Dynamic variable selection in dynamic logistic regression: An application to Internet subscription

Andrés Ramírez-Hassan*

September 11, 2018

Abstract

We extend the dynamic model averaging framework for dynamic logistic regression proposed by McCormick et al. (2012) to incorporate variable selection. This method of accommodating uncertainty regarding predictors is particularly appealing in scenarios where relevant predictors change through time, and there are potentially many of them, as a consequence, the computational burden is high. Simulation experiments demonstrate that our greedy variable selection strategy works well in identifying the relevant regressors. We apply our algorithm to uncover the determinants of Internet subscription in Medellín (Colombia) among 18 potential factors, and thus 262,144 potential models. Our results suggest that subscription to pay TV, household members studying, years of education and number of household members are positively associated with Internet subscription.

JEL Classification: C11, C15, L86
Keywords: Bayes factor, dynamic model averaging, Internet subscription, logistic model, MCMC, variable selection.

*Department of Economics, School of Economics and Finance, Universidad EAFIT, Medellín, Colombia. email: aramir21@eafit.edu.co

1 Introduction

Econometric inference should take into account model uncertainty (Hansen, 2005). In particular, uncertainty regarding regressors can be crucial in settings where the number of potential regressors is very large (Fernandez et al., 2001), as the probability of finding a statistically significant relationship is substantially increased just by chance (Yeung et al., 2005). Bayesian model averaging (BMA) introduces model uncertainty founded on statistical theory following the rules of probability, and its prediction minimizes convex loss functions (Hoeting et al., 1999; Raftery and Zheng, 2003). In contrast to common model selection techniques, such as the Akaike and Bayesian information criteria, the BMA strategy avoids the choice of a single model as well as the underestimation of the uncertainty that derives from that (Hoeting et al., 1999). In addition, it is well known that model averaging improves forecasts (Madigan et al., 1994; Timmermann, 2006). In discrete response models, which is our focus here, BMA predictive distributions also minimize the posterior expected loss under logarithmic scoring (Madigan and Raftery, 1994; Hoeting et al., 1999; Raftery and Zheng, 2003).

These characteristics make BMA a strong statistical tool. Although BMA enjoys a long tradition in statistics (Leamer, 1978), its application in economics has only recently come into its own (Fernandez et al., 2001; Sala-i Martin et al., 2004; Eicher et al., 2012; Moral-Benito, 2013a; Jetter and Ramírez Hassan, 2015). Moral-Benito (2013b) provides a detailed survey on the use of BMA methods in economics.

However, model averaging in situations in which the number of regressors is large is computationally burdensome, with K possible regressors implying 2^K models. For instance, $K = 20$ implies more than one million models.

We propose an evolutionary variable selection heuristic algorithm (greedy algorithm) based on models with the highest posterior probabilities that substantially reduces the computational burden due to not visiting every point in the model space. Our strategy is designed for dynamic model averaging, and naturally extends the framework proposed by McCormick et al. (2012), where the set of models is fixed.

Our variable selection is an approach that produces a solution to variable selection uncertainty in a computationally efficient manner. The method adopts an algorithm based on the Bayes factor, which is founded on strong decision-theoretic arguments, and avoids *ad hoc* selection of significance levels that do not take into account the implicit balance of losses (Jeffreys, 1961; Bernardo and Smith, 1994; Benjamin et al., 2017). In addition, the Bayes factor evaluates the evidence in favor of a null hypothesis relative to an explicit alternative (Kass and Raftery, 1995). Moreover, it minimizes the weighted sum of type I and type II error probabilities, avoiding the use of a fixed type I error irrespective of the sample size (DeGroot, 1975; Pericchi and Pereira, 2016). It has been proved in many settings that the Bayes factor is consistent (Chib and Kuffner, 2016), that is, it identifies the true model under the true probability distribution. Finally, Bayes factors follow the rules of probability, and, as a consequence, are a coherent basis on which to build a model selection strategy (Lindley, 2000).

We find in our simulations exercises that we are able to identifying the best set of relevant *contemporaneous* variables of the *data generating process* using our variable selection strategy in conjunction with the guidelines of Kass and Raftery (1995).[1]

We use our greedy algorithm to uncover the determinants of Internet subscription in Medellín (Colombia). As Internet has taken unforeseen dimensions, determinants of its adoption is gaining relevance. The fact that empirical research shows positive effects of Internet use in aspects such as education (Agarwal and Day, 1998; Becker et al., 2010), income (Schreyer, 2000; Harris, 1998) and information dissemination (Allen, 1977; Davis, 1989), illustrates the importance of Internet adoption to foster development in societies. However, to uncover the determinants of Internet adoption has not obtained the attention that deserves. But if Internet use generates benefits in the development of a society, then it is useful to understand what determines Internet adoption in the first place. For example, if the truly predictors lie in the range of policies (such

[1]Other two sensible criteria to select relevant variables are the median probability model (Barbieri and Berger, 2004) and the highest posterior probability model (Clyde and George, 2004). The former minimizes a quadratic predictive loss, and it is the model composed by variables that have a posterior inclusion probability greater than 0.5. The latter is the model with the highest posterior model probability, and minimizes a 0–1 loss function for correct selection.

as, education or social programs) then the level of Internet adoption can be affected. But if society–specific conditions (such as, gender or age) play the most important role, little remains to be done for policymakers.

We find that the number of determinants in Internet subscription seems to vary between 7 and 24 (see Table 5). We collect information for 18 regressors, which implies a total of 262,144 different potential models. Thus, this paper constitutes a step toward understanding the determinants of Internet subscription.

We implement our variable selection strategy with 15,000 iterations, and as the binary processes are inherently dynamic, we use dynamic variable selection (DVS) for dynamic logistic regressions with the best 25 models. We find that subscription to pay TV, household members studying, years of education, and number of household members have a positive effect on Internet subscription.

After this introduction, the paper is organized as follows: in Section 2 we describe the methodology, and give an heuristic proof why our dynamic selection approach works; In Section 3 some simulation exercises are provided in order to show the performance of our algorithm compared with standard alternatives; Section 4 shows the outcomes of our application, Internet subscription. Section 5 give some concluding remarks.

2 Methodology

The starting point is a dynamic linear logistic model. In particular, y_t is a *Bernoulli* response that is related to a set of covariates, $\boldsymbol{x}_t' = (x_{t1}, x_{t2}, \ldots, x_{tK})$, using the *logit* link.

$$logit(p_t) = \ln\left(\frac{p_t}{1-p_t}\right) = \boldsymbol{x}_t'\boldsymbol{\beta}_t, \tag{1}$$

where $p_t \equiv \Lambda(\boldsymbol{x}_t'\boldsymbol{\beta}_t) = P[Y_t = 1|\boldsymbol{x}_t'\boldsymbol{\beta}_t] = \frac{\exp(\boldsymbol{x}_t'\boldsymbol{\beta}_t)}{1+\exp(\boldsymbol{x}_t'\boldsymbol{\beta}_t)}$, and $\boldsymbol{\beta}' = (\beta_{t1}, \beta_{t2}, \ldots, \beta_{tK})$ is a vector of coefficients to estimate at time $t = 1, 2, \ldots, T$.

This model is specified by a normal initial distribution for the K-dimensional state vector at time $t = 0$, $\boldsymbol{\beta}_0 \sim \mathcal{N}_K(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{C}}_0)$, where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{C}}_0$ emerge from a Maximum Likelihood estimation of a logistic model using an initial sample training.

Assuming $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{w}_t$, where $\boldsymbol{w}_t \sim \mathcal{N}_K(\boldsymbol{0}, \boldsymbol{W}_t)$, $\pi(\boldsymbol{\beta}_{t-1}|\boldsymbol{y}_{t-1:1}) \sim \mathcal{N}_K(\hat{\boldsymbol{\beta}}_{t-1}, \hat{\boldsymbol{C}}_{t-1})$, and the prediction equation is $\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1}) \sim \mathcal{N}_K(\hat{\boldsymbol{\beta}}_{t-1}, \boldsymbol{R}_t)$, where $\boldsymbol{R}_t = \hat{\boldsymbol{C}}_{t-1} + \boldsymbol{W}_t$ (Petris et al., 2007). However, McCormick et al. (2012) define the covariance matrix of the prediction equation as $\boldsymbol{R}_t = \hat{\boldsymbol{C}}_{t-1}/\lambda$, where λ is slightly less than one, and is called the forgetting parameter.[2] This parameter allows incorporating more information from past time periods when the process is stable.

Following the updating property of Bayesian inference, we have that $\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1}) \propto f(\boldsymbol{y}_{t-1:1}|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t)$ where $f(\boldsymbol{y}_{t-1:1}|\boldsymbol{\beta}_t)$ is the likelihood function, and $\pi(\boldsymbol{\beta}_t)$ is the prior distribution of the state vector. Therefore,

$$\begin{aligned}
\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t:1}) &\propto f(\boldsymbol{y}_{t:1}|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t) \\
&= f(y_t|\boldsymbol{y}_{t-1:1}, \boldsymbol{\beta}_t) f(\boldsymbol{y}_{t-1:1}|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t) \\
&\propto f(y_t|\boldsymbol{y}_{t-1:1}, \boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1}) \\
&= f(y_t|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1}),
\end{aligned} \tag{2}$$

where the last line uses the fact that conditionally on $\boldsymbol{\beta}_t$, the $\boldsymbol{y}_{t:1}$'s are independent and y_t depends on $\boldsymbol{\beta}_t$ only.

We see from Equation (2) that the predictive density acts as the prior. However, the likelihood function of a logistic process does not allow for a standard expression for the posterior distribution. Therefore, strategies such as MCMC, filtering algorithms (sequential Monte Carlo), variational Bayes or Laplace based approximations can be used to tackle this situation. The three former implies a huge computational burden for each possible model, whereas the latter is less computational expensive, at the cost of an error term that is $O(T^{-1})$. Then, McCormick

[2] McCormick et al. (2012) set this parameter equal to 0.99 by default.

et al. (2012) approximates equation (2) using a normal distribution.

Applying the Newton–Raphson algorithm,

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t-1} - \left[\bigtriangledown^2 l(\hat{\boldsymbol{\beta}}_{t-1})\right]^{-1} \bigtriangledown l(\hat{\boldsymbol{\beta}}_{t-1}), \tag{3}$$

where $l(\boldsymbol{\beta}_t) = \ln\{f(y_t|\boldsymbol{\beta}_t)\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1})\} = y_t \ln p_t + (1-y_t)\ln(1-p_t) - \frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{R}_t| - \frac{1}{2}\left\{(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_{t-1})'\boldsymbol{R}_t^{-1}(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_{t-1})\right\}$. Then,

$$\bigtriangledown l(\hat{\boldsymbol{\beta}}_{t-1}) = \left(y_t - \frac{\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}{1+\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}\right)\boldsymbol{x}_t, \tag{4}$$

$$\bigtriangledown^2 l(\hat{\boldsymbol{\beta}}_{t-1}) = \boldsymbol{R}_t^{-1} + \frac{\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}{1+\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}\left(1 - \frac{\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}{1+\exp(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1})}\right)\boldsymbol{x}_t\boldsymbol{x}_t'. \tag{5}$$

In addition, the state variance is approximated by $\hat{\boldsymbol{C}}_{t-1} = \left[-\bigtriangledown^2 l(\hat{\boldsymbol{\beta}}_{t-1})\right]^{-1}$.

One econometric issue that has recently concerned researchers is variable selection uncertainty (Hansen, 2005). BMA provides an elegant solution to this issue. Thus, given K possible predictors, the number of possible models becomes 2^K. Write $\mathcal{M}_t = \left\{M_t^{(1)}, M_t^{(2)}, \dots, M_t^{(2^K)}\right\}$ for the set of indexes associated with model k at time t, such that $M_t^{(k)}$ indicates that the process is driven by model $M^{(k)}$ at time t, where each model depends on a vector of parameters $\boldsymbol{\beta}_t^{(k)}$, $k = 1, 2, \dots, 2^K$. Now the state space at each time consists of the pair $(M_t^{(k)}, \boldsymbol{\beta}_t^{(k)})$, and using standard probabilistic rules,

$$\pi(\boldsymbol{\beta}_t|\boldsymbol{y}_{t-1:1}) = \sum_{k=1}^{2^K} \pi(\boldsymbol{\beta}_t^{(k)}|\boldsymbol{y}_{t-1:1}, M_t^{(k)})\pi(M_t^{(k)}|\boldsymbol{y}_{t-1:1}), \tag{6}$$

where $\pi(\boldsymbol{\beta}_t^{(k)}|\boldsymbol{y}_{t-1:1}, M_t^{(k)})$ is the predictive distribution for model k and

$$\pi(M_t^{(k)}|\boldsymbol{y}_{t-1:1}) = \frac{\left[\pi(M_{t-1}^{(k)}|\boldsymbol{y}_{t-1:1})\right]^{\alpha}}{\sum_{l=1}^{2^K}\left[\pi(M_{t-1}^{(l)}|\boldsymbol{y}_{t-1:1})\right]^{\alpha}}, \tag{7}$$

where α is slightly less than one, and is known as the model forgetting factor. This strategy avoids calculating the $2^K \times 2^K$ transition matrix.

The posterior model probability is equal to

$$\pi(M_t^{(k)}|\boldsymbol{y}_{t:1}) = \frac{\pi(M_t^{(k)}|\boldsymbol{y}_{t-1:1})f(y_t|\boldsymbol{y}_{t-1:1}, M^{(k)})}{\sum_{l=1}^{2^K}\pi(M_t^{(l)}|\boldsymbol{y}_{t-1:1})f(y_t|\boldsymbol{y}_{t-1:1}, M^{(l)})}, \tag{8}$$

where

$$f(y_t|\boldsymbol{y}_{t-1:1}, M^{(k)}) = \int_{\boldsymbol{\beta}_t^{(k)}} f(y_t|\boldsymbol{\beta}_t^{(k)}, M^{(k)})\pi(\boldsymbol{\beta}_t^{(k)}|\boldsymbol{y}_{t-1:1}, M^{(k)})d\boldsymbol{\beta}_t^{(k)}. \tag{9}$$

A Laplace approximation is used to find this expression.

Observe that the numerator in expression 8 is the product of the predictive, which acts as the prior distribution, and the marginal likelihood.

We can calculate the posterior mean value and posterior variance,

$$E\left(\boldsymbol{\beta}_t\right) = \sum_{k=1}^{2^K}\pi(M_t^{(k)}|\boldsymbol{y}_{t:1})E\left(\boldsymbol{\beta}_t^{(k)}\right), \tag{10}$$

$$Var\left(\boldsymbol{\beta}_t\right) = \sum_{k=1}^{2^K}\pi(M_t^{(k)}|\boldsymbol{y}_{t:1})Var\left(\boldsymbol{\beta}_t^{(k)}\right) + \sum_{k=1}^{2^K}\pi(M_t^{(k)}|\boldsymbol{y}_{t:1})\left(E\left(\boldsymbol{\beta}_t^{(k)}\right) - E\left(\boldsymbol{\beta}_t\right)\right)^2, \tag{11}$$

where $E\left(\boldsymbol{\beta}_t^{(k)}\right)$ and $Var\left(\boldsymbol{\beta}_t^{(k)}\right)$ are the posterior mean value and variance associated with model k at time t, respectively.

Observe that equation 11 implies that posterior coefficient estimates increase their variance when there is heterogeneity regarding different models estimates. This is the way how BMA takes into account model uncertainty.

We follow Kass and Raftery (1995) to define the relevance of regressors. In particular, Kass and Raftery (1995) suggest that posterior inclusion probabilities (PIP) less than 0.5 is evidence against the regressor, $0.5 \leq PIP < 0.75$ is weak evidence, $0.75 \leq PIP < 0.95$ is positive evidence, $0.95 \leq PIP < 0.99$ is strong evidence, and $PIP \geq 0.99$ is very strong evidence. The PIP is given by

$$PIP(\boldsymbol{x}_{tq}) = \sum_{k=1}^{2^K} \pi(M_t^{(k)}|\boldsymbol{y}_{t:1}) \times I_q^{(k)}, \tag{12}$$

where $I_q^{(k)} = \left\{ \begin{array}{ccc} 1 & if & x_q \in M^{(k)} \\ & & \\ 0 & if & x_q \notin M^{(k)} \end{array} \right\}, q = 1, 2, \ldots, K.$

However, we should take into consideration that sorting through all possible combinations of variables implies two computational problems. First, the processing requirements are very time consuming, and second, the physical memory requirements increase substantially. Thus, we select a set of models characterized by a high posterior model probability. In particular, we propose a Markov chain Monte Carlo variable selection procedure in a dynamic setting. This procedure reduces enormously the computational burden of the algorithm, as well as the physical memory requirements.

In particular, we propose a heuristic algorithm based on evolutionary ideas where the best–fit model, among two competing models, survives (see Algorithm 1). Then, we evaluate the performance of the latter, and replace the less–fit models ($M^{(Min)}$) with the new models in the case of performance improvement. Our “fit” criterion is based on Bayes factors, which take into account model fit, and a penalty term associated with model complexity (number of regressors).

Acceptance of candidate models in our heuristic algorithm is based on a multiple testing approach using posterior model probabilities comparing candidate and less–fit models. Casella and Moreno (2006) proposed variable selection using this idea in a linear regression framework with intrinsic priors (Berger and Pericchi, 1996). Our proposal is also related to Markov chain

[3]Observe that there is implicitly in our algorithm a symmetric constant transition kernel $q(M^{(Min)} \rightarrow M^{(c)})$, such that $q(M^{(Min)} \rightarrow M^{(c)}) = q(M^{(c)} \rightarrow M^{(Min)}) = \frac{1}{2^K-(J-1)-1}$. We also impose that each model should have as minimum 1 regressor, this excludes the model without regressors.

Algorithm 1 Variable selection algorithm for a dynamic logistic regression.

1: *Initial set*: Set $\boldsymbol{X}_{J\times K}$, $[x_{jq}] = \{0,1\}$, 1 indicates predictor $q = 1,2,\dots,K$ is in model $j = 1,2,\dots,J$. We use a Bernoulli distribution with probability equal to 0.5 to build this initial set of models with the restriction that all models have to be different.

2: *Marginal likelihood*: Calculate the marginal likelihood for the set of J models in the last H periods,

$$m_H(M^{(j)}) = \prod_{t=T+1-H}^{T} \pi(M_t^{(j)}|\boldsymbol{y}_{t-1:1}) f(y_t|\boldsymbol{y}_{t-1:1}, M^{(j)}), \ j = 1,2,\dots,J.$$

3: *Minimum marginal likelihood*: Calculate $m_H(M^{(Min)}) = Min\left\{m_H(M^{(j)})\right\}$. This is the less-fit model.

4: *Candidate model*: Draw a candidate model, $M^{(c)}$ using a Bernoulli distribution with probability 0.5 of including each regressor under the restriction of no being equal to any of the J models in $\boldsymbol{X}$, except the less-fit model.[3]

5: *Candidate marginal likelihood*: Calculate $m_H(M^{(c)})$.

6: *Acceptance rate*: Calculate the acceptance rate $\alpha = Min\left\{PO_H^{c,Min}, 1\right\}$, where $PO_H^{c,Min} = \frac{m_H(M^{(c)})}{m_H(M^{(Min)})}$.

7: *Selection*: Draw $u \sim \mathcal{U}(0,1)$, if $\alpha \geq u$, $M^{(c)}$ replaces $M^{(Min)}$ in matrix $\boldsymbol{X}$, otherwise $\boldsymbol{X}$ is kept the same.

8: *Iteration*: Repeat steps 2 to 7, S times, $S << 2^K$.

Monte Carlo model composition (MC^3) (Madigan et al., 1995). However, we do not generate candidate models from neighbours of the actual model. It seems that in dynamic settings like us, random–walk kind proposals have a tendency to get stuck in local models; our simulation exercises suggest that MC^3 convergences slower than our proposal. There are other options to our approach such as reversible jump Markov chain Monte Carlo computation (Green, 1995), stochastic search variable selection (Spike and Slab) (George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005), and using nonlocal prior densities (Jhonson and Rossell, 2012). Future research should consider these sensible alternatives.

The acceptance probability in our algorithm is $\alpha = Min\left\{PO_H^{c,Min}, 1\right\}$, where[4]

[4]We work with the geometric mean of the marginal likelihoods for computational convenience. Observe that this does not affect the ordering of models. In addition, we consider an initial training sample.

$$
\begin{aligned}
PO_H^{k,j} &= \frac{m_H(M^{(k)})}{m_H(M^{(j)})} \\
&= \prod_{t=T+1-H}^{T} \frac{\pi(M_t^{(k)}|\boldsymbol{y}_{t:1})}{\pi(M_t^{(j)}|\boldsymbol{y}_{t:1})} \\
&= \prod_{t=T+1-H}^{T} \frac{\pi(M_t^{(k)}|\boldsymbol{y}_{t-1:1})f(y_t|\boldsymbol{y}_{t-1:1}, M^{(k)})}{\pi(M_t^{(j)}|\boldsymbol{y}_{t-1:1})f(y_t|\boldsymbol{y}_{t-1:1}, M^{(j)})} \\
&= \prod_{t=T+1-H}^{T} \frac{(\pi(M_{t-1}^{(k)}|\boldsymbol{y}_{t-1:1}))^{\alpha}f(y_t|\boldsymbol{y}_{t-1:1}, M^{(k)})}{(\pi(M_{t-1}^{(j)}|\boldsymbol{y}_{t-1:1}))^{\alpha}f(y_t|\boldsymbol{y}_{t-1:1}, M^{(j)})} \\
&= \prod_{t=0}^{T-H} \left(\frac{f(\boldsymbol{y}_{T-t:T+1-H-t}|\boldsymbol{y}_{T-H-t:1}, M^{(k)})}{f(\boldsymbol{y}_{T-t:T+1-H-t}|\boldsymbol{y}_{T-H-t:1}, M^{(j)})}\right)^{\alpha^t} \times \left(\frac{f(\boldsymbol{y}_{H-1:1}|M^{(k)})}{f(\boldsymbol{y}_{H-1:1}|M^{(j)})}\right)^{\alpha^{T+1-H}} \times \cdots \times \left(\frac{f(y_1|M^{(k)})}{f(y_1|M^{(j)})}\right)^{\alpha^{T-1}} \\
&= \prod_{t=0}^{T-H} \left(BF_{T-t:T+1-H-t}^{k,j}\right)^{\alpha^t} \times \left(BF_{H-1:1}^{k,j}\right)^{\alpha^{T+1-H}} \times \cdots \times \left(BF_1^{k,j}\right)^{\alpha^{T-1}},
\end{aligned} \tag{13}
$$

where $T \geq H$, second equality takes into account that in this dynamic setting the ratio between marginal likelihoods is equal to the ratio between posterior model probabilities. Third and fourth equalities use equations 8 and 7, respectively. Fifth equality uses the fact that $\pi(M_t|\boldsymbol{y}_{t:1}) \propto \prod_{l=0}^{t-2} f(y_{t-l}|\boldsymbol{y}_{t-1-l:1}, M) \times f(y_1|M)$ (the proportionality constant is the same for every model), $BF_{T-t:T+1+H-t}^{k,j}$ are conditional Bayes factors using information up to $\boldsymbol{y}_{T-H-t:1}$ (Dickey and Gunel, 1978), and $BF_{H-1:1}^{k,j}, \dots, BF_1^{k,j}$ are unconditional Bayes factors.

Thus our acceptance criterion depends on age-weighted product of sample-specific conditional Bayes factors, that is, *contemporaneous* information has more influence to make decisions regarding model choice; this is intuitively appealing (Raftery et al., 2010). It has been proved in many $i.i.d$ settings that the Bayes factor is pairwise consistent (see Chib and Kuffner (2016) for an excellent review).[5] Given $f^{(k)}(\boldsymbol{y}|\boldsymbol{\beta})$ and $\pi^{(k)}(\boldsymbol{\beta})$, the likelihood and prior distributions under model k ($M^{(k)}$), respectively, then Bayes factor almost sure pairwise consistency requires that when comparing models $M^{(k)}$ and $M^{(j)}$, the two following statements hold with probability one under the true sampling distribution ($P_{M^{(d.g.p)}}$ with density $p_{M^{(d.g.p)}}$),

[5] Jhonson and Rossell (2012) highlight the important difference between pairwise consistency, and model selection consistency. The latter requires consistency of a sequence of pairwise nested comparisons.

1. $BF_{t:1}^{k,j} \overset{p_{M^{(d.g.p)}}}{=\!\!\longrightarrow} \infty, T \to \infty$, when $M^{(k)}$ contains the true model ($M^{(d.g.p)}$).

2. $BF_{t:1}^{k,j} \overset{p_{M^{(d.g.p)}}}{\longrightarrow} 0, T \to \infty$, when $M^{(j)}$ contains the true model ($M^{(d.g.p)}$).

where $BF_{t:1}^{k,j} = \frac{\int f(\boldsymbol{y}_{t:1}|\boldsymbol{\beta}^{(k)}, M^{(k)})\pi(\boldsymbol{\beta}^{(k)}, M^{(k)})d\boldsymbol{\beta}^{(k)}}{\int f(\boldsymbol{y}_{t:1}|\boldsymbol{\beta}^{(j)}, M^{(j)})\pi(\boldsymbol{\beta}^{(j)}, M^{(j)})d\boldsymbol{\beta}^{(j)}}$.

In the case that the set of models does not contain the true model ($M^{(d.g.p)}$), the Bayes factor convergences to the "best" approximating model among those being considered. The definition of "best" depends on specific settings; for instance, Walker (2004a,b) considers the Kullback–Leibler (δ) property of prior distributions in a nonparametric setting, and Casella et al. (2009) consider a measure of fit based on residual sum of squares in linear regression models using intrinsic priors.

Most of the work regarding consistency of Bayes factors has been done assuming $i.i.d$ sampling. Recently, Chatterjee et al. (2018) show consistency of Bayes factors in a general set–up; this includes dependent data and misspecified settings. Their optimality criterion is based on the Kullback–Leibler divergence between the true sampling distribution, and the distribution generated by the model $M^{(k)}$, that is, given $d_{KL}(M^{(k)}) = \lim_{t\to\infty} \frac{1}{t} \int log \left\{ \frac{p_{t,M^{d.g.p}}}{p_{t,M^{(k)}}} \right\} dP_{t,M^{d.g.p}}$, Chatterjee et al. (2018) prove that $\lim_{t\to\infty} \frac{1}{t} log \left\{ BF_t^{k,j} \right\} = d_{KL}(M^{(j)}) - d_{KL}(M^{(k)})$.[6] The Bayes factor selects $M^{(k)}$ if $d_{KL}(M^{(k)}) < d_{KL}(M^{(j)})$, and selects $M^{(j)}$ if $d_{KL}(M^{(j)}) < d_{KL}(M^{(k)})$.[7] The proof of this result uses Chib and Kuffner (2016)'s approach, and the intuition behind is that there is a type of log likelihood ratio, which is bounded in probability ($O_p(1)$), the ratio of priors, which is also bounded ($O(1)$), as its effect is negligible for large samples, and the log ratio of marginal likelihoods, which converges to the K–L divergence exponentially fast.

We will now provide an heuristic explanation why our algorithm may work based on the previous statements.[8] Following same arguments as Shalizi (2009) and Chatterjee et al. (2018) in equation 13, we would expect that if $\exists t_0(\alpha) \to \infty$ such that $\alpha^{t_0(\alpha)+1} \approx 0$, $T \geq H + t_0$, then,

[6] A similar result is given by Walker (2004a) in a nonparametric $i.i.d$ setting.

[7] Observe that $d_{KL}(M^{(k)}) \geq 0$, $d_{KL}(M^{(k)}) = 0$ when $p_{t,M^{d.g.p}} = p_{t,M^{(k)}}$.

[8] A formal proof is beyond the scope of this paper.

$$
\begin{aligned}
\lim_{H\to\infty}\frac{1}{H}log\left\{PO_H^{k,j}\right\} &= \lim_{H\to\infty}\frac{1}{H}\left\{\sum_{t=0}^{T-H}\alpha^t log\left\{BF_{T-t:T+1-H-t}^{k,j}\right\}+\right.\\
&\qquad \left.\alpha^{T+1-H}log\left\{BF_{H-1}^{k,j}\right\}+\cdots+\alpha^{T-1}log\left\{BF_1^{k,j}\right\}\right\}\\
&= \lim_{H\to\infty}\frac{1}{H}\left\{\sum_{t=0}^{t_0(\alpha)}\alpha^t BF_{T-t:T+1-H-t}^{k,j}\right\}\\
&= \sum_{t=0}^{t_0(\alpha)}\alpha^t\left\{d_{KL}(M_{T-t:T+1-H-t}^{(j)})-d_{KL}(M_{T-t:T+1-H-t}^{(k)})\right\} \qquad (14)
\end{aligned}
$$

where $d_{KL}(M_{T-t:T+1-H-t}^{(k)})$ and $d_{KL}(M_{T-t:T+1-H-t}^{(j)})$ is the Kullback–Leibler divergence between the *data generating process* governing at $T-t$, and models k and j, respectively, using the conditional Bayes factor with information between $T-t$ and $T+1-H-t$.

Observe that $t_0(\alpha)$ depends on the forgetting parameter, $0<\alpha<1$. $\alpha\to 1$ implies $\alpha^t\to 0$ slowly, that is, past observations have a high impact at t. This can be problematic if the variables which drive the process are changing through time such that $d_{KL}(M_{T-t:T+1-H-t}^{(k)})-d_{KL}(M_{T-t:T+1-H-t}^{(j)})$ does not have a clear pattern, that is, the algorithm does not identify a "better" model. Then, $t_0(\alpha)$ should converge to ∞ faster to identify in a better way the model driven the process at t. However, $t_0(\alpha)$ converging to ∞ very fast or α very small implies that we cannot borrow strength from past observations to identify the *d.g.p.* To borrow strength from past observations is good under stable environments.

There are six parameters to define in algorithm 1, two forgetting parameters (λ and α), the training sample, number of models in $\boldsymbol{X}$ (J), the marginal likelihood bandwidth (H), and the number of iterations (S). We set the forgetting parameters equal to 0.99 as McCormick et al. (2012) set by default, and use a training sample equal to 50% of the sample in our simulation exercises. We follow Raftery et al. (1997), who found that the number of relevant models is normally less than 25, so we set $J=25$.

H depends on the stability of the process through time. If the model is very stable, that is, same variables are relevant always, then $T = H$ (this implies $t_0(\alpha) = 0$). In this scenario, it is computationally more efficient to perform static variable selection (SVS). On the other hand, if the process is unstable such that relevant variables change through time, we should select $H < T$ (this implies $t_0(\alpha) > 0$). Thus, our approach borrows strength from previous periods for estimation, but selects models based on recent performance. However, a small H generates instability in the variable selection process (increases variability in the algorithm), although it reduces variable selection bias. On the other hand, a large H induces variable selection bias, if there were changes regarding relevant variables in H span, but reduces model selection variability.

In practice, we do not know about the stability of the *data generating process*, but we can get some clues from preliminary exercises based on static variable selection techniques. Although, this may fail if the relevant variables change in short time periods. We find in simulation exercises that our proposal performs better to select *contemporaneous* relevant variables than model selection using static techniques when we do not have the oracle property of knowing when there is a change regarding relevant variables. Although future research should consider optimal bandwidth choice.

Regarding the number of iterations, it is common belief that it should be based on convergence criteria. However, there are two caveats here. First, it is important to note that the objective of variable selection is somewhat different from the objective of estimating a posterior distribution. Our main focus is to find good regressors from a huge space of potential combinations, and we are less interested in estimating all of the potential combination (Casella and Moreno, 2006). Therefore, it is also possible to find promising regressors even though one has explored a very small fraction of the model space (Brown et al., 2002). Second, if converge would be an issue in variable selection, it seems that there are no available diagnostics to determine whether the MCMC chain has either converged or explored a sufficient proportion of models in the model space (Jerrum and Sinclair, 1996; Ormerod et al., 2017). Thus, we propose to perform a sensitivity analysis regarding stability of PIP with number of iterations, and make decisions using the guidelines of Kass and Raftery (1995). For instance, we found in our simulation exercises that there is stability regarding variable selection after $S \geq 1,000$ (we show results using $S = 1,000$ and $S = 5,000$). However, this analysis should be case specific. For instance, we found in our application that stability is achieved after 10,000 iterations (we show outcomes using 15,000 iterations).

3 Simulation exercises

We evaluate the performance of Algorithm 1 to identify relevant *contemporaneous* variables of the *data generating process* under three different scenarios: Complete and incomplete uncorrelated regressors, and complete correlated regressors. In this setting, complete regressors means that all variables in the *data generating process* are in the set of potential explanatory variables. On the other hand, incomplete regressors means that all relevant variables in the *data generating process* are not in the set of potential explanatory variables.

Complete uncorrelated regressors

In this experiment the *data generating process* is given by

$$y_t^* = \beta_{t0} + \beta_{t1}x_{t1} + \beta_{t2}x_{t2} + \beta_{t3}x_{t3} + \beta_{t4}x_{t4} + \beta_{t5}x_{t5} + \beta_{t6}x_{t6} + \epsilon_t$$

$$y_t = \begin{Bmatrix} 1, & y_t^* > 0 \\ 0, & y_t^* \leq 0 \end{Bmatrix}$$

where $x_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $i = 1,2,3,4$, $x_{ti} \overset{i.i.d}{\sim} \mathcal{B}(0.7)$, $i = 5,6$ and $\epsilon_t \overset{i.i.d}{\sim} \mathcal{LG}(0,1)$, $t = 1,2,\ldots,5000$. We can see from Table 1 that the relevant variables in the *d.g.p.* change over time. In particular, x_{1t} and x_{4t} are not relevant in the second sub-sample, whereas x_{6t} is only relevant in the second sub-sample. In addition, β_{3t} increases 1/5000 each period from 1 to 2, and β_{5t} decreases one unit in the second sub-sample.

Our design matrix also includes 14 additional predictors that are no part of the *d.g.p.*, such that $x_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $i = 7,\ldots,16$, and $x_{ti} \overset{i.i.d}{\sim} \mathcal{B}(0.5)$, $i = 18,\ldots,20$. We set T=5,000,

Table 1: Location parameters of the latent structure

	β_{t0}	β_{t1}	β_{t2}	β_{t3}	β_{t4}	β_{t5}	β_{t6}
$t = 1, \ldots, 4000$	0.5	-2	-1	$1 + t/5000$	2.5	-1	0
$t = 4001, \ldots, 5000$	0.5	0	-1	$1 + t/5000$	0	-2	-1

and 20 regressors in our simulation exercises to have a setting similar to our Internet application.

Our setting implies that there are 2^{20} possible models, that is, 1,048,576 models. Our goal is to implement Algorithm 1 to identify the best *contemporaneous* potential regressors using the guidelines of Kass and Raftery (1995). Observe that this can be challenging in our setting, where in the second part of the sample two regressors are dropped (x_{t1} and x_{t4}), and a new regressor becomes relevant (x_{t6}).

We can see in Table 2 the posterior inclusion probabilities using three different approaches: Naive dynamic variable selection (Naive DVS, Algorithm 1), MC^3 DVS, and static variable selection (SVS). The former two procedures are set with 1,000 and 5,000 iterations, and both are based on the best 25 models (J=25).

It seems from Table 2 that there is strong evidence that variables $\boldsymbol{x}_2$, $\boldsymbol{x}_3$ and $\boldsymbol{x}_5$ are associated with $\boldsymbol{y}$ under different marginal likelihood bandwidths ($H = \{100, 200, 500, 2000\}$). However, $H = 100$ is no large enough to identify the relevance of $\boldsymbol{x}_6$ on $\boldsymbol{y}$. This is common to all procedures. However, setting $H = 200$ and $H = 500$, the dynamic model selection procedures identify that there is strong evidence that this variable is also associated with $\boldsymbol{y}$. Remember that this variable is relevant in the period 4,001–5,000, that is, the oracle H is 1,000. Using a marginal likelihood bandwidth equal to 2,000, SVS establishes that there is very strong evidence for variables $\boldsymbol{x}_1$ to $\boldsymbol{x}_6$, whereas DVS excludes $\boldsymbol{x}_6$ from this set. Thus, it seems from this experiment that DVS techniques are better to identify *contemporaneous* relevant regressors than common static approaches, but robustness regarding marginal likelihood bandwidth should be performed.

Regarding differences between naive DVS and MC^3 DVS, it seems that the former converges faster in this setting. We can observe in Table 2 that using 1,000 iterations, the PIP of irrelevant regressors associated with our naive DVS are always less than 0.75 ($H = \{100, 200, 500\}$), this means weak evidence or evidence against regressors $\boldsymbol{x}_1$, $\boldsymbol{x}_4$, $\boldsymbol{x}_7$ to $\boldsymbol{x}_{20}$. This is no the case when MC^3 DVS is used with 1,000. However, using 5,000 iterations, there are not meaningful differences between these algorithms.

Table 2: Posterior inclusion probability: Complete uncorrelated set.*

Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
$H = 100$																				
Naive DVS, $S = 1,000$	38.1	100	100	34.9	100	0.0	21.1	44.2	23.9	35.9	24.0	40.0	33.0	71.5	18.7	43.0	33.4	28.6	9.7	16.6
Naive DVS, $S = 5,000$	26	100	100	26.3	100	0.0	25.3	19.2	44.5	37.2	35.0	2.5	27.8	37.1	23.0	2.0	13.8	17.6	2.8	26.1
MC^3 DVS, $S = 1,000$	4.4	100	100	25.7	100	0.0	19.2	33.4	22.7	426	43.4	19.0	50.0	40.1	12.4	27.7	87.5	39.4	3.2	11.3
MC^3 DVS, $S = 5,000$	0.0	100	100	14.3	100	0.0	48.6	25.1	6.0	37.4	20.7	8.9	4.4	11.8	28.5	26.2	42.8	42.5	26.1	7.6
SVS	2.8	96.5	100	3.4	100	3.3	3.2	5.5	3.4	3.0	7.9	3.1	4.7	2.8	6.4	2.8	6.3	2.7	4.9	3.0
$H = 200$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	67.7	100	100	36.1	100	94.4	37.9	34.2	21.1	21.5	15.9	31.8	12.6	23.7	48.5	17.9	20	29.9	41.9	16.4
Naive DVS, $S = 5,000$	16.9	100	100	48.2	100	96.0	22.5	19.7	9.5	38.3	15.2	15.1	5.1	41.7	45.8	12.0	49.5	41.8	26.7	13.4
MC^3 DVS, $S = 1,000$	0.0	100	100	24.0	100	97.9	72.5	56.4	15.0	0.0	7.9	48.2	11.3	49.5	30.9	86.1	3.1	30.0	75.7	3.5
MC^3 DVS, $S = 5,000$	20.2	100	100	20.6	100	99.6	53.8	26.9	0.0	11.0	10.5	23.4	51.5	19.3	28.8	46.8	20.0	19.4	29.4	2.5
SVS	2.6	100	100	3.1	100	11.1	2.7	3.7	3.5	7.2	3	2.5	4.2	7.3	3.8	2.6	2.4	2.5	2.9	2.5
$H = 500$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	28.7	100	100	41.8	100	97.3	38.7	3.0	44.0	43.4	14.1	60.5	19.5	27.6	13.4	21.5	25.3	19.3	14.6	22.0
Naive DVS, $S = 5,000$	47.4	100	100	58.4	100	97.9	14.2	6.1	17.2	37.4	4.6	56.1	5.1	27.7	43.4	13.9	21.6	20.1	5.9	3.0
MC^3 DVS, $S = 1,000$	27.9	100	100	3.8	100	96.2	0.0	56.7	44.2	22.5	31.6	14.4	46.4	72.5	30.4	43.0	17.6	24.7	30.7	37.7
MC^3 DVS, $S = 5,000$	3.4	100	100	31.3	100	98.2	1.0	6.8	20.3	54.8	44.9	10.3	10.3	32.4	17.6	16.4	35.4	27.1	39.2	3.6
SVS	0.0	100	100	2.4	100	91.2	7.7	1.9	2.4	1.9	0.0	3.4	0.0	31.0	2.6	1.9	0.0	6.6	0.0	0.0
$H = 2,000$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	100	100	100	100	100	62.2	26.6	58	27.2	60.5	24.6	17.9	22.4	16.3	10.7	17.8	15.7	14.8	24.3	17.4
Naive DVS, $S = 5,000$	100	100	100	100	100	61.0	15.1	57.4	19.0	26.8	21.8	36.1	21.5	9.7	4.3	7.6	17.0	15.9	30.1	4.9
MC^3 DVS, $S = 1,000$	100	100	100	100	96.2	34.5	80.6	42.0	23.3	47.0	34.0	85.3	55.5	0.5	57.3	33.0	14.7	0.5	45.8	37.8
MC^3 DVS, $S = 5,000$	100	100	100	100	100	66.8	38.0	17.2	21.3	20.7	11.7	35.8	39.0	42.6	25.3	27.3	40.0	9.3	9.3	42.1
SVS	100	100	100	100	100	100	0.0	0.0	0.0	0.0	0.0	0.0	5.5	12.7	4.7	0.0	0.0	5.1	0.0	0.0

* We use the best 25 models in MC^3. SVS uses BIC approximation.

Incomplete uncorrelated regressors

We have same configuration as previous experiment, except that we omit $\boldsymbol{x}_5$ from the potential set of regressors, and introduce a new regressor $\boldsymbol{x}_{21}$. Note that $\boldsymbol{x}_5$ is part of the $d.g.p$, that is why there is an incomplete set of regressors in this setting.

We can see in Table 3 that there is strong evidence that $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are associated with $\boldsymbol{y}$ for all marginal likelihood bandwidths, and there is positive evidence ($H = 200$), and strong evidence ($H = 500$) that $\boldsymbol{x}_6$ is associated with $\boldsymbol{y}$ using dynamic variable selection approaches. This is no present in this setting using the static variable selection approach, which identifies $\boldsymbol{x}_6$ using $H = 2,000$. Observe that in this setting, DVS approaches identify the relevant *contemporaneous* variables that are part of the potential set of regressors despite the fact of omitting one relevant variable in this set. It seems also that naive DVS is slightly faster than MC^3 DVS to discard irrelevant variables.

Table 3: Posterior inclusion probability: Incomplete uncorrelated set.*

										$H=100$										
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}	x_{t21}
Naive DVS, $S=1,000$	12.7	100	100	18.4	80	8.8	5.4	5.4	19.3	9.3	87.4	4.1	7.3	85.1	1.5	20.2	4.9	17.3	4.4	74.3
Naive DVS, $S=5,000$	84.1	100	100	14.1	0.0	39.4	5.5	34.8	25.3	8.9	33.1	9.9	25.8	10.6	15.4	33.5	12.8	13.5	1.7	7.3
MC^3 DVS, $S=1,000$	86.0	100	100	26.3	0.0	23.3	58.2	38.6	22.7	13.6	22.8	22.3	66.1	30.9	11.3	15.2	2.2	26	8.5	71.1
MC^3 DVS, $S=5,000$	70.0	100	100	46.5	0.0	9.0	13.2	22.9	63	12.2	21.2	0.0	26.4	26.9	0.0	12.3	26.9	10.2	7.4	28.3
SVS	26.7	96.3	100	2.4	1.8	2.3	4.7	1.9	8.5	3.4	19.9	3.6	2.0	5.7	1.9	7.9	3.0	4.8	2.3	3.2
										$H=200$										
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}	x_{t21}
Naive DVS, $S=1,000$	40.0	100	100	18.2	90.0	25.0	35.5	35.9	53.9	13.4	52.6	35.1	54.2	27.7	25.1	56.0	2.4	30.0	42.0	2.5
Naive DVS, $S=5,000$	54.7	100	100	25.6	91.9	0.0	19.2	1.5	55.1	13.9	29.3	21.7	45.2	32.4	39.9	19.5	7.5	16.9	13.7	4.0
MC^3 DVS, $S=1,000$	51.0	100	100	47.9	92.7	92.4	11.7	0.0	51.4	67.0	39.1	14.9	52.6	31.6	53.7	9.1	52.5	17.1	18.4	1.0
MC^3 DVS, $S=5,000$	49.6	100	100	76.5	94.0	18.1	0.1	5.3	21.4	92.5	26.4	20.6	16.2	1.4	10.3	31.7	0.3	51.0	6.8	15.0
SVS	5.5	100	100	2.6	10.2	2.6	3.1	3.1	5.8	2.5	4.6	2.7	3.0	3.7	2.5	2.5	2.5	2.6	2.6	3.2
										$H=500$										
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}	x_{t21}
Naive DVS, $S=1,000$	23.3	100	100	39.5	81.8	27.8	32.0	49.1	3.5	7.2	13.0	36.1	55.0	20.7	44.6	29.4	42.5	20.4	0.0	22.5
Naive DVS, $S=5,000$	37.9	100	100	70.3	95.9	32.5	8.3	35.0	7.2	16.7	46.0	5.5	29.3	15	20.4	10.7	48.5	11.2	0.0	3.1
MC^3 DVS, $S=1,000$	41.6	100	100	35.9	97.9	5.7	14.0	34.6	10.2	0.0	27.7	44.1	29.7	10.6	52.8	0.0	1.1	63.4	9.3	62.7
MC^3 DVS, $S=5,000$	53.9	100	100	62.8	96.9	4.9	0.0	6.6	23.8	7.5	0.8	15.3	29.2	2.2	53.8	12.6	2.9	13.8	0.0	44.3
SVS	1.4	100	100	4.0	80.0	3.7	0.0	1.7	3.5	0.0	5.3	0.0	59.8	1.3	3.8	1.5	3.4	1.4	0.0	1.6
										$H=2,000$										
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}	x_{t21}
Naive DVS, $S=1,000$	100	100	100	100	66.7	59.3	54.2	38.5	23.9	20.4	26.4	20.1	47.0	30.7	15.3	26.1	38.7	49.1	19.4	41.3
Naive DVS, $S=5,000$	100	100	100	100	81.5	35.3	31.1	39.6	12.8	20.4	39.9	25.2	32.8	25.7	27.2	22.7	6.2	19.1	7.2	40.5
MC^3 DVS, $S=1,000$	100	100	100	100	36.1	77.9	60.5	53.9	38.5	42.9	59.5	54.2	77.3	19.0	50.2	26.5	74.5	53.4	33.2	51.7
MC^3 DVS, $S=5,000$	100	100	100	100	100	21.8	64.5	5.2	48.8	17.1	31.7	18.0	60.1	12.9	16.3	28.3	11.2	8.5	22.3	44.6
SVS	100	100	100	100	100	0.0	0.0	0.0	0.0	0.0	0.0	4.9	11.0	4.1	0.0	0.0	3.9	0.0	0.0	0.0

* We use the best 25 models in MC^3. SVS uses BIC approximation.

Complete correlated regressors

We have in this setting same structure as in the first simulation exercise, the difference is that $[x_{t1}\ x_{t2}\ x_{t3}\ x_{t7}\ x_{t8}\ x_{t9}]' \sim \mathcal{N}_6(\mathbf{0}, \mathbf{\Sigma})$, and $[x_{t4}\ x_{t5}\ x_{t6}\ x_{t10}\ x_{t11}\ x_{t12}]' \sim \mathcal{N}_6(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = [\sigma_{ij}]$, $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5,\ i \neq j$, and $x_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$, $i = 13, \ldots, 20$.

We can see in Table 4 that all variable selection procedures give very strong evidence that variables $\boldsymbol{x}_3$ and $\boldsymbol{x}_5$ are associated with $\boldsymbol{y}$. In addition, DVS approaches give strong evidence that $\boldsymbol{x}_2$ and $\boldsymbol{x}_6$ are associated with $\boldsymbol{y}$ using $H = 200$, SVS approach does not have this characteristic. However, MC^3 DVS is not as stable as naive DVS when looking at $H = 500$ results. At this stage, SVS gives very strong evidence that variables $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_5$ and $\boldsymbol{x}_6$ (relevant *contemporaneous* variables) are associated with $\boldsymbol{y}$. Using a bandwidth equal to 2,000, DVS approaches identify variables $\boldsymbol{x}_1$ to $\boldsymbol{x}_5$ as very strong predictors, and SVS adds $\boldsymbol{x}_6$ to this set. It seems from this exercise that naive DVS converges faster to identify relevant *contemporaneous* regressors.

4 Internet subscription

Although there is a consensus on the benefits of Internet use for aspects related to development (Agarwal and Day, 1998; Becker et al., 2010; Schreyer, 2000; Harris, 1998), the literature on the determinants of Internet adoption is less consistent. Previous works suggest various potential factors. Table 5 summarizes the most prominent results, but at the same time reveals a persistent econometric issue, variable selection uncertainty. Consider the number of included determinants, which varies between 7 and 24. Only education and income appear having a positive significant effect in all the studies, and it remains unclear which other variables should form part of a standard regression for determining Internet adoption. Just to illustrate the situation, there is not conclusive evidence of having children or gender on Internet adoption.

Beyond the problem of variable selection uncertainty, we should take into consideration that Internet adoption is a dynamic phenomenon. As a consequence, the determinants of Internet adoption may change through time as well as the effects of its determinants.

Table 4: Posterior inclusion probability: Complete correlated set.*

$H = 100$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	44.7	68.2	100	35.4	100	100	97.8	88.9	5.1	42.7	76.9	78.3	20.2	3.7	60.3	6.3	6.9	40.7	36.5	5.6
Naive DVS, $S = 5,000$	27.7	0.0	100	53.7	100	100	100	4.1	9.5	42.3	19.5	19.5	8.5	66.6	29	28.4	28.4	13.6	18.7	11.1
MC^3 DVS, $S = 1,000$	0.9	97	100	0.1	100	100	99.3	0.5	0.2	97	2.1	100	0.1	0.0	0.4	0.9	97	2.4	0.6	0.7
MC^3 DVS, $S = 5,000$	5.9	0.0	100	9.6	100	100	0.3	0.3	14.2	11.8	77.5	76.5	16.6	43.1	40.5	16.9	33.3	54.5	31.5	10.4
SVS	0.0	0.0	100	0.0	100	100	38.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$H = 200$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	46.9	96.2	100	2.9	100	100	47.8	37.8	33.4	1.2	23.2	60.1	23.9	37.6	66.4	13.9	9.4	25.8	63.5	28.7
Naive DVS, $S = 5,000$	0.0	98.1	100	55.0	100	100	22.1	24.0	0.4	1.1	76.0	78.5	0.3	0.3	45.1	0.5	23.4	0.0	0.0	21.0
MC^3 DVS, $S = 1,000$	0.1	98.8	100	50.0	100	96.3	85.9	29.6	0.3	46.2	4.3	49.7	60.1	49.4	59.6	24.5	14.4	46.5	75.5	35.7
MC^3 DVS, $S = 5,000$	0.5	99.5	100	0.0	100	100	67.1	9.5	48.2	0.0	45.0	55.2	7.9	0.2	61.3	55.0	16.3	44.9	0.0	21.3
SVS	5.1	30.4	100	1.9	100	86.7	25.7	5.7	5	2.1	9.5	14.8	2.1	1.2	11.3	7.1	11.3	2.1	1.1	1.2
$H = 500$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	56.4	100	100	51.4	100	96.8	38.2	48.7	17.9	29.6	27.6	46.0	17.9	9.8	51.1	6.1	34.2	38.2	17.5	16.4
Naive DVS, $S = 5,000$	52.9	100	100	47.3	100	100	51.4	25.5	7.5	27.3	16.7	32.9	19.7	19.8	27	12.3	5.7	13.0	19.1	24.0
MC^3 DVS, $S = 1,000$	75.5	100	100	60.4	100	87.7	19.6	43.2	3.9	43.1	52.9	48.2	26.4	20.7	30.8	9.2	29.3	61.3	21.8	2.0
MC^3 DVS, $S = 5,000$	64.8	100	100	25.7	100	92.7	10.9	21.4	0.0	22.9	55	62.7	62.5	8.0	36.4	3.3	17.7	42.9	6.8	5.2
SVS	4.4	100	100	2.6	100	100	33.0	2.6	0.0	2.1	10.9	0.0	2.0	2.4	0.0	0.0	0.0	2.0	5.3	2.3
$H = 2,000$																				
Algorithm	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}	x_{t11}	x_{t12}	x_{t13}	x_{t14}	x_{t15}	x_{t16}	x_{t17}	x_{t18}	x_{t19}	x_{t20}
Naive DVS, $S = 1,000$	100	100	100	100	100	80.9	25.2	28.9	35.5	29.7	28.4	43.2	51.1	49.1	58.7	46.0	47.0	46.6	37.7	59.7
Naive DVS, $S = 5,000$	100	100	100	100	100	67.5	24.8	19.0	27.7	22.8	44.2	37.1	10.0	26.6	38.3	34.8	29.1	37.0	12.1	12.7
MC^3 DVS, $S = 1,000$	100	100	100	100	100	48.4	52.9	61.8	31.5	69.9	100	51.2	37.0	11.3	58.0	15.8	52.5	27.0	53.0	4.9
MC^3 DVS, $S = 5,000$	100	100	100	100	100	81.1	49.5	29.3	38.0	18.2	60.6	37.9	28.9	24.9	28.5	33.1	32.4	35.0	40.3	12.7
SVS	100	100	100	100	100	100	0.0	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.6	0.0	0.0	6.1	0.0

* We use the best 25 models in MC^3. SVS uses BIC approximation.

We analyse 18 potential determinants of Internet subscription in a survey that covers 9 years (2006–2014) using information from 5,000 households in Medellín (Colombia).

We can see in tables 6 and 7 descriptive statistics of the data set. In particular, we have that on average 35.4% of the households have Internet with a standard deviation equal to 0.48. However, we can see in Table 6 that the proportion of Internet subscribers has a positive trend. The proportion of subscribers was 19% in 2006, whereas this figure was 53% in 2014. We can see in Table 7 that households without Internet have lower expenses (proxy of income), less members studying, more children, 45% are in stratum 2,[9] less pay TV, and a higher proportion

[9]Strata are a socio–economic classification designed by the Colombian government to implement social programs. A lower classification implies more household vulnerability.

Table 5: Recent studies on the determinants of Internet subscription.

Paper	Authors	# of predictors	Significant determinants
Internet adoption and usage patterns in Africa: Evidence from Cameroon	(Penard et al., 2015)	16	Gender (+), Age (-), Education (+), Family in Africa (+), Family in World (+), Type of Job, English-Speaking (+), Use Software (+), Install Software (+), PC (+), MP3 (+), Membership (+), City
Determinants of adoption of Internet in Africa: Case of 17 sub-Saharan countries	(Birba and Diagne, 2012)	24	Internet connection at home, Male, Education level, Social network, Rich, Young, Household with a computer x Rate of households with an Internet connection, Household with a computer x Rate of households with a computer, Youth x urban, Density, Residence in urban areas (urban)
Internet Adoption and usage patterns are different: Implications for the digital divide	(Goldfarb and Prince, 2008)	19	Income (+), schooling (+), Married (+), White (+), Age (-), population of city (+), Leisure time (+), Operates a business from home (+), Brings work home (+), Telecommutes (+), Work usage (+)
A segment-based analysis of Internet service adoption among UK households	(Robertson et al., 2007)	9	Dial-up price (-), Broadband price (-), Household income [dial-up] (+), Household income [broadband] (+), General higher schooling (+), General normal schooling (+), Other qualification (+), No qualifications (+), Children (+)
Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts	(Rice and Katz, 2003)	7	Income (+), Age (-), Education (+), Phone calls (-), religious organizations (+), Gender (+), Children (+)
Understanding internet adoption as telecommunications behavior	(Atkin et al., 1998)	21	Education (+), Age (-), Income (+), Communication Activities & Orientations: Localite (-), Pro-technology (+), Computer in home (+), TV Viewing (-), Film Viewing [at Theater] (+)

Table 6: Proportion Internet subscribers: Internet survey, Medellín 2006–2014.

Variable	Year								
	2006	2007	2008	2009	2010	2011	2012	2013	2014
Mean	0.19	0.27	0.28	0.29	0.38	0.49	0.45	0.52	0.53
Std. Dev.	0.39	0.45	0.45	0.45	0.49	0.50	0.50	0.50	0.50
Sample size	625	704	753	755	456	400	419	443	445

have business in their houses. In addition, the head of household have less years of education, less proportion of Caucasian and employees.

Table 7: Descriptive statistics: Internet survey, Medellín 2006–2014.

Variable	**Subscribers**		**No–Subscribers**		**t-statistic***
	Mean	Std. Dev.	Mean	Std. Dev.	
Gender head of household (male)	0.49	0.50	0.49	0.50	0.45
Age	51.23	14.27	51.42	16.43	-0.43
Log expenses household	6.44	0.58	5.74	0.60	39.96*
Years of education	11.51	5.26	7.26	5.07	27.71*
Caucasian	0.20	0.40	0.13	0.33	6.81*
Employed	0.61	0.49	0.55	0.50	4.09*
Pay TV	0.92	0.27	0.55	0.50	33.94*
Business at home	0.06	0.24	0.07	0.26	-1.68
Any household member studying	0.65	0.48	0.54	0.50	7.42*
Number of children	0.79	0.97	1.01	1.19	-6.88*
Number of household members	3.75	1.54	3.69	1.80	1.12
Single	0.37	0.48	0.46	0.50	-5.82*
Stratum 1	0.04	0.19	0.15	0.36	15.39*
Stratum 2	0.22	0.41	0.45	0.50	-18.12*
Stratum 3	0.34	0.48	0.28	0.45	4.47*
Stratum 4	0.18	0.39	0.06	0.25	11.52*
Stratum 5	0.17	0.37	0.04	0.19	13.57*
Stratum 6	0.05	0.23	0.01	0.07	8.99*
Sample size	1,769		3,231		

* Means difference is statistically significant at 5%.

We exclude Stratum 1 to avoid perfect multicollinearity, and include as regressor squared age of head of household.

We can see in Table 8 the posterior inclusion probabilities associated with regressors of Internet subscription using 15,000 iterations[10], and two marginal likelihood bandwidths (445 and

[10]We perform robustness regarding number of iterations. Available upon request.

888). We can see that there is very strong evidence that pay TV and having any household member studying are associated with Internet subscription. In addition, there is positive evidence and strong evidence that years of education of head of household, and number of household members are also associated with Internet subscription, respectively.

Table 8: Posterior inclusion probabilities: Internet subscribers, Medellín.

Variable	Bandwidth=445*	Bandwidth=888*
Gender: Head of household (male)	0.30	0.13
Age	0.41	0.49
Squared age	0.00	0.00
Log expenses household	0.44	0.50
Years of education	**0.89**	**0.76**
White	0.07	0.62
Employed	0.29	0.18
Pay TV	**1.00**	**1.00**
Business at home	0.22	0.14
Any household member studying	**1.00**	**1.00**
Number of children	0.17	0.31
Number of household members	**1.00**	**0.96**
Single: Head of household	0.41	0.42
Stratum 2	0.22	0.04
Stratum 3	0.16	0.16
Stratum 4	0.19	0.16
Stratum 5	0.35	0.36
Stratum 6	0.00	0.00

*Bandwidth equal to 445 and 888 correspond to periods 2014 and 2013–2014, respectively.

Figure 1: Posterior means: Relevant regressors

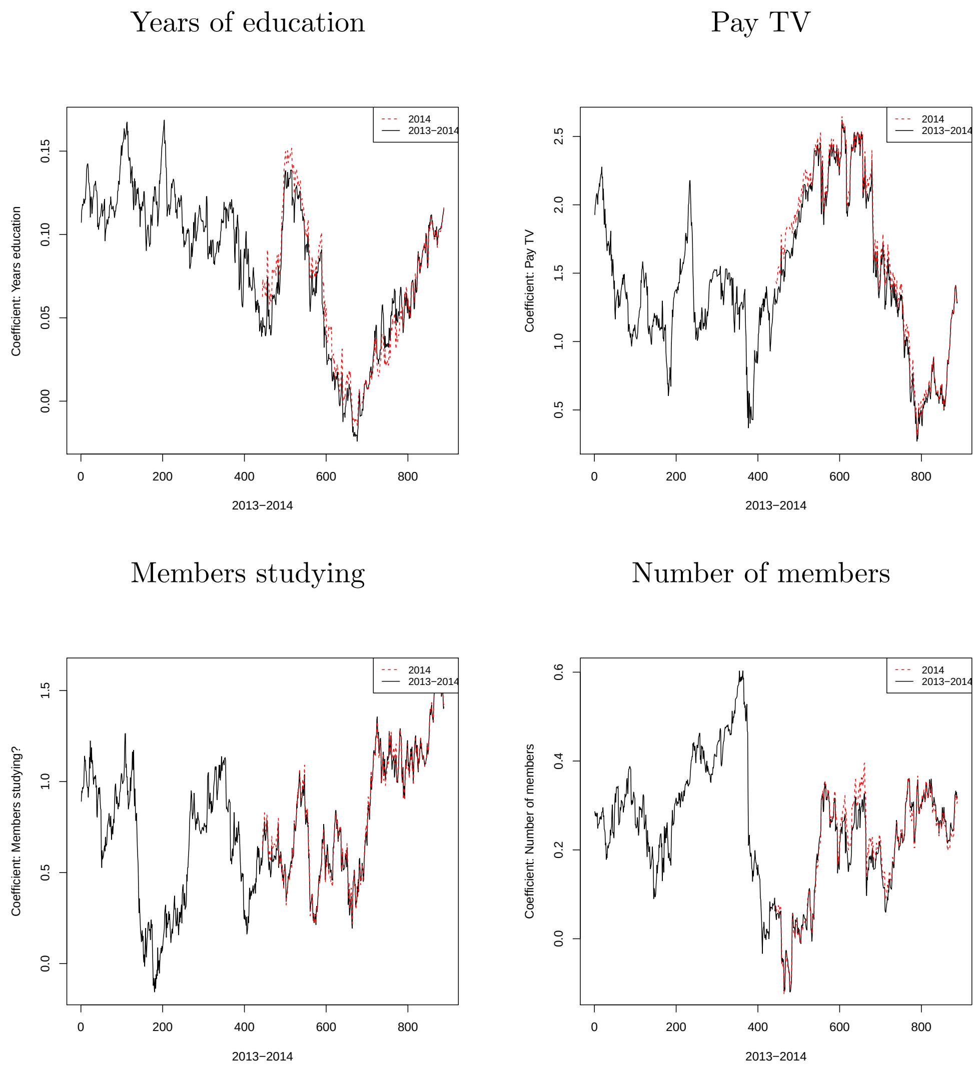


Figure 1 displays mean value estimates of the relevant regressors for Internet subscription using information from the best 25 models (see equation 10). We observe that the outcomes are robust regarding bandwidth choice. We focus now on results using $H = 445$, that is, 2014.

Figure 2: 95% Credible intervals: Relevant regressors, 2014

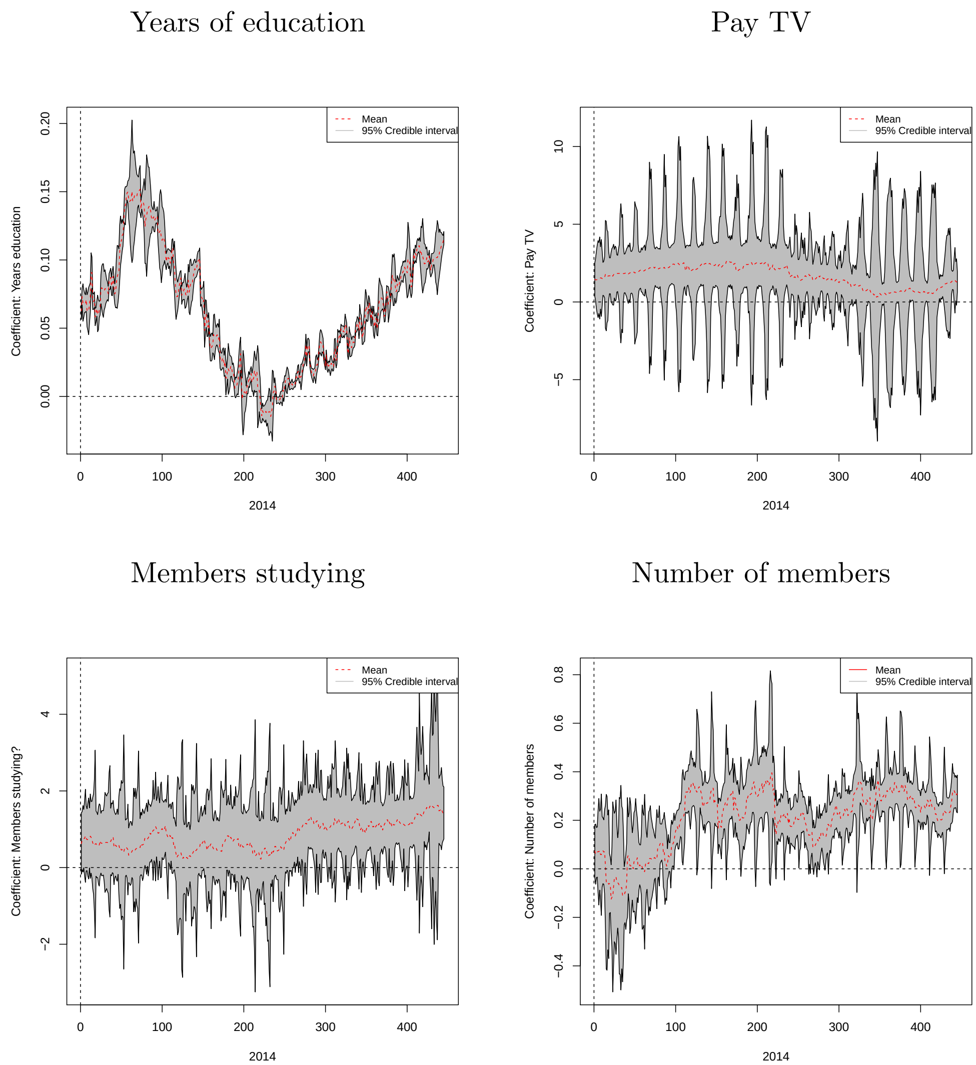
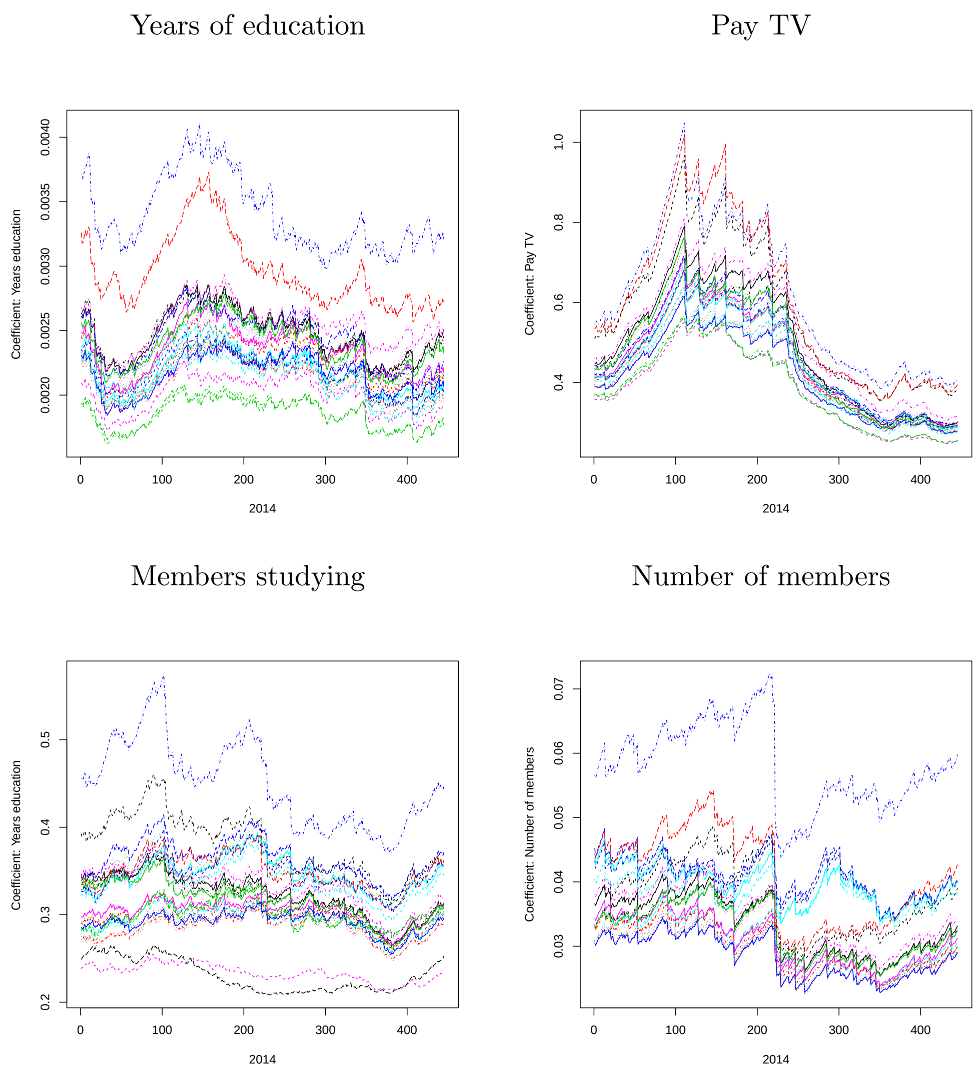


Figure 2 displays 95% credible intervals associated with the relevant regressors. We can see that the most statistically significant posterior estimates between different models are related to years of education followed by number of household members. On the other hand, posterior estimates associated with pay TV and members studying exhibits a high variability level between the best 25 models (see Figure 3 and equation 11). This explains why the width of the credible intervals. However, from economic relevance perspective (magnitude), the former variables are the most relevant.

Figure 3: Coefficient estimates 25 best models: Relevant regressors, 2014



5 Concluding remarks

"Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena."

Box (1976, pp 792)

We agree with Box's advice to employ an Occam's razor driven by marginal likelihoods using

Bayesian model averaging, which is well suited to yield predictive improvements when the entire class of models is well specified, and misspecified (Geweke and Amisano, 2011). In particular, we average over the models with the highest posterior probabilities. We identify them using variable selection in such a way that we mitigate the computational burden by an avoidance of going over the entire model space, but taking into account predictors' uncertainty.

We give an heuristic proof why our approach may work well in complete and incomplete sets of regressors settings, and simulation exercises seem to corroborate this. It seems also from simulation exercises that our algorithm converges faster than MC^3 in dynamic settings. However, future research regarding optimal marginal likelihood bandwidth should be done, as well as extension to other strategies to model variable uncertainty, and dynamic variable selection in other non–linear and non-Gaussian state–space representations.

Regarding Internet subscription in Medellín (Colombia), it seems that the most relevant explanatory variables are years of education of head of households, subscription to pay TV, presence of members studying, and total household members. All these variables have a positive effect on Internet subscription. However, the most stable posterior estimates are associated with years of education.

From a policymaker view, we see that variables associated with education (years of education of head of household and household members studying) have a positive effect on Internet subscription. Therefore, government interventions promoting education would have a positive effect on Internet subscription, as a consequence, the positive externalities that come from it.

Compliance with Ethical Standards

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

Agarwal, R. and Day, A. E. (1998). The impact of the internet on economic education. *The Journal of Economic Education*, 29(2):99–110.

Allen, T. J. (1977). *Managing the flow of technology: technology transfer and the dissemination of technological information within the R and D organization.* Massachusetts Institute of Technology, Cambridge, MA, United States.

Atkin, D. J., Jeffres, L. W., and Neuendorf, K. A. (1998). Understanding Internet adoption as telecommunications behavior. *Journal of Broadcasting & Electronic Media*, 42(4):475–490.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870—-897.

Becker, S., Crandall, M. D., Fisher, K. E., Kinney, B., Landry, C., and Rocha, A. (2010). Opportunity for all: How the american public benefits from internet access at us libraries. *Institute of Museum and Library Services.*

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.-J., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C., Clyde, M., Cook, T., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A., Forster, M., George, E., Gonzalez, R., Goodman, S., Green, E., Green, D., Greenwald, A., Hadfield, J., Hedges, L., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D., Imai, K., Imbens, G., Ioannidis, J., Jeon, M., Jones, J., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S., McCarthy, M., Moore, D., Morgan, S., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D., Winship, C., Wolpert, R., Xie, Y., Young, C., Zinman, J., and Johnson, V. (2017). Redefine statistical significance. *Nature Human Behaviour*, forthcoming.

Berger, J. and Pericchi, L. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* Wiley, Chichester, England.

Birba, O. and Diagne, A. (2012). Determinants of adoption of Internet in Africa: Case of 17 sub-Saharan countries. *Structural Change and Economic Dynamics*, 23(4):463–472.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71:791–799.

Brown, P. J., Vanucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Ser. B*, 64:519–536.

Casella, G., Girón, J., Martinez, L., and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207–1228.

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167.

Chatterjee, D., Maitra, T., and Bhattacharya, S. (2018). A short note on almost sure convergence of bayes factors in the general set–up. Working paper.

Chib, S. and Kuffner, T. (2016). Bayes factor consistency. Working paper.

Clyde, M. and George, E. (2004). Model uncertatinty. *Statistical Science*, 19(1):81–94.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340.

DeGroot, M. H. (1975). *Probability and Statistics*. Addison-Wesley, London.

Dickey, J. M. and Gunel, E. (1978). Bayes factors from mixed probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):43–46.

Eicher, T. S., Henn, C., and Papageorgiou, C. (2012). Trade creation and diversion revisited: Accounting for model uncertainty and natural trading partner effects. *Journal of Applied Econometrics*, 27(2):296–321.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

George, E. and McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164:130—-141.

Goldfarb, A. and Prince, J. (2008). Internet adoption and usage patterns are different: Implications for the digital divide. *Information Economics and Policy*, 20(1):2–15.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Hansen, B. (2005). Challenges for econometric model selection. *Econometric Theory*, 21:60–68.

Harris, R. G. (1998). The internet as a gpt: factor market implications. *General purpose technologies and economic growth*, pages 145–166.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, London.

Jerrum, M. and Sinclair, A. (1996). The markov chain monte carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-Hard Problems*, pages 482—-487. Boston: PWS Publishing.

Jetter, M. and Ramírez Hassan, A. (2015). Want export diversification? educate the kids first. *Economic Inquiry*. Forthcoming.

Jhonson, V. E. and Rossell, D. (2012). Bayesian model selection in high–dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Leamer, E. (1978). *Specification Searches*. New York: Weley, New York.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49(3):293–337.

Madigan, D., Raftery, A., York, J., and Bradshaw, J. (1994). Strategies for graphical model selection. Selecting Models from Data: AI and Statistics, pages 91–100. New York: Springer–Verlag.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232.

McCormick, T. H., Raftery, A. E., Madigan, D., and Burd, R. S. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, 68(1):23–30.

Moral-Benito, E. (2013a). Determinants of economic growth: A Bayesian panel data approach. *The Review of Economics and Statistics*, 94(2):566–579.

Moral-Benito, E. (2013b). Model averaging in economics: an overview. *Journal of Economic Surveys*.

Ormerod, J. T., You, C., and Muller, S. (2017). A variational bayes approach to variable selection. *Electronic Journal of Statistics*, 11:3549—-3594.

Penard, T., Poussing, N., Mukoko, B., and Piaptie, G. B. (2015). Internet adoption and usage patterns in Africa: Evidence from Cameroon. *Technology in Society*, 42:71–80.

Pericchi, L. and Pereira, C. (2016). Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics*, 30(1):70–90.

Petris, G., Petrone, S., and Campagnoli, P. (2007). *Dynamic Linear Models with R*. Springer.

Raftery, A. and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, 98(464):931–938.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52—-66.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Rice, R. E. and Katz, J. E. (2003). Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts. *Telecommunications Policy*, 27(8):597–623.

Robertson, A., Soopramanien, D., and Fildes, R. (2007). A segment-based analysis of Internet service adoption among UK households. *Technology in society*, 29(3):339–350.

Sala-i Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, pages 813–835.

Schreyer, P. (2000). The contribution of information and communication technology to output growth.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074.

Timmermann, A. (2006). Chapter 4 forecast combinations. volume 1 of *Handbook of Economic Forecasting*, pages 135–196. Elsevier.

Walker, S. G. (2004a). Modern Bayesian asymptotics. *Statistical Science*, 19(1):111–117.

Walker, S. G. (2004b). New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5):2028–2043.

Yeung, K., Bumgarner, R., and Raftery, A. (2005). Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *BIOINFORMATICS*, 21(10):2394–2402.