European Economic Review 81 (2016) 189–202

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

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], Mathias Moser [a]

[a] Department of Economics, Vienna University of Economics and Business (WU), Welthandelsplatz 1, 1020 Vienna, Austria
[b] Austrian Institute of Economic Research (WIFO), Austria
[c] International Institute of Applied System Analysis (IIASA), Austria
[d] Wittgenstein Centre for Demography and Global Human Capital (WIC), Austria
[e] Department of Applied Statistics, Johannes Kepler University Linz (JKU), Austria

ARTICLE INFO

Available online 3 April 2015

JEL classification:
C11
C21
O47

Keywords:
Economic growth determinants
Bayesian model averaging
Latent class analysis
Dirichlet processes

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.

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 assesses 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. Fernández 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

* Corresponding author at: Department of Economics, Vienna University of Economics and Business (WU), Welthandelsplatz 1, 1020 Vienna, Austria. Tel.: +43 131336 4530; fax: +43 131336 728.
E-mail address: jcrespo@wu.ac.at (J. Crespo Cuaresma).

http://dx.doi.org/10.1016/j.euroecorev.2015.03.009
0014-2921/

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 Doppelhofer and Weeks in a working paper of 2005 (Doppelhofer and Weeks, 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 desirable 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 is not accounted for in the indices of Ley and Steel (2007) or Strachan (2009).[1]

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 Fernández et al., 2001; 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., 2015).

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 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.

[1] Interestingly, the measures proposed by Doppelhofer and Weeks (2009a), Ley and Steel (2007) and Strachan (2009) were independently developed earlier in the context of data mining. The statistic of Doppelhofer and Weeks (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's (2009) measure is closely related to the so-called *two-way support* (see Tan et al., 2004; Glass, 2013).