combine the BMA approach to regression variable uncertainty and the TSLS procedure. More precisely, I derive Conjugate Informative Proper priors for all the endogenous covariates in order to model uncertainty and investigate endogeneity issues at both the instrument and covariate levels in the ROB procedure. Then, the validity of instruments in the TSLS-ROB procedure is addressed by constructing a Bayesian test of the identification restrictions based on model averaged posterior predictive p-values. In a context of dynamic multicountry setups, I build on Pacifico (2019b) and estimate a simplified version of the Structural Panel Bayesian VAR by defining hierarchical prior specification strategy and MC^F implementations in order to extend variable selection procedure to a wide array of candidate models and thus candidate density forecast combinations. Bayesian methods are used to reduce the dimensionality of the model, structure the time variations, and evaluate issues of endogeneity. In this way, policies designed to protect the economy against the worst-case consequences of misspecified dynamics are less aggressive and good approximations of the estimated rule.

Acknowledgments

I gratefully thank the two anonymous referees for their useful suggestions.

References

- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy* 76(2): 169–217. doi:[10.1086/259394](https://doi.org/10.1086/259394)
- Berger, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference* 25(3):303–328. doi:[10.1016/0378-3758\(90\)90079-A](https://doi.org/10.1016/0378-3758(90)90079-A)
- Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., Dasgupta, A., Gustafson, P., Wasserman, L., Kadane, J. B., Srinivasan, C., Lavine, M., O'Hagan, A., Polasek, W., Robert, C. P., Goutis, C., Ruggeri, F., Salinetti, G., Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test* 3(1):5–124. doi:[10.1007/BF02562676](https://doi.org/10.1007/BF02562676)
- Bhattacharya, A., Chakraborty, A., Mallick, B. K. (2016). Fast sampling with gaussian scale-mixture priors in high-dimensional regression. *Biometrika* 103(4):985–991. doi:[10.1093/biomet/asw042](https://doi.org/10.1093/biomet/asw042)
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* 87(419):738–754. doi:[10.1080/01621459.1992.10475276](https://doi.org/10.1080/01621459.1992.10475276)

- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37(4):373–384. doi:[10.1080/00401706.1995.10484371](https://doi.org/10.1080/00401706.1995.10484371)
- Breiman, L., Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique* 60(3):291–319. doi:[10.2307/1403680](https://doi.org/10.2307/1403680)
- Brier, S. S., Fienberg, S. E. (1980). Recent econometric modeling of crime and punishment: Support for the deterrence hypothesis? *Evaluation Review* 4(2):147–191. doi:[10.1177/0193841X8000400201](https://doi.org/10.1177/0193841X8000400201)
- Canova, F., Ciccarelli, M. (2004). Forecasting and turning point predictions in a Bayesian panel var model. *Journal of Econometrics* 120(2):327–359. doi:[10.1016/S0304-4076\(03\)00216-1](https://doi.org/10.1016/S0304-4076(03)00216-1)
- Canova, F., Ciccarelli, M. (2009). Estimating multicountry var models. *International Economic Review* 50(3): 929–959. doi:[10.1111/j.1468-2354.2009.00554.x](https://doi.org/10.1111/j.1468-2354.2009.00554.x)
- Canova, F., Ciccarelli, M., Ortega, E. (2007). Similarities and convergence in G7 cycles. *Journal of Monetary Economics* 54(3):850–878. doi:[10.1016/j.jmoneco.2005.10.022](https://doi.org/10.1016/j.jmoneco.2005.10.022)
- Canova, F., Ciccarelli, M., Ortega, E. (2012). Do institutional changes affect business cycles? *Journal of Economic Dynamics and Control* 36(10):1520–1533. doi:[10.1016/j.jedc.2012.03.017](https://doi.org/10.1016/j.jedc.2012.03.017)
- Chib, S., Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician* 49(4): 327–335. doi:[10.2307/2684568](https://doi.org/10.2307/2684568)
- Ciccarelli, M., Ortega, E., Valderrama, M. T. (2018). Commonalities and cross-country spillovers in macroeconomic-financial linkages. *Journal of Macroeconomics* 16(1):231–275.
- Clyde, M. (1999). Bayesian model averaging and model search strategies. *Bayesian Statistics* 6:157–185.
- Clyde, M., George, E. I. (2004). Model uncertainty. *Statistical Science* 19(1):81–94. doi:[10.1214/0883423040000000035](https://doi.org/10.1214/0883423040000000035)
- Ehrlich, I. (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy* 81(3):521–565. doi:[10.1086/260058](https://doi.org/10.1086/260058)
- George, E. I., McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88(423):881–889. doi:[10.1080/01621459.1993.10476353](https://doi.org/10.1080/01621459.1993.10476353)
- George, E. I., McCulloch, R. E. (1997). Approaches for bayesian model selection. *Statistica Sinica* 7(2):339–373.
- Groen, J. J. J., Paap, R., Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics* 31(1):29–44. doi:[10.1080/07350015.2012.727718](https://doi.org/10.1080/07350015.2012.727718)
- Hansen, B. E. (2007). Notes and comments: Least squares model averaging. *Econometrica* 75(4):1175–1189. doi:[10.1111/j.1468-0262.2007.00785.x](https://doi.org/10.1111/j.1468-0262.2007.00785.x)
- Kadiyala, R. K., Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian var models. *Journal of Applied Econometrics* 12(2):99–132. doi:[10.1002/\(SICI\)1099-1255\(199703\)12:2<99::AID-JAE429>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1099-1255(199703)12:2<99::AID-JAE429>3.0.CO;2-A)
- Kass, R. E., Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* 90(430):773–795. doi:[10.1080/01621459.1995.10476572](https://doi.org/10.1080/01621459.1995.10476572)
- Kleibergen, F., Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics* 114(1):29–72. doi:[10.1016/S0304-4076\(02\)00219-1](https://doi.org/10.1016/S0304-4076(02)00219-1)
- Koop, G., Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review* 53(3):867–886. doi:[10.1111/j.1468-2354.2012.00704.x](https://doi.org/10.1111/j.1468-2354.2012.00704.x)
- Lenkoski, A., Eicher, T. S., Raftery, A. E. (2014). Two-stage Bayesian model averaging in endogenous variable models. *Econometric Reviews* 33(1–4):122–124. doi:[10.1080/07474938.2013.807150](https://doi.org/10.1080/07474938.2013.807150)
- Madigan, D., Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89(428):1535–1546. doi:[10.1080/01621459.1994.10476894](https://doi.org/10.1080/01621459.1994.10476894)
- Madigan, D., York, J., Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 63(2):215–232. doi:[10.2307/1403615](https://doi.org/10.2307/1403615)
- Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society: Series A* 147(3):389–425. doi:[10.2307/2981576](https://doi.org/10.2307/2981576)
- Mitchell, T. J., Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83(404):1023–1032. doi:[10.1080/01621459.1988.10478694](https://doi.org/10.1080/01621459.1988.10478694)
- Pacifico, A. (2019a). International co-movements and business cycles synchronization across advanced economies: A SPBVAR evidence. *International Journal of Statistics and Probability* 8(4):68–85. doi:[10.5539/ijsp.v8n4p68](https://doi.org/10.5539/ijsp.v8n4p68)
- Pacifico, A. (2019b). Structural panel Bayesian var model to deal with model misspecification and unobserved heterogeneity problems. *Econometrics* 7(1):8–24. doi:[10.3390/econometrics7010008](https://doi.org/10.3390/econometrics7010008)
- Raftery, A., Karny, M., Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* 52(1):52–66. doi:[10.1198/TECH.2009.08104](https://doi.org/10.1198/TECH.2009.08104)
- Raftery, A. E., Madigan, D. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89(428):1535–1546. doi:[10.1080/01621459.1994.10476894](https://doi.org/10.1080/01621459.1994.10476894)
- Raftery, A. E., Madigan, D., Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437):179–191. doi:[10.1080/01621459.1997.10473615](https://doi.org/10.1080/01621459.1997.10473615)

- Raftery, A. E., Madigan, D., Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 6:323–349.
- Stigler, G. J. (1970). The optimum enforcement of laws. *Journal of Political Economy* 78(3):526–536. doi:[10.1086/259646](https://doi.org/10.1086/259646)
- Vandaele, W. (1987). Participation in illegitimate activities: Ehrlich revisited, 1960. In *Deterrence and Incapacitation*, Ann Arbor: Springer, pp. 270–335.
- Weisberg, S. (2005). *Applied Linear Regression*. Hoboken, NJ: John Wiley & Sons, Inc., 3rd ed.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* 81(394):446–451. doi:[10.1080/01621459.1986.10478289](https://doi.org/10.1080/01621459.1986.10478289)