Unveiling Covariate Inclusion Structures In Economic Growth Regressions Using Latent Class Analysis*

Jesus Crespo Cuaresma[a,b,c,d], Bettina Grün[e], Paul Hofmarcher[a], Stefan Humer[a], and Mathias Moser[a]

[a]Department of Economics, Vienna University of Economics and Business (WU)

[b]Austrian Institute of Economic Research (WIFO)

[c]International Institute of Applied System Analysis (IIASA)

[d]Wittgenstein Centre for Demography and Global Human Capital (WIC)

[e]Department of Applied Statistics, Johannes Kepler University Linz (JKU)

Abstract

We propose the use of Latent Class Analysis methods to analyze the covariate inclusion patterns across specifications resulting from Bayesian Model Averaging exercises. Using Dirichlet Process clustering, we are able to identify and describe dependency structures among variables in terms of inclusion in the specifications that compose the model space. We apply the method to two datasets of potential determinants of economic growth. Clustering the posterior covariate inclusion structure of the model space formed by linear regression models reveals interesting patterns of complementarity and substitutability across economic growth determinants.

JEL Classification: C11, C21, O47.

Keywords: Economic Growth Determinants, Bayesian Model Averaging, Latent Class Analysis, Dirichlet Processes.

*The authors would like to thank Sylvia Frühwirth-Schnatter, two anonymous referees and the participants at the *Second Bayesian Young Statisticians Meeting* for helpful comments. Corresponding Author: Jesus Crespo Cuaresma. Address: Welthandelsplatz 1, 1020 Vienna, Austria, Tel: +43(0)131336-4530, Fax: +43(0)131336-728, Email: jcrespo@wu.ac.at. Paul Hofmarcher's research is supported by funds of the Oesterreichische Nationalbank (Oesterreichische Nationalbank, Anniversary Fund, project number: 14663).

1. Introduction

Bayesian Model Averaging (BMA) has become a popular tool for economic growth applications in economics (for a comprehensive introduction to BMA, see Hoeting et al., 1999). The so-called *open-endedness* of economic theory concerning the factors driving income per capita differences across countries (Brock and Durlauf, 2001) allows the empirical researcher to specify a large number of models to quantify the effect of potential drivers on economic growth. The use of techniques that explicitly assess model uncertainty (mostly within the class of linear regression models) has thus become widespread in econometric research dealing with the empirical determinants of income growth differences across countries (for some seminal contributions to this literature, see e.g. Fernandez et al., 2001; Sala-i Martin et al., 2004; Masanjala and Papageorgiou, 2008; Durlauf et al., 2008; Ley and Steel, 2009b).

Economic growth applications of BMA tend to quantify the relative importance of a given covariate by calculating its so-called posterior inclusion probability (PIP), which is defined as the sum of posterior probabilities of specifications which contain that particular variable. Such a statistic has become a standard tool in econometric applications of BMA and is routinely used to measure the relative importance of different potential drivers of income growth differences across economies. While standard PIPs are intuitive measures that provide valuable insights into the overall importance of individual covariates as economic growth determinants, they face a number of shortcomings. The PIP neglects the heterogeneity across typical model specifications and accordingly does not inform about whether the degree of *importance* of the variable is evenly spread across potential specifications (that is, it is relatively independent of whether other covariates are part of the model) or, on the contrary, it is particular to specific combinations of explanatory variables.

Previous work assessing joint covariate inclusion in BMA applications has focused on capturing relevant dependency structures using bivariate measures, that is, concentrating on the analysis of the joint posterior distribution of the inclusion of pairs of variables over the model space. Such a concept has been quantified in the form of bivariate *jointness* measures in the context of BMA, put forward first by G. Doppelhofer and M. Weeks in a working paper of 2005 (Doppelhofer et al., 2005), which got published in a slightly different version as Doppelhofer and Weeks (2009a). Ley and Steel (2007), Strachan (2009) and Ley and Steel (2009a) offer alternative measures of jointness. In particular, Ley and Steel (2007) formulate a set of properties for jointness measures and show that Doppelhofer and Weeks's statistics do not fulfill them. Additionally, they propose two other indices which satisfy all of their suggested properties. Strachan (2009) shows that the interpretability of the jointness measure of Doppelhofer and Weeks (2009a) may be limited in contexts where one or both of the analyzed variables have a negligible PIP and offers yet another measure in order to tackle this shortcoming. Doppelhofer and Weeks (2009b), on the other hand, argue that another desirable property of jointness measures happens to be fulfilled by their indicator but not accounted for in the indices of Ley and Steel (2007) or Strachan (2009).[1]

[1]Interestingly, the measures proposed by Doppelhofer and Weeks (2009a), Ley and Steel (2007) and Strachan

In this paper we propose an alternative approach aimed at succinctly and comprehensibly describing the dependency structure across variables in the model space using *latent class analysis* (LCA, see, e.g., Vermunt and Magidson, 2002) and apply it to economic growth regressions. This method was first introduced by Lazarsfeld (1950) to describe dependency structures between observed discrete variables based on latent traits and has gained widespread popularity in such research fields as psychometrics or political science (see, e.g., Breen, 2000; Blaydes and Linzer, 2008). The main idea behind LCA is to relate the realizations of observed variables to an unobserved, categorical latent variable which captures the dependency structure between the observed variables. This latent variable groups observations in such a way that the dependency between variables is reduced to a minimum within groups. By applying LCA methods to the covariate inclusion structure of best models identified by BMA, we are able to capture the dependency patterns across included covariates through a (unobserved) latent variable which induces classes with independent covariates conditional on class membership. Such a setting implies that PIPs within clusters constitute sufficient information to describe the importance of the variables and the differences of PIPs between clusters are representative of the dependencies in the inclusion of a covariate with respect to (all) other variables.

The method proposed in this paper provides a tool for applied econometricians that goes beyond the identification of individual robust determinants of socioeconomic variables by distilling the joint covariate structures that underlie the distribution of the posterior model probability across specifications. Suitable theoretical frameworks based on the results of the clustering can then be inferred based on the identity of the corresponding groups of variables. In the spirit of Durlauf et al. (2008), the applied researcher may be interested in incorporating prior beliefs about the relative importance of some theoretical frameworks (defined over the joint prior inclusion probability of certain covariate groups) in order to assess the evidence for or against them. The modeling tool provided by our method is able to incorporate this information in a straightforward manner.

We apply this approach to the two datasets that have been most widely used for assessing the robustness of economic growth determinants (those in Fernandez et al., 2001, and Sala-i Martin et al., 2004, henceforth FLS and SDM, respectively). Our results for the FLS dataset reveal patterns of complementarity and substitutability across geographical, institutional and religious variables. For the SDM dataset, we find that the importance of the variable related to malaria prevalence is highly dependent on the inclusion of other covariates in the specification. The insights gained from the clustering exercise for the SDM dataset partly reconcile some of the contradictory results found in the literature concerning the importance of malaria prevalence as a determinant of income growth differences across countries (see for example Sala-i Martin et al., 2004; Schneider and Wagner, 2012; Hofmarcher et al., 2014).

The remainder of this paper is structured as follows. In Section 2, we present the econometric setting used to analyze the anatomy of covariate inclusion over the model space within BMA

(2009) were independently developed earlier in the context of data mining. The statistic of Doppelhofer et al. (2005) is known as *log-ratio*, the measures of Ley and Steel (2007) are related to the *Jaccard index*. The index of Doppelhofer and Weeks (2009a) is known as *odds-ratio* and Strachan (2009)'s measure is closely related to the so-called *two-way support* (see Tan et al., 2004; Glass, 2013).

applications and outline the LCA approach. Section 3 presents the results of the LCA analysis applied to the set of best models identified for the FLS and SDM datasets. Section 4 concludes and proposes further paths of research.

2. Evaluating Covariate Inclusion Dependency Using Latent Class Analysis

2.1. Model Uncertainty and Economic Growth Determinants: The Econometric Framework

The standard setting for BMA analysis in the framework of cross-country growth regressions assumes that the growth rate of income per capita (y) can be linearly related to a group of covariates (X_j) chosen from a set of potential growth determinants (X). Assuming that n observations are available, a typical linear regression model (M_j) is given by

$$y|\alpha, \beta_j, \sigma \sim N(\alpha\iota + X_j\beta_j, \sigma^2 I), \tag{1}$$

where ι is a column vector of ones of dimension n. Assuming that a total of K variables are available, inference on a quantity of interest (Δ) is given by

$$p(\Delta|y) = \sum_{j=1}^{2^K} p(\Delta|y, M_j)p(M_j|y), \tag{2}$$

where $p(M_j|y)$ is the posterior model probability, which in turn is proportional to the product of the prior model probability $p(M_j)$ and its marginal likelihood $p(y|M_j)$. After eliciting priors over model-specific parameters ($p(\beta_j|M_j)$ and $p(\sigma|M_j)$), as well as over models ($p(M_j)$), posterior model probabilities and thus the posterior distributions given by equation (2) can be computed. The problems caused by the exorbitantly large number of summands in equation (2) when K is not small can be overcome in a straightforward manner by sampling from the model space using Markov Chain Monte Carlo (MCMC) methods (Madigan and York, 1995).

In the spirit of the literature on jointness in BMA applications, we propose to analyze the anatomy of the set of models sampled by the Markov chain in order to carry out inference about the covariate inclusion structures existing in the model space. While existing jointness measures tend to concentrate on the analysis of the $K \times K$ matrix of bivariate inclusion frequencies in the Markov chain, we aim at gaining knowledge about the overall structure of covariate inclusion by analyzing the full $M \times K$ matrix of inclusion profiles of the specifications sampled by the Markov chain, where M is the number of sampled models. A model profile γ_i, for $i = 1, \ldots, M$ (that is, one of the rows of the matrix), is a K-dimensional vector of ones and zeros indicating the variables which are included in model i, with typical element $\gamma_{ik} = 1$ if variable k is part of model i and $\gamma_{ik} = 0$ otherwise. We propose to perform the analysis of the inclusion patterns over the model space assuming the existence of implicit latent groups to which model specifications are assigned

depending on their covariate inclusion pattern.

2.2. Latent Classes and Covariate Inclusion: A Bayesian Approach Using Dirichlet Processes

We propose to use a method that resembles existing BMA applications dealing with the computation of jointness measures among covariates. It takes a two-step approach in terms of analyzing the posterior probability distribution over model specifications obtained using standard BMA methods. Using clustering methods based on LCA, it aims at unveiling clusters of model profiles among the specifications sampled in the Markov chain Monte Carlo model composite procedure.

Following the methods put forward by Molitor et al. (2010), we apply Dirichlet Process Clustering (DPC) in order to carry out inference about the latent classes governing covariance inclusion structures in economic growth regressions. Compared to other methods in the literature (Forgy, 1965; Hartigan and Wong, 1979; Patterson et al., 2002), DPC eliminates the need to set the number of latent classes a priori. While selecting a suitable number of clusters has been a widely discussed problem in the LCA and finite mixture literatures (McLachlan and Peel, 2000, Chap. 6), the nature of Bayesian inference using DPC allows for the automatic selection of an optimal number of clusters for given prior settings.

We assume that γ_i, the K-dimensional vector summarizing the variable inclusion profile for model i, has elements that arise from a mixture of infinitely, but countably many distributions,

$$p(\gamma_i) = \sum_{c=1}^{\infty} p(g_i = c) \prod_{k=1}^{K} p(\gamma_{ik} | g_i = c), \tag{3}$$

where $p(g_i = c)$ denotes the probability that model i is assigned to cluster c and $p(\gamma_{ik} \mid g_i = c)$ governs the inclusion probability of the k-th covariate in cluster c. In turn, for our application we use

$$\begin{aligned} p(\gamma_{ik} \mid g_i = c) &\sim \text{Bern}(\pi_{ck}), \\ \pi_{ck} &\sim \text{Beta}(\delta, \delta), \\ p(g_i = c) = V_c \prod_{j=1}^{c-1} (1 - V_j), \\ V_c &\sim \text{Beta}(1, \alpha). \end{aligned}$$

Such a mixture model implies, that given assignment to a cluster, the inclusion of covariate k resembles the probabilistic process proposed, for example, in Ley and Steel (2009b). The inclusion probability of covariate k in a given cluster c is thus governed by a Bernoulli distribution whose parameter follows a Beta distribution. The probabilistic structure that governs assignment to the different clusters, $p(g_i = c)$, on the other hand, corresponds to the so-called stick-breaking process formulation of the Dirichlet process (see Sethuraman, 1994; Papaspiliopoulos, 2008; Liverani et al.,

2013). This representation can be interpreted as determining the mixing proportions $p(g_i = c)$ by successive divisions of the unit interval whose relative sizes are determined by independent draws from the Beta$(1, \alpha)$ distribution.

Posterior inference for DPC can be carried out using MCMC methods. Papaspiliopoulos and Roberts (2008), for instance, present an approach using retrospective sampling. However, identifying a DPC model is difficult due to label switching (Redner and Walker, 1984). We follow Molitor et al. (2010) and derive a suitable partitioning of the set of sampled model profiles using the information on co-assignment to the same clusters during sampling. This information is collapsed into an association matrix that can be interpreted as a similarity matrix between model profiles when assuming that model specifications often assigned to the same cluster are *similar*. A clustering technique relying only on similarity measures between specification profiles can then be used to find the final clustering, for instance Partitioning Around Medoids (PAM, Kaufman and Rousseeuw, 1990), which is the approach used in our empirical application.

Once a partition has been chosen, several statistics can be used to assess the goodness of fit of the clustering. In our application we rely on measures based on the likelihood ratio chi-squared test statistic (G^2), which measures goodness-of-fit by relating the observed counts of specification profiles in each cluster to the counts predicted by the estimated model. The test statistic is given by $G^2 = 2\sum_j^{2^K} q_j \ln \frac{q_j}{Q_j}$, where q_j refers to the observed number of counts of specification profile γ_j and Q_j is the expected number of counts assuming independency of the explanatory variables (see for example Brier, 1980). We calculate this G^2 statistic separately for each cluster and the aggregated BMA results.

In addition, in order to identify substitutability/complementarity of variables based on the cluster solution, we compute a simple measure of *interestingness* of a variable (*IM*) in the spirit of the literature on association rules. The interestingness measure *IM* is determined as the square root of the mean squared deviation of PIPs with respect to the unclustered case across clusters, weighted by the cluster-specific mass of posterior model probabilities. Thus, this measure reflects the stability of the relative importance of the variable across model structures and is able to give an indication of the existence of substitutability/complementarity inclusion patterns across covariates in the model space.

2.3. A Simulation Exercise

We assess the performance of the method by making use of a small-scale simulation exercise. We consider a set of ten potential covariates, x_k, $k = 1, \ldots, 10$ and two settings based on different data generating processes. In the first setting, the dependent variable is a linear combination of the first five covariates and a random error term, $y_i = \sum_{k=1}^{5} x_{ik} + \varepsilon_i$, where $\varepsilon_i \sim N(0, 0.01)$ and all covariates are drawn from standard normal distributions. In the second setting, the dependent variable can be represented by two different linear combinations of predictors, so that $y_i = \sum_{k=1}^{5} x_{ik} + \varepsilon_i = -\sum_{j=6}^{10} x_{ik} + \varepsilon_i$.[2]

[2]Technically, we implement this setting by defining $x_{i,10} \approx -\sum_{k=1}^{9} x_{ik}$.

Using simulated datasets with 50 observations for each one of the settings, we perform standard BMA (assuming a single cluster of model specifications) as well as the clustering procedure proposed over the sampled model profiles. We use a Beta-Binomial prior for covariate inclusion (Ley and Steel, 2009b) and a unit information prior for the parameters in the BMA application. For this small example with $K = 10$ a complete enumeration of all models is performed. For the clustering procedure, we use a Gamma(2, 1) prior over α, elicit $\delta = 90$ and retain the top 500 models. The posterior inference is based on 1,500 MCMC iterations, after 1,000 burn-in runs. The results for the first (single cluster) setting are presented in Table 1, where we report the posterior inclusion probabilities and the mean of the posterior distribution of the parameters associated to each one of the covariates, averaged over 100 simulated datasets.

The standard BMA method (see results in panel (a) of Table 1) correctly identifies the covariates included in the true model and the mean of the posterior distribution of the relevant parameters appear very close to the true value of unity. The clustering approach identifies two clusters, with the first one covering over 99% of the models in the BMA procedure and reproducing the same results as those in the non-clustered case in terms of PIP and means of the posterior distribution of the associated parameters (see panels (b) and (c) in Table 1). In the second setting, whose results are presented in Table 2, the standard BMA procedure *averages out* the effects of the two alternative data generating processes. The PIP values obtained using BMA are around 0.6 for all variables and the mean of the posterior distribution over the parameters is approximately 0.5 for the first five covariates and -0.5 for the rest of the variables. DPC is able to disentangle the two competing data generating processes, assigning roughly the same posterior mass to each one of the two clusters found. The mean of the posterior distribution of the parameters are in line with the actual values in the true model(s) and the covariates which are not included in the alternative specification have a relatively low PIP and an expected effect which is very close to zero.

3. Covariate Inclusion Clustering in Economic Growth Regressions

The clustering method presented in Section 2 is applied to the datasets compiled by Fernandez et al. (2001) and Sala-i Martin et al. (2004) (henceforth, FLS and SDM datasets). These two datasets comprise cross-country information on a large number of potential determinants of income growth and have been extensively used to assess empirically the role played by model uncertainty in economic growth regressions. In addition to GDP per capita growth figures, the FLS dataset is composed by 41 covariates and spans information for 72 countries, while the SDM dataset includes information on 67 different determinants for 88 economies. The variables in both datasets are presented in the Appendix A.

The BMA analysis of both datasets is carried out using a Beta-Binomial prior on covariate inclusion probabilities with a prior average model size of $K/2$ (20.5 for the FLS dataset and 33.5 for the SDM dataset) and the hyper g-prior proposed in Liang et al. (2008) for the regression coefficients. We base our inference concerning the inclusion probability of covariates on five million MCMC model draws, whereby the first two million draws were discarded. Alternatively,

Table 1: Simulation Results: Single cluster

(a) Standard BMA

	PIP	Post. Mean
β_1	1.0000	0.9799
β_2	1.0000	0.9822
β_3	1.0000	0.9810
β_4	1.0000	0.9828
β_5	1.0000	0.9796
β_6	0.2030	−0.0001
β_7	0.2053	−0.0007
β_8	0.2032	0.0003
β_9	0.2063	−0.0002
β_{10}	0.2053	−0.0001

(b) DPC: Cluster 1 (>99%)

	PIP	Post. Mean
β_1	1.0000	0.9799
β_2	1.0000	0.9822
β_3	1.0000	0.9810
β_4	1.0000	0.9828
β_5	1.0000	0.9796
β_6	0.2009	−0.0001
β_7	0.2033	−0.0007
β_8	0.2011	0.0003
β_9	0.2042	−0.0002
β_{10}	0.2034	−0.0001

(c) DPC: Cluster 2 (<1%)

	PIP	Post. Mean
β_1	1.0000	0.9799
β_2	1.0000	0.9824
β_3	1.0000	0.9811
β_4	1.0000	0.9825
β_5	1.0000	0.9804
β_6	0.9975	−0.0004
β_7	1.0000	−0.0025
β_8	1.0000	0.0010
β_9	1.0000	−0.0013
β_{10}	1.0000	−0.0010

Simulation results averaged over 100 simulated datasets. Data generating process: $y_i = \sum_{k=1}^{5} x_{ik} + \varepsilon_i$. Column labelled *PIP* reports the posterior inclusion probability, column labelled *Post. Mean* reports the mean of the posterior distribution of the corresponding parameter. See text for details on the setting of the simulation.

we also implemented dilution priors over the model space following George (1999) (see also Durlauf et al., 2008). Such a model prior assigns relatively lower prior probability to specifications with highly correlated covariates by weighting the prior model probability using the determinant of the correlation matrix of the explanatory variables. The results obtained using such a dilution prior are not qualitatively different from those with the standard Beta-Binomial prior which are presented below.[3]

Using the top 500 unique models visited by the Markov chain (weighted by their posterior model probability), we apply the clustering procedure described in Section 2 in order to unveil clusters of inclusion patterns across specifications. Technically, we create an auxiliary dataset composed by the 500 top model profiles drawn where the number of observations of each model profile is proportional to its posterior probability. We normalize this auxiliary dataset so that the profile corresponding to the 500^{th} top model is included exactly once and the relative importance of the rest of the models is preserved. For the FLS and SDM dataset the weighted top 500 model profiles in the auxiliary datasets span 33,480 and 28,800 model profile observations, respectively.[4]

Concerning prior elicitation for DPC, we use a setting that implies a relative preference for a smaller number of broad clusters over a multitude of clusters populated by few model structures,

[3]For the FLS dataset, for instance, the correlation between the posterior inclusion probabilities obtained with the dilution prior and the standard Beta-Binomial prior, as well as between the means and standard deviations of the posterior parameter distributions, tend to be above 0.8. Detailed results of the BMA exercise using George (1999)'s dilution prior are available from the authors upon request.

[4]Expanding the set of top models to cover a larger part of the posterior model probability leads to significant computational complications. For the case of the FLS dataset, which contains less covariates, we also implemented the method for the top 1,000 models, leading to similar results as those presented for the top 500 specifications. Such a result is not very surprising given the fact that the increase in the posterior model probability covered by the top models is very modest when moving from the top 500 to the top 1,000 specifications.

Table 2: Simulation Results: Two clusters

(a) Standard BMA

	PIP	Post. Mean
β_1	0.6030	0.4951
β_2	0.6038	0.4949
β_3	0.6032	0.4931
β_4	0.6039	0.4945
β_5	0.6028	0.4941
β_6	0.5983	−0.4854
β_7	0.5989	−0.4874
β_8	0.5985	−0.4853
β_9	0.5983	−0.4851
β_{10}	0.5983	−0.4857

(b) DPC: Cluster 1 (49%)

	PIP	Post. Mean
β_1	0.1995	0.0024
β_2	0.2009	0.0023
β_3	0.2000	0.0012
β_4	0.2014	0.0018
β_5	0.1991	0.0020
β_6	1.0000	−0.9774
β_7	1.0000	−0.9808
β_8	1.0000	−0.9772
β_9	1.0000	−0.9771
β_{10}	1.0000	−0.9782

(c) DPC: Cluster 2 (51%)

	PIP	Post. Mean
β_1	1.0000	0.9817
β_2	1.0000	0.9814
β_3	1.0000	0.9789
β_4	1.0000	0.9809
β_5	1.0000	0.9801
β_6	0.2009	0.0006
β_7	0.2025	0.0000
β_8	0.2018	0.0006
β_9	0.2011	0.0008
β_{10}	0.2009	0.0006

Simulation results averaged over 100 simulated datasets. Data generating process: $y_i = \sum_{k=1}^{5} x_{ik} + \varepsilon_i = -\sum_{k=6}^{10} x_{ik} + \varepsilon_i$. Column labelled *PIP* reports the posterior inclusion probability, column labelled *Post. Mean* reports the mean of the posterior distribution of the corresponding parameter. See text for details on the setting of the simulation.

which may eventually lack interpretability. We use a Gamma(2, 1) prior over α and $\delta = 90$. Posterior inference is based on 1,500 MCMC iterations, after 1,000 burn-in runs. This choice of priors is relatively standard in LCA applications (see e.g. Liverani et al., 2013).[5]

3.1. Results for the FLS Dataset

DPC identifies an optimal partition of seven clusters of models by inclusion structure in the FLS dataset. Table 3 provides an overview of the main characteristics of these different model clusters regarding the number of model specifications in the cluster, as well as the mean model size and the average adjusted R^2 for specifications within the cluster. These statistics are also presented for the unclustered model space considered. Although the top 500 models used for the analysis only cover approximately 8% of the posterior model probability in the space of potential specifications, the overall unclustered results are very similar to those in Fernandez et al. (2001) concerning the PIP of individual variables.[6]

The first two clusters capture more than half of the posterior mass covered by the set of specifications considered, while clusters 6 and 7 cover a negligible part of the model space in terms of posterior model probability. Cluster 7 is composed by very large models and due to its minimal importance in terms of posterior probability does not appear particularly relevant in terms of interpretation. The cluster-specific G^2 statistics imply an improvement in fit as compared to the

[5] We have carried out several robustness checks changing the elicitation of the priors which did not lead to any significant differences in the inference results as long as the prior setting implies a preference for a relatively small number of clusters.

[6] It should be noted that, in contrast to Fernandez et al. (2001) and Ley and Steel (2007), we employ a hyperprior for prior inclusion probabilities and model-specific parameters, following Ley and Steel (2009b) and Liang et al. (2008), respectively.

unclustered results once the covariate inclusion structures are assigned to the classes identified. The reduction in the G^2 statistic is very sizable and widespread across the clusters.

Table 3: Summary of FLS clusters

	Overall	1	2	3	4	5	6	7
$\sum$ Posterior model prob.	0.08	0.03	0.02	0.01	0.01	0.01	0.00	0.00
Average model size	10.46	10.46	8.68	8.44	11.59	10.95	18.15	41.00
Average adjusted R^2	0.83	0.84	0.81	0.80	0.85	0.84	0.90	0.91
G^2 statistic ($\times 10^5$)	3.52	0.24	0.24	0.13	0.15	0.09	0.19	0.00

Figure 1 offers a graphical representation of the differences in PIPs for individual covariates across the identified clusters. The covariates are sorted by their PIPs in the standard (unclustered) BMA exercise, which are plotted as a solid line together with their corresponding within-cluster PIPs, depicted as bars. It should be noted that the patterns of PIP across variables in all clusters differ structurally from the unclustered BMA results, so that no individual cluster mimics the PIPs obtained by the standard BMA exercise closely. The color of the bars in Figure 1 corresponds to the value of the *IM* statistic.

Table 4: FLS dataset: Weighted correlation of cluster-specific PIPs for variables with $IM > 0.5 \max(IM)$

	SubSahara	EcoOrg	YrsOpen	Muslim	RuleofLaw
SubSahara	1.00	0.50	−0.65	−0.34	0.73
EcoOrg		1.00	−0.87	−0.33	0.87
YrsOpen			1.00	0.45	−0.96
Muslim				1.00	−0.31
RuleofLaw					1.00

The PIPs of the four most robust variables of the FLS dataset (`Confucian`, `GDP60`, `LifeExp` and `Equipinv`) appear stable across clusters. The variables with a higher degree of variability in PIPs across clusters tend to be related to geography (`SubSahara`), institutions (`EcoOrg`, `RuleofLaw` and `YrsOpen`[7]) and religion (`Muslim`). The characteristics of the inclusion structure of these variables across clusters can be best understood by computing the weighted correlation matrix of cluster-specific PIPs, which is presented in Table 4. The correlation among covariate inclusion variables reveals that `SubSahara`, `EcoOrg` and `RuleofLaw` tend to contain complementary information in the sense of appearing together in specifications. The same is true for the group of variables formed by `YrsOpen` and `Muslim`, while the inclusion of these two sets of variables presents sizable

[7]The variable `YrsOpen` is based on the Sachs-Warner index of openness, which has a strong institutional component. For example, socialist economies are automatically considered closed to trade by this indicator.

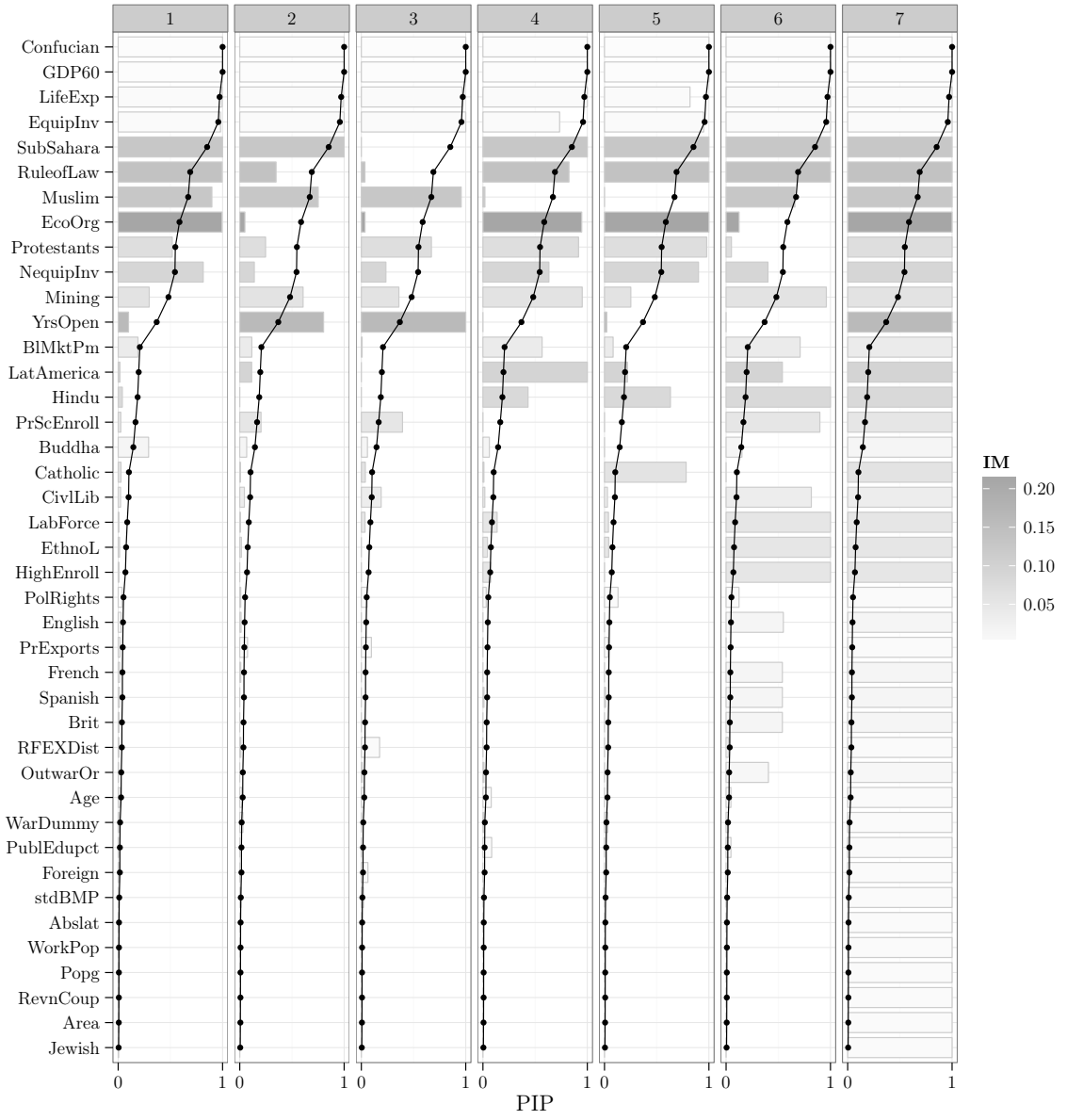

Figure 1: FLS dataset: PIPs in unclustered BMA (solid line) and by identified cluster (bars)

negative correlation. This result indicates that some of the effects of institutions and geographical variables on economic growth can be alternatively modeled using these two groups of covariates in a robust manner, but that once that they are controlled for, the inclusion of variables of the other group appears redundant.

The interplay of changes in PIPs across clusters presented in Figure 1 indicates that the set of religious, institutional and geographical variables used in cross-country growth regressions often contain redundant information which can be replicated using different subgroups of them. An example of such a phenomenon is observed when comparing clusters 1 and 3. The importance of `SubSahara` and `RuleofLaw` as growth determinants which can be inferred from the results in cluster 1 disappears in cluster 3 and their fall in PIPs occurs in parallel to a strong increase in PIP for `YrsOpen`. The set of religious variables (`Muslim`, `Catholic`, `Protestants`, `Hindu` and, to a lesser extent, `Buddha`) also presents large variation in PIPs across clusters.

3.2. Results for the SDM Dataset

Ley and Steel (2009a) found very weak (bivariate and/or trivariate) jointness in the group of covariates included in the SDM dataset. Our procedure splits the model space into three different model clusters by covariate inclusion patterns. Table 5 presents the summary statistics for the identified clusters. The top 500 unique specifications cover 40% of the posterior model probability, a much larger proportion than in the case of the FLS dataset. The structure of variable inclusion for the SDM dataset appears to have a different nature as compared to the results for the FLS dataset. In addition to the lower number of identified clusters, the first two classes of inclusion structures identified exhibit relatively similar characteristics in terms of the posterior model probability covered. As in the case of the FLS dataset, the cluster specific G^2 statistics are lower than the corresponding value for the model without clustering, thus supporting the method employed.

Figure 2 depicts the PIPs of the variables in the SDM dataset computed using the top 500 models, as well as those derived from the models in the single clusters.[8] The results show a large degree of variability in PIPs across clusters for many of the covariates, including those presenting the highest PIPs in the unclustered case.

Given the large posterior probability mass over models covered by the first two clusters, we concentrate on the differences in PIPs observed between these two. Remarkable differences in PIPs across these two clusters can be observed for the `MALFAL66` variable, which presents a much higher PIP in the second cluster, making it the second most important variable for models within that cluster. Such a phenomenon is accompanied by a sizable decrease in PIP for `P60, IPRICE1, TROPICAR, GDPCH60L` and `DENS65C`. The empirical literature on model uncertainty in cross-country growth regressions which analyzes the SDM dataset often reports on the effect that the use of different approaches to parameter shrinkage has on the importance of `MALFAL66`. Schneider and

[8] Variables with PIP lower than 5% have been excluded in order to improve the readability of the graph. For these variables no remarkable changes could be detected when comparing the BMA results with the cluster-specific PIPs.

Table 5: Summary of SDM clusters

	Overall	1	2	3
$\sum$ Posterior model prob.	0.40	0.21	0.17	0.03
Average model size	5.47	6.54	3.95	6.77
Average adjusted R^2	0.67	0.71	0.62	0.70
G^2 statistic ($\times 10^5$)	10.25	1.33	2.36	0.58

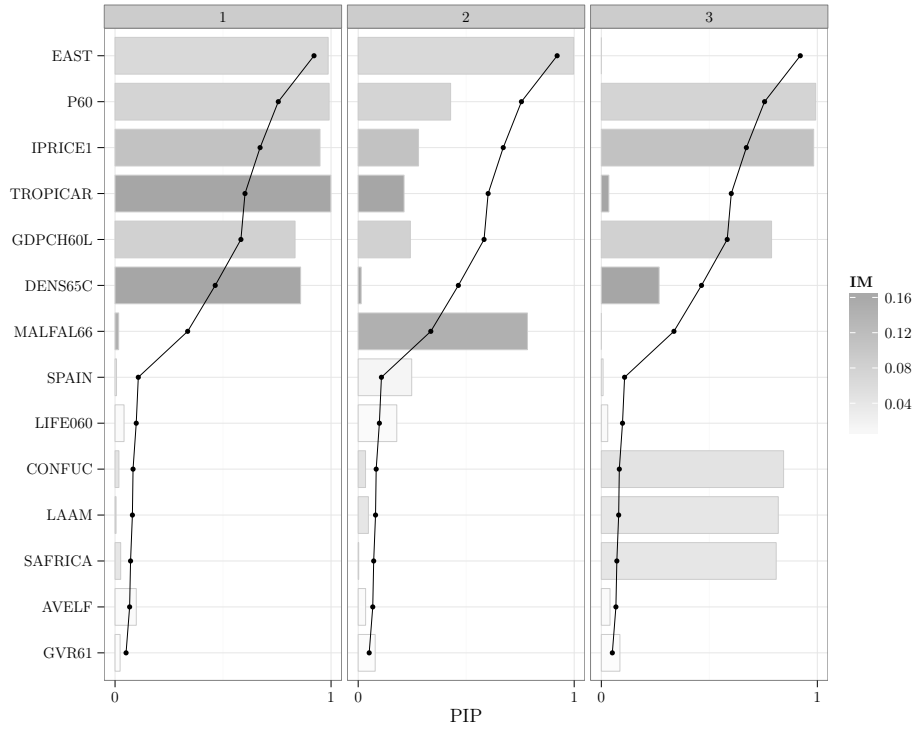

Figure 2: SDM dataset: PIPs in unclustered BMA (solid line) and by identified cluster (bars)

Wagner (2012) as well as Hofmarcher et al. (2014), for instance, find that the robustness of `MALFAL66` as a determinant of income growth differences across countries improves when estimation methods based on LASSO and elastic nets are used. In addition, the results in Schneider and Wagner (2012) and Hofmarcher et al. (2014) also indicate a loss of importance of `DENS65C` when methods implying a more stringent shrinkage are used in the estimation. These are precisely two of the variables which present the highest values of *IM* in our results, hinting to the fact that their relative importance depends on the *type* of model (as represented by the variable inclusion structure cluster) considered.

Such a pattern of substitutability across covariates is easily recognizable from the weighted correlation matrix of cluster-specific PIPs for the group of variables with the highest *IM* values, which is presented in Table 6. The correlation patterns present in the model space indicate that

Table 6: SDM dataset: Weighted correlation of cluster-specific PIPs for variables with $IM > 0.5 \max(IM)$

	DENS65C	GDPCH60L	IPRICE1	MALFAL66	TROPICAR
DENS65C	1.00	0.95	0.92	−0.93	0.97
GDPCH60L		1.00	1.00	−1.00	0.83
IPRICE1			1.00	−1.00	0.79
MALFAL66				1.00	−0.80
TROPICAR					1.00

`MALFAL66` tends to act as a substitute of the group of variables composed by `IPRICE1, TROPICAR, GDPCH60L` and `DENS65C`. The difference in average model size across these two important clusters in the space of posterior inclusion probability structures is in line with the strong impact of different parameter shrinkage approaches on the relative importance of the variables which is highlighted in previous literature. In addition, in their study of pairwise jointness measures, Doppelhofer and Weeks (2009a) report that `P60`, `IPRICE1`, `DENS65C` and `TROPICAR` exhibit significant negative bivariate jointness with `MALFAL66`, a result that can be easily reconciled with the output of our analysis. While Ley and Steel (2007) find very limited evidence for jointness structures in the SDM dataset, the only triplets of important variables for which disjointness is reported also involve `TROPICAR` and `MALFAL66`.

In spite of the fact that the third cluster that DPC identifies covers a very small part of the posterior mass over models, its PIP structure also reveals interesting patterns as compared to the other two clusters. In this group of models, two of the most relevant variables in terms of (unclustered) PIP, `EAST` and `TROPICAR`, lose their importance and their information is captured by a different set of geographical and religious variables (`CONFUC`, `LAAM` and `SAFRICA`). The results in Doppelhofer and Weeks (2009a) concerning the complementarity of `EAST` and `TROPICAR` and the substitutability of `EAST` with respect to `CONFUC`, `LAAM` and `SAFRICA` are perfectly in line with these results. In addition, Doppelhofer and Weeks (2009a) find the latter to be complements, which is also supported by the comparison of the PIPs in our third cluster with those in the other two.

4. Conclusions and Future Paths of Research

In this contribution we are concerned with covariate inclusion patterns of BMA exercises with large model spaces. Recent research on such *jointness* structures tends to choose a low-dimensional approach to such an analysis and thus concentrates on bivariate or trivariate approaches, by calculating the inclusion relationships of few explaining factors at a time. We propose a novel approach by utilizing LCA techniques and apply DPC to two well known datasets in the BMA growth literature. The clustering method put forward in our contribution aims at unveiling commonalities in the joint inclusion of variables and thus offering the applied econometrician

evidence about the competing structures (as formed by *groups* of variables that appear together) that are covered by the posterior over the model space.

Our results indicate that within the set of models sampled by the Markov chain in the BMA analysis of determinants of economic growth, several distinct clusters of models by covariate inclusion can be identified. For the FLS data, we identify seven clusters of models which differ in the inclusion structure for geographic, institutional and religious covariates. In contrast, the SDM dataset only reveals three latent classes with very different dependency structures. The inclusion of the variable measuring malaria prevalence is shown to vary strongly across clusters, with its effect on economic growth being captured often by other factors such as the fraction of tropical area and coastal population density.

We show that the study of dependency structures in covariate inclusion for large model spaces appears particularly relevant in order to understand the nature of the factors affecting global patterns of income growth. The proposed method lends itself to further straightforward expansions such as the use of low-dimensional jointness measures for the analysis of within-cluster inclusion patterns for small groups of covariates. The assessment of covariate inclusion clusters in the model space under different shrinkage priors can also shed light on the effects of multicollinearity on the robustness of economic growth determinants to model uncertainty.

In order to make our method and results comparable to those in the literature on *jointness* measures, we decided to follow a two-step procedure and use the clustering method on the model profiles visited by the Markov chain of the BMA procedure. The LCA and DPC methods proposed in this contribution would also lend themselves to create priors over suitable covariate combinations in the specifications that compose the model space. This path of further research, which we are pursuing at the moment, appears particularly promising in order to unify the literature on jointness and dilution priors in BMA applications.

References

Blaydes, L. and Linzer, D. A. (2008). The political economy of women's support for fundamentalist Islam. *World Politics*, 60(4):576–609.

Breen, R. (2000). Why is support for extreme parties underestimated by surveys? A latent class analysis. *British Journal of Political Science*, 30(2):375–382.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67(3):591–596.

Brock, W. A. and Durlauf, S. N. (2001). What have we learned from a decade of empirical research on growth? Growth empirics and reality. *The World Bank Economic Review*, 15(2):229–272.

Doppelhofer, G. and Weeks, M. (2009a). Jointness of growth determinants. *Journal of Applied Econometrics*, 24(2):209–244.

Doppelhofer, G. and Weeks, M. (2009b). Jointness of growth determinants: Reply to comments by Rodney Strachan, Eduardo Ley and Mark FJ Steel. *Journal of Applied Econometrics*, 24(2):252–256.

Doppelhofer, G., Weeks, M., et al. (2005). Jointness of determinants of economic growth. In *Money Macro and Finance (MMF) Research Group Conference 2005*, number 54. Money Macro and Finance Research Group.

Durlauf, S. N., Kourtellos, A., and Tan, C. M. (2008). Are any growth theories robust? *The Economic Journal*, 118(527):329–346.

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

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21(3):768–769.

George, E. (1999). Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6*, pages 157–177.

Glass, D. H. (2013). Confirmation measures of association rule interestingness. *Knowledge-Based Systems*, 44:65–77.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Hofmarcher, P., Crespo Cuaresma, J., Grün, B., and Hornik, K. (2014). Last night a shrinkage saved my life: Economic growth, model uncertainty and correlated regressors. *Mimeo, Vienna University of Economics and Business*.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. John Wiley & Sons, Inc., New York, USA.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In Stouffer, S. A., editor, *Measurement and Prediction*, pages 362–412. John Wiley & Sons, New York.

Ley, E. and Steel, M. F. (2007). Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics*, 29(3):476–493.

Ley, E. and Steel, M. F. (2009a). Comments on 'Jointness of growth determinants'. *Journal of Applied Econometrics*, 24(2):248–251.

Ley, E. and Steel, M. F. (2009b). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481).

Liverani, S., Hastie, D. I., Papathomas, M., and Richardson, S. (2013). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *arXiv preprint arXiv:1303.2836*.

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

Masanjala, W. and Papageorgiou, C. (2008). Rough and lonely road to prosperity: A reexamination of the sources of growth in Africa using Bayesian model averaging. *Journal of Applied Econometrics*, 23(5):671–682.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, 11(3):484–498.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet mixture models. Technical report 8, crism paper, University of Warwick. Centre for Research in Statistical Methodology.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Patterson, B. H., Dayton, C. M., and Graubard, B. I. (2002). Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association*, 97(459):721–741.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

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*, 94(4):813–835.

Schneider, U. and Wagner, M. (2012). Catching growth determinants with the adaptive LASSO. *German Economic Review*, 13:71–85.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Strachan, R. W. (2009). Comment on 'Jointness of growth determinants' by Gernot Doppelhofer and Melvyn Weeks. *Journal of Applied Econometrics*, 24(2):245–247.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

Vermunt, J. K. and Magidson, J. (2002). Latent class cluster analysis. In Hagenaars, J. and McCutcheon, A., editors, *Applied Latent Class Analysis*, pages 89–106. Cambridge University Press, Cambridge, UK.

A. Datasets

Table A.1: Variable names and descriptive statistics — FLS

	Abbreviation	Variable	Mean	Std. Dev.
1	Abslat	Absolute latitude	25.73	17.250
2	Age	Age	23.71	37.307
3	Area	Area (Scale Effect)	972.92	2051.976
4	BlMktPm	Black Market Premium	0.16	0.291
5	Brit	British Colony dummy	0.32	0.470
6	Buddha	Fraction Buddhist	0.06	0.184
7	Catholic	Fraction Catholic	0.42	0.397
8	CivlLib	Civil Liberties	3.47	1.712
9	Confucian	Fraction Confucian	0.02	0.087
10	EcoOrg	Degree of Capitalism	3.54	1.266
11	English	Fraction of Pop. Speaking English	0.08	0.239
12	EquipInv	Equipment investment	0.04	0.035
13	EthnoL	Ethnolinguistic fractionalization	0.37	0.296
14	Foreign	Fraction speaking foreign language	0.37	0.422
15	French	French Colony dummy	0.12	0.333
16	GDP60	GDP level in 1960	7.49	0.885
17	HighEnroll	Higher education enrollment	0.04	0.052
18	Hindu	Fraction Hindu	0.02	0.101
19	Jewish	Fraction Jewish	0.01	0.097
20	LabForce	Size labor force	9305.38	24906.056
21	LatAmerica	Latin American dummy	0.28	0.451
22	LifeExp	Life expectancy	56.58	11.448
23	Mining	Fraction GDP in mining	0.04	0.077
24	Muslim	Fraction Muslim	0.15	0.295
25	NequipInv	Non-Equipment Investment	0.15	0.055
26	OutwarOr	Outward Orientation	0.39	0.491
27	PolRights	Political Rights	3.45	1.896
28	Popg	Population Growth	0.02	0.010
29	PrExports	Primary exports, 1970	0.67	0.299
30	Protestants	Fraction Protestant	0.17	0.252
31	PrScEnroll	Primary School Enrollment, 1960	0.80	0.246
32	PublEdupct	Public Education Share	0.02	0.009
33	RevnCoup	Revolutions and coups	0.18	0.238
34	RFEXDist	Exchange rate distortions	121.71	41.001
35	RuleofLaw	Rule of law	0.55	0.335
36	Spanish	Spanish Colony dummy	0.22	0.419
37	stdBMP	SD of black-market premium	45.60	95.802
38	SubSahara	Sub-Saharan dummy	0.21	0.409
39	WarDummy	War dummy	0.40	0.494
40	WorkPop	Ratio workers to population	−0.95	0.189
41	y	GDP per capita growth	0.02	0.018
42	YrsOpen	Number of Years open economy	0.44	0.355

Table A.2: Variable names and descriptive statistics — SDM

	Abbreviation	Variable	Mean	Std. Dev.
1	ABSLATIT	Absolute latitude	23.21	16.843
2	AIRDIST	Air distance to big cities	4324.17	2613.763
3	AVELF	Ethnolinguistic fractionalization	0.35	0.302
4	BRIT	British colony	0.32	0.468
5	BUDDHA	Fraction Buddhist	0.05	0.168
6	CATH00	Fraction Catholic	0.33	0.415
7	CIV72	Civil liberties	0.51	0.326
8	COLONY	Colony dummy	0.75	0.435
9	CONFUC	Fraction Confucian	0.02	0.079
10	DENS60	Population density costal 1960's	108.07	201.445
11	DENS65C	Population density 1960	146.87	509.828
12	DENS65I	Interior density	43.37	88.063
13	DPOP6090	Population growth rate 1960-1990	0.02	0.009
14	EAST	East Asian dummy	0.11	0.319
15	ECORG	Capitalism	3.47	1.381
16	ENGFRAC	English-speaking population	0.08	0.252
17	EUROPE	European dummy	0.22	0.414
18	FERTLDC1	Fertility in 1960's	1.56	0.419
19	GDE1	Defense spending share	0.03	0.025
20	GDPCH60L	GDP 1960 (log)	7.35	0.901
21	GEEREC1	Public education spending share in GDP in 1960's	0.02	0.010
22	GGCFD3	Government consumption share deflated with GDP prices	0.05	0.039
23	GOVNOM1	Nominal government GDP share 1960's	0.15	0.058
24	GOVSH61	Government share of GDP	0.17	0.071
25	GR6096	Average growth rate of GDP per capita 1960-1996	0.02	0.019
26	GVR61	Government consumption share 1960's	0.12	0.075
27	H60	Higher education in 1960	0.04	0.050
28	HERF00	Religous intensity	0.78	0.193
29	HINDU00	Fraction Hindu	0.03	0.125
30	IPRICE1	Investment price	92.47	53.678
31	LAAM	Latin American dummy	0.23	0.421
32	LANDAREA	Land area	867188.52	1814688.290
33	LANDLOCK	Landlocked country dummy	0.17	0.378
34	LHCPC	Hydrocarbon deposits in 1993	0.42	4.351
35	LIFE060	Life expectancy in 1960	53.72	12.062
36	LT100CR	Fraction of land area near navigable water	0.47	0.380
37	MALFAL66	Malaria prevalence in 1960's	0.34	0.431
38	MINING	Fraction GDP in mining	0.05	0.077
39	MUSLIM00	Fraction Muslim	0.15	0.296
40	NEWSTATE	Time of independence	1.01	0.977
41	OIL	Oil-producing country dummy	0.06	0.233
42	OPENDEC1	(Imports+exports)/GDP	0.52	0.336
43	ORTH00	Fraction Orthodox	0.02	0.098
44	OTHFRAC	Fraction speaking foreign language	0.32	0.414
45	P60	Primary schooling 1960	0.73	0.293
46	PI6090	Average inflation 1960-1990	13.13	14.990
47	POP1560	Fraction population less than 15	0.39	0.075
48	POP60	Population in 1960	20308.08	52538.387
49	POP6560	Fraction population over 65	0.05	0.029
50	PRIEXP70	Primary exports in 1970	0.72	0.283
51	PRIGHTS	Political rights	3.82	1.997
52	PROT00	Fraction Protestant	0.14	0.285
53	RERD	Real exchange rate distortions	125.03	41.706
54	REVCOUP	Revolution and coups	0.18	0.232
55	SAFRICA	African dummy	0.31	0.464
56	SCOUT	Outward orientation	0.40	0.492
57	SIZE60	Size of the economy	16.15	1.820
58	SOCIALIST	Socialist dummy	0.07	0.254
59	SPAIN	Spanish colony	0.17	0.378
60	SQPI6090	Square of inflation 1960-1990	394.54	1119.699
61	TOT1DEC1	Terms of trade growth in 1960's	0.00	0.035
62	TOTIND	Terms of trade ranking	0.28	0.190
63	TROPICAR	Fraction of tropical area	0.57	0.472
64	TROPPOP	Fraction population in tropics	0.30	0.373
65	WARTIME	Fraction spent in war 1960-1990	0.07	0.152
66	WARTORN	War participation 1960-1990	0.40	0.492
67	YRSOPEN	Years open	0.36	0.344
68	ZTROPICS	Tropical climate zone	0.19	0.269

B. Posterior Inclusion Probabilities by Cluster

Table B.1: PIPs within detected clusters — FLS

	Overall	1	2	3	4	5	6	7	IM
GDP level in 1960	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Fraction Confucian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Life expectancy	0.97	0.99	0.99	0.97	1.00	0.82	1.00	1.00	0.00
Equipment investment	0.96	0.98	1.00	1.00	0.73	0.95	1.00	1.00	0.01
Sub-Saharan dummy	0.85	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.13
Rule of law	0.69	1.00	0.35	0.03	0.82	1.00	1.00	1.00	0.14
Fraction Muslim	0.67	0.90	0.75	0.96	0.02	0.00	0.66	1.00	0.13
Degree of Capitalism	0.59	0.99	0.05	0.04	0.95	1.00	0.12	1.00	0.21
Fraction Protestant	0.55	0.52	0.25	0.67	0.92	0.98	0.05	1.00	0.07
Non-Equipment Investment	0.54	0.81	0.14	0.24	0.63	0.90	0.40	1.00	0.09
Fraction GDP in mining	0.48	0.30	0.61	0.36	0.95	0.25	0.96	1.00	0.06
Number of Years open economy	0.37	0.10	0.80	1.00	0.00	0.02	0.00	1.00	0.16
Black Market Premium	0.21	0.19	0.12	0.01	0.57	0.08	0.71	1.00	0.04
Latin American dummy	0.20	0.02	0.11	0.00	1.00	0.22	0.54	1.00	0.10
Fraction Hindu	0.19	0.04	0.00	0.00	0.43	0.63	1.00	1.00	0.09
Primary School Enrollment, 1960	0.17	0.02	0.20	0.39	0.00	0.00	0.90	1.00	0.05
Fraction Buddhist	0.15	0.29	0.07	0.06	0.06	0.00	0.15	1.00	0.02
Fraction Catholic	0.10	0.03	0.01	0.04	0.01	0.78	0.00	1.00	0.06
Civil Liberties	0.10	0.02	0.04	0.19	0.02	0.03	0.82	1.00	0.04
Size labor force	0.09	0.01	0.01	0.04	0.14	0.04	1.00	1.00	0.05
Ethnolinguistic fractionalization	0.08	0.01	0.02	0.00	0.04	0.04	1.00	1.00	0.05
Higher education enrollment	0.07	0.01	0.00	0.00	0.08	0.00	1.00	1.00	0.05
Political Rights	0.05	0.04	0.01	0.06	0.03	0.13	0.12	1.00	0.00
Fraction of Pop. Speaking English	0.05	0.03	0.01	0.00	0.00	0.00	0.55	1.00	0.02
Primary exports, 1970	0.04	0.02	0.07	0.10	0.00	0.03	0.00	1.00	0.00
French Colony dummy	0.04	0.01	0.01	0.00	0.00	0.00	0.54	1.00	0.02
Spanish Colony dummy	0.04	0.01	0.00	0.00	0.01	0.01	0.54	1.00	0.02
British Colony dummy	0.04	0.01	0.00	0.00	0.00	0.00	0.54	1.00	0.02
Exchange rate distortions	0.03	0.01	0.01	0.17	0.00	0.00	0.02	1.00	0.01
Outward Orientation	0.03	0.00	0.00	0.01	0.01	0.00	0.41	1.00	0.01
Age	0.03	0.02	0.02	0.02	0.08	0.00	0.05	1.00	0.00
War dummy	0.02	0.01	0.03	0.00	0.01	0.03	0.02	1.00	0.00
Public Education Share	0.02	0.00	0.00	0.00	0.08	0.00	0.05	1.00	0.00
Fraction speaking foreign language	0.02	0.01	0.00	0.06	0.00	0.00	0.00	1.00	0.00
SD of black-market premium	0.01	0.00	0.01	0.02	0.00	0.00	0.00	1.00	0.00
Absolute latitude	0.01	0.01	0.00	0.01	0.00	0.00	0.00	1.00	0.00
Ratio workers to population	0.01	0.01	0.00	0.01	0.00	0.00	0.00	1.00	0.00
Population Growth	0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Revolutions and coups	0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Area (Scale Effect)	0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Fraction Jewish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00

Table B.2: PIPs within detected clusters — SDM

	Overall	1	2	3	IM
East Asian dummy	0.92	0.99	1.00	0.00	0.06
Primary schooling 1960	0.76	0.99	0.43	0.99	0.08
Investment price	0.67	0.95	0.28	0.98	0.11
Fraction of tropical area	0.60	1.00	0.21	0.03	0.17
GDP 1960 (log)	0.58	0.83	0.24	0.79	0.08
Population density 1960	0.46	0.86	0.01	0.27	0.17
Malaria prevalence in 1960's	0.34	0.02	0.78	0.00	0.14
Spanish colony	0.11	0.01	0.25	0.01	0.01
Life expectancy in 1960	0.10	0.04	0.18	0.03	0.00
Fraction Confucian	0.08	0.02	0.03	0.84	0.04
Latin American dummy	0.08	0.00	0.05	0.82	0.04
African dummy	0.07	0.03	0.00	0.81	0.04
Ethnolinguistic fractionalization	0.07	0.10	0.03	0.04	0.00
Government consumption share 1960's	0.05	0.02	0.08	0.09	0.00
Political rights	0.05	0.09	0.00	0.00	0.00
Years open	0.05	0.04	0.05	0.04	0.00
Fraction Muslim	0.04	0.03	0.03	0.15	0.00
Fraction Buddhist	0.04	0.04	0.00	0.28	0.00
Fraction GDP in mining	0.04	0.03	0.02	0.15	0.00
Population density costal 1960's	0.03	0.06	0.00	0.08	0.00
Higher education in 1960	0.03	0.03	0.03	0.00	0.00
(Imports+exports)/GDP	0.03	0.03	0.02	0.02	0.00
Government share of GDP	0.02	0.01	0.02	0.14	0.00
Fraction speaking foreign language	0.02	0.02	0.02	0.03	0.00
Primary exports in 1970	0.02	0.00	0.04	0.00	0.00
Air distance to big cities	0.02	0.04	0.00	0.00	0.00
Real exchange rate distortions	0.02	0.02	0.01	0.02	0.00
Fraction population less than 15	0.02	0.03	0.01	0.00	0.00
Government consumption share deflated with GDP prices	0.01	0.01	0.00	0.10	0.00
Fraction Protestant	0.01	0.01	0.02	0.01	0.00
Fraction population in tropics	0.01	0.01	0.01	0.03	0.00
Absolute latitude	0.01	0.01	0.01	0.00	0.00
Civil liberties	0.01	0.01	0.00	0.00	0.00
Colony dummy	0.01	0.01	0.01	0.00	0.00
Revolution and coups	0.01	0.01	0.01	0.00	0.00
Outward orientation	0.01	0.01	0.00	0.00	0.00
Fraction Hindu	0.01	0.01	0.00	0.00	0.00
Average inflation 1960-1990	0.01	0.01	0.00	0.00	0.00
European dummy	0.00	0.00	0.00	0.01	0.00
Size of the economy	0.00	0.01	0.00	0.00	0.00
Hydrocarbon deposits in 1993	0.00	0.01	0.00	0.00	0.00
Fertility in 1960's	0.00	0.01	0.00	0.00	0.00
Fraction population over 65	0.00	0.00	0.00	0.00	0.00
British colony	0.00	0.01	0.00	0.00	0.00
English-speaking population	0.00	0.00	0.01	0.00	0.00
Square of inflation 1960-1990	0.00	0.01	0.00	0.00	0.00
Defense spending share	0.00	0.01	0.00	0.00	0.00
Landlocked country dummy	0.00	0.01	0.00	0.00	0.00
Religous intensity	0.00	0.01	0.00	0.00	0.00
Oil-producing country dummy	0.00	0.01	0.00	0.00	0.00
Time of independence	0.00	0.01	0.00	0.00	0.00
Socialist dummy	0.00	0.01	0.00	0.00	0.00
Fraction Catholic	0.00	0.00	0.00	0.01	0.00
Population growth rate 1960-1990	0.00	0.00	0.00	0.00	0.00
Nominal government GDP share 1960's	0.00	0.01	0.00	0.00	0.00
Public education spending share in GDP in 1960's	0.00	0.00	0.00	0.00	0.00
Capitalism	0.00	0.01	0.00	0.00	0.00
Terms of trade growth in 1960's	0.00	0.00	0.00	0.00	0.00
Tropical climate zone	0.00	0.00	0.00	0.00	0.00
Fraction spent in war 1960-1990	0.00	0.00	0.00	0.00	0.00
War participation 1960-1990	0.00	0.00	0.00	0.00	0.00
Land area	0.00	0.00	0.00	0.00	0.00
Population in 1960	0.00	0.00	0.00	0.00	0.00
Fraction Orthodox	0.00	0.00	0.00	0.00	0.00
Fraction of land area near navigable water	0.00	0.00	0.00	0.00	0.00
Interior density	0.00	0.00	0.00	0.00	0.00
Terms of trade ranking	0.00	0.00	0.00	0.00	0.00