The Determinants of Long-Run Economic Growth: A Conceptually and Computationally Simple Approach

Jaroslava Hlouskova[a] and Martin Wagner[b,c]

JEL-Classification: C31, C52, O11, O18, O47
Keywords: economic growth, economic convergence, frequentist model averaging, growth regressions, principal components augmented regression

1. Introduction

Durlauf, Johnson, and Temple (2005) forcefully argue that the empirical analysis of economic growth is one of the areas of economics in which progress seems to be hardest to achieve and where only few definite results are established. Large sets of potentially relevant candidate variables have been used in empirical analysis to capture what Brock and Durlauf (2001) refer to as theory open endedness of economic growth. A large variety of different approaches has been and is used to identify variables relevant for economic growth. Many of the contributions employ model averaging estimators to tackle the uncertainty about the relevant variables. Sala-i-Martin (1997a) runs two million regressions and uses a modification of the extreme bounds test of Leamer (1985), used in the growth context earlier also by Levine and Renelt (1992), to single out what he calls 'significant' variables. Fernandez, Ley, and Steel (2001) and Sala-i-Martin, Doppelhofer, and Miller (2004) use Bayesian model averaging (BMA) techniques to identify important growth determinants. The former perform Bayesian averaging of Bayesian estimates, introduced by Leamer (1978), whereas the latter perform Bayesian averaging of classical estimates, proposed by Raftery (1995).

a Institute for Advanced Studies, Vienna, Austria.
b Faculty of Statistics, Technical University Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany; & Frisch Centre for Economic Research, Oslo, Norway. Email: mwagner@statistik.tu-dortmund.de.
c We thank Mark Steel for providing us with posterior means and posterior inclusion probabilities for the Fernandez, Ley, and Steel (2001) data. Furthermore, we thank the editor and an anonymous referee for very helpful comments that led to significant improvements of the paper. Partial financial support from the Anniversary Fund of the Oesterreichische Nationalbank under Grant Nr. 11688 is gratefully acknowledged. The usual disclaimer applies.

More recently Magnus, Powell, and Prüfer (2010) provide a computationally simple approach to BMA that bears some resemblences with our approach (compare the discussion in Section 2.3 below).

Alternatively, several approaches have been followed in the growth regressions literature that instead intend to provide only one selected model together with one set of parameter estimates. Hendry and Krolzig (2004) use, similar to Hoover and Perez (2004), a general-to-specific modeling strategy to cope with the large amount of regressors while avoiding the estimation of a large number of equations. Clearly, also in a general-to-specific analysis a certain number of regressions, typically greater than one, has to be estimated. Schneider and Wagner (2012) apply the adaptive LASSO estimator, which has the computational cost of one OLS estimation including all variables, in the growth regressions context.

The typical situation in the empirical analysis of economic growth is the availability of a data set where the number of variables is close to (or in one of our data sets, which is an extension and update of the Sala-i-Martin, Doppelhofer, and Miller (2004) data set, precisely equal to) the number of countries considered. Given the above mentioned uncertainty about which variables are relevant the researcher faces a difficult situation. Any regression including only few regressors (which are usually the ones considered with high probability in BMA growth regression exercises) risks to suffer from omitted variables biases. On the other hand any regression with many variables included runs the risk of large estimator variance, in particular in case of near multi-collinearity of the data. As an example, for one of the data sets used in this paper, the one originally used in Sala-i-Martin, Doppelhofer, and Miller (2004), the reciprocal condition number of the full regressor matrix including all 67 regressors is 9.38×10^{-20}. In the extended and updated data set the number of variables is, as mentioned before, equal to the number of countries which allows for perfect (but meaningless) fit from the full regression. Thus, it is important to find a good trade-off between parsimony of the regression (to achieve low estimator variance but potentially high bias) and the inclusion of as many variables as possible (to achieve low bias at the price of potentially high variance). An optimal positioning on the bias variance trade-off is achieved, *given* a choice concerning which variable(s) the researcher is interested in, by so-called principal components augmented regressions (discussed in detail in Section 2 and considered also in Wagner and Hlouskova (2012)). In the simplest case, when one is interested in understanding the individual variables' conditional effect on growth (which is usually the object of interest in the empirical growth literature) one estimates regressions of GDP growth on this variable and principal components extracted from all other variables. We refer to such regressions as principal components augmented regressions (PCARs). These regressions contain (in a sense made precise in the following section) as much information as possible (for a given number of included additional regressors) from the data set and thus are least prone to omitted variables bias whilst being well-conditioned since the principal components are mutually orthogonal.[1]

Clearly, principal components augmentation is also useful when combined with model averaging, be it frequentist or Bayesian. Once the set of *focus* variables of interest is specified, which can be motivated by a specific theoretical model or also by the quest of understanding the contribution to growth of certain factors like human capital, for which several proxies may be available in the data set, one can perform model averaging over these focus variables, whilst including in all regressions principal components computed from the remaining variables. Compared to the usual BMA analysis in which the priors are set such that the expected prior model size is very small the augmentation by principal components makes model averaging estimators less prone to suffer from omitted variables biases whilst keeping estimator variance low. In the empirical analysis in this paper we partition the extended and updated growth data set into twelve groups and perform model averaging over principal components augmented regressions for these groups. The selection of a relatively small set of *focus* variables drastically reduces the model space (to two to the power number of focus variables) and allows for estimating all models rather than just a random selection of models, as would be the case in the unrestricted model space with no partitioning of the variables into focus and auxiliary variables. The coefficients to the focus variables in a PCAR measure the effect on growth of each of these variables when considered jointly, whilst in addition conditioning on the information contained in the principal components and are in this sense robust estimates.

In this paper we perform model averaging based upon PCAR in a frequentist framework, using recent advantages in the statistics literature which allow to perform valid frequentist inference in a model averaging context, see in particular Claeskens and Hjort (2008, Section 7.5). In our analysis we consider four different weighting schemes. One, as a benchmark, uses equal weights for each model and the three others are based on weights derived from information criteria computed for the individual models. These are smoothed AIC and smoothed BIC weights considered by Buckland, Burnham, and Augustin (1997) and studied in detail also in Claeskens and Hjort (2008) and Mallows

1 Thus, problems of multi-collinearity can only arise if the variable whose effect is studied is highly correlated with the principal components, which can be easily checked upfront.

model averaging (MMA) advocated by HANSEN (2007). Furthermore, we calculate frequentist analogs to quantities considered to be informative in a Bayesian model averaging framework. We compute e.g. for any given weighting scheme the so-called inclusion weight as the classical counterpart of the Bayesian posterior inclusion probability of a variable.[2] Clearly, either Bayesian or frequentist model averaging could be employed and we have chosen frequentist model averaging to try popularizing it in the growth empirics literature, in which typically Bayesian or pseudo-Bayesian model averaging approaches are employed to date.[3]

We apply the methodology to three data sets. Two of them have been widely studied and are used to 'benchmark' our findings. These are the data of SALA-I-MARTIN, DOPPELHOFER, and MILLER (2004) and of FERNANDEZ, LEY, and STEEL (2001). For these data sets we compute the (conditional) individual effects of all variables and compare the findings with those in the original papers as well as with the estimates found from the simple bivariate regressions (in order to highlight the ensuing omitted variables biases). The third data set is, as has been mentioned already, an updated and extended data set based on the SALA-I-MARTIN, DOPPELHOFER, and MILLER (2004) data with the number of variables equal to the number of countries. This clearly necessitates thinking about how to perform regression analysis in a sensible way. For this data set we also compute the variables' individual conditional effects, but in addition as mentioned also study the joint effects of 12 groups of variables (comprising between 4 and 10 variables) by combining PCAR with model averaging. The 'final' analysis then considers all significant variables from the 12 groups jointly. The results show that estimating well-behaved regressions that include the relevant information from the available variables is important to obtain robust estimates of the variables' effect on economic growth. For the SALA-I-MARTIN, DOPPELHOFER, and MILLER (2004) data set our findings differ from those in the original paper in that we find more core economic variables related to economic growth and an implied convergence speed that is about twice as high as found by SALA-I-MARTIN, DOPPELHOFER, and MILLER (2004). For the FERNANDEZ, LEY, and

2 Similarly one can also compute the distribution of model weights over model sizes, see also WAGNER and HLOUSKOVA (2012). That paper also performs model averaging over PCARs with the focus variables taken from the results of the partial LASSO estimates in SCHNEIDER and WAGNER (2012). KOOP and POTTER (2004) perform Bayesian model averaging over principal components computed from all variables and employ their approach to a macroeconomic forecasting exercise. Thus, in our language they include zero focus variables.

3 WAGNER and ZEUGNER (2012) develop a Bayesian framework for model averaging with principal components augmented regressions.

Steel (2001) data set our findings very strongly coincide with the BMA findings obtained in the original paper, yet they are obtained at a negligible fraction of computational cost (and are independent of any – in the growth literature inevitably – ad hoc choices concerning priors on coefficients or model sizes). For the extended data set we consider the joint effect of thematic groups of variables in addition to each variables' effect. From an individual perspective the following 7 variables are significantly related (at the 5% level) to long-run economic growth: initial GDP (with an implied convergence speed that is about three times as high as in Sala-i-Martin , Doppelhofer, and Miller, 2004), the male labor force participation rate, the fraction of Confucian in the population, the share of government spending in GDP, the relative investment price, the share of mining in GDP and mobile phone subscribers per person (as a proxy for modern communication technologies). When considering the variables separated into 12 thematic groups there are 5 groups in which only one variable is found to be significant and 4 groups in which no variable is significant (information technology, education, health, historical and political data) and altogether there are only 16 variables significant (when considering the 10% and only 14 when considering the 5% level). Analyzing these 16 variables' joint conditional effect on economic growth results again in 7 variables significantly related to economic growth at the 5% level. Combining the different pieces of evidence (both the individual and the group-wise conditional effects and their significance) leads to only 6 variables related to economic growth: initial GDP, the population growth rate, the share of mining in GDP, the losses due to climate disasters, the relative investment price and the share of Confucian in the population. All coefficients have the correct sign and sensible magnitudes. Furthermore, the inclusion weights (which are the frequentist counterparts of the Bayesian posterior inclusion probabilities) of the significant variables are typically very high, confirming their importance from another angle. The findings show that appropriate conditioning on the relevant information in well-behaved regressions can help to uncover the determinants of economic growth in a computationally extremely cheap fashion.

The paper is organized as follows: The following section contains a description of the econometric approach. In Section 3 the results obtained with the three mentioned data sets are discussed in three subsections. Section 4 briefly summarizes and concludes. The appendix contains a detailed description of the extended data set including the data sources. Two supplementary appendices as well as the dataset are available as supplementary material on the website of the Journal (www.sjes.ch). Appendix B contains additional empirical results and Appendix C describes the computation of (frequentist) confidence intervals for model average estimators.

2. Description of the Econometric Approach

2.1 Principal Components Augmented Regressions

Let $y \in \mathbb{R}^N$ denote the variable to be explained (in our application average per capita GDP growth for N countries) and collect all explanatory variables in $X = [X_1 X_2] \in \mathbb{R}^{N \times k}$, with the focus variables given in $X_1 \in \mathbb{R}^{k_1}$ and the auxiliary variables in $X_2 \in \mathbb{R}^{k_2}$ with $k = k_1 + k_2$. Without loss of generality we assume that all variables have zero mean, since in growth regressions an intercept is typically included. As is well known, by the Frisch-Waugh theorem, the regressions can therefore equivalently be estimated with demeaned variables. The regression including all variables is given by

$$y = X_1\beta_1 + X_2\beta_2 + u. \tag{1}$$

The information for regression (1) contained in X_2 is equivalently summarized in the set of (orthogonal) principal components computed from X_2. The principal components are the set of transformed variables $\breve{X}_2 = X_2 O$, with $O \in \mathbb{R}^{k_2 \times k_2}$ computed from the eigenvalue decomposition of $\Sigma_{X_2} = X_2' X_2$ (due to the assumption of zero means):

$$\begin{aligned} \Sigma_{X_2} &= X_2' X_2 = O \Lambda O' = [O_1 O_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} O_1' \\ O_2' \end{bmatrix} \\ &= O_1 \Lambda_1 O_1' + O_2 \Lambda_2 O_2', \end{aligned} \tag{2}$$

where $O'O = OO' = I_{k_2}$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_{k_2})$, $\lambda_i \geq \lambda_{i+1}$ for $i = 1, \ldots, k_2 - 1$. The partitioning into variables with subscripts 1 and 2 in (2) will become clear in the discussion below. From (2) the orthogonality of the variables in $\breve{X}_2$ is immediate, since $\breve{X}_2' \breve{X}_2 = \Lambda$.

Let us consider the case of multi-collinearity in X_2 first (which e.g. necessarily occurs when $k_2 > N$) and let us denote the rank of X_2 with r. Take $\Lambda_1 \in \mathbb{R}^{r \times r}$, hence $\Lambda_2 = 0$ and $X_2' X_2 = O_1 \Lambda_1 O_1'$. The space spanned by the columns of $X_2 \in \mathbb{R}^{N \times k_2}$ coincides with the space spanned by the orthogonal regressors $\tilde{X}_2 = X_2 O_1 \in \mathbb{R}^{N \times r}$, i.e. with the space spanned by the r principal components. Thus, in this case regression (1) is equivalent to the regression

$$y = X_1\beta_1 + \tilde{X}_2\tilde{\beta}_2 + u \tag{3}$$

in the sense that both regressions lead to exactly the same fitted values and residuals. Furthermore, in case $[X_1 \ \tilde{X}_2]$ has full rank, regression (3) leads to unique coefficient estimates of β_1 and β_2. Therefore, the use of principal components is one computationally efficient way of overcoming multicollinearity. The facts discussed above are, of course, well known results from linear regression theory.

Using principal components in case of full rank of X_2 and hence of Σ_{X_2}, however, also has a clear interpretation and motivation. In such a situation replacing X_2 by the first r principal components $\tilde{X}_2$ leads to a regression where the set of regressors $\tilde{X}_2$ spans that r-dimensional subspace of the space spanned by the columns of X_2 which minimizes the approximation error to the full space in a least squares sense. More formally the following holds true, resorting here to the population level.[4] Let $x_2 \in \mathbb{R}^{k_2}$ be a mean zero random vector with covariance matrix Σ_{X_2} (using here the same notation for both the sample and the population covariance matrix for simplicity). For a given value of r consider a decomposition of x_2 into a factor component and a noise component, i.e. a decomposition $x_2 = Lf + \nu$, where $f \in \mathbb{R}^r$ is random, $L \in \mathbb{R}^{k_2 \times r}$ is non-random and $\nu \in \mathbb{R}^{k_2}$ is noise. If the decomposition is such that the factors f and the noise ν are uncorrelated, i.e. orthogonal, then $\Sigma_{X_2} = L\Sigma_f L' + \Sigma_\nu$, with Σ_f denoting the covariance matrix of f and Σ_ν denoting the covariance matrix of ν. Principal components analysis performs such an orthogonal decomposition of x_2 into Lf and ν so that the noise component is as small as possible, i.e. it minimizes $\mathbb{E}(\nu'\nu) = tr(\Sigma_\nu)$. As is well known, the solution is given by $f = O_1'x_2$, $L = O_1$, with $O_1 \in \mathbb{R}^{k_2 \times r}$ and $\nu = O_2O_2'x_2$, using the same notation for the spectral decomposition as above.

Therefore, including only r principal components $\tilde{X}_2$ instead of all regressors X_2 has a clear interpretation: the principal components augmented regression (PCAR) includes 'as much information as possible' with r linearly independent regressors contained in the space spanned by the columns of X_2. We write the PCAR as:

$$y = X_1\beta_1 + \tilde{X}_2\tilde{\beta}_2 + \tilde{u}, \tag{4}$$

neglecting in the notation the dependence upon the (chosen) number of principal components r, but indicating by using $\tilde{u}$ the fact that the residuals of (4) in general differ from the residuals of (3). Including only the information contained in the first r principal components of X_2 in the regression when the rank of X_2 is

4 I.e. we now consider the k_2-dimensional random vector x_2 for which a sample X_2 of size N is available.

larger than r of course amounts to neglecting some information and hence leads to different, larger residuals. Thus, in comparison to the full regression (1), if it can be estimated, the PCAR regression will potentially incur some bias in the estimates which has to be weighed against the benefits of a lower estimator variance. It is immediate that the choice of r is a key issue. The larger r, the more information is included but the fewer degrees of freedom are left (i.e. a lower bias but a higher variance). The choice concerning r is generally based on the eigenvalues λ_i, where 'large' eigenvalues are typically attributed to the factors and 'small' ones to the noise. The literature provides many choices in this respect and we have experimented with several thereof.[5] A classical, descriptive approach is given by the so-called variance proportion criterion (VPC),

$$r_{VPC(\alpha)} = \min_{j=1,\dots,k_2} \left\{ j \mid \frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{k_2} \lambda_i} \geq 1-\alpha \right\}, \tag{5}$$

with $\alpha \in [0,1]$. Thus, $r_{VPC(\alpha)}$ is the smallest number of principal components such that a fraction $1-\alpha$ of the variance is explained. For our applications setting $\alpha = 0.2$, i.e. explaining 80% of the variance, leads to reasonable numbers of principal components included. In the context of growth regressions there is no underlying theoretical factor model explaining the second-moment structure of the auxiliary variables X_2 available. Thus, any choice has to a certain extent heuristic character and has to trade off good approximation (necessary to capture the information contained in all explanatory variables to have small bias) with a sufficiently small number of principal components (necessary for well-behaved regression analysis with low estimator variance).

When computing the principal components from the regressors $X_2 \in \mathbb{R}^{N \times k_2}$ in our growth application, we split this set of variables in 2 groups. One group contains the quantitative or cardinal variables and the other includes the dummy or qualitative variables. We separate these two groups to take into account their

5 In addition to the results reported in the paper the number of principal components has also been determined using the testing approaches of LAWLEY and MAXWELL (1963), MALINOWSKI (1989), FABER and KOWALSKI (1997), SCHOTT (2006) and KRITCHMAN and NADLER (2008). In a variety of simulations, however, the VPC criterion and a simple eigenvalue test based on the correlation matrix (see below) have performed best. Using the VPC criterion can also be interpreted as a *regularization* device for linear regression when the data are in fact not generated by a factor model (as is potentially the case in our application), whereas the mentioned tests have been derived explicitly for factor models.

different nature when computing principal components. For both groups the principal components are computed based on the correlation matrix of the variables. Computing the principal components based on the correlation matrix is especially important for the group of quantitative variables. These differ considerably in magnitude, due to their scaling which we keep unchanged for the Fernandez, Ley, and Steel (2001) and Sala-i-Martin, Doppelhofer, and Miller (2004) data to use exactly the same data as in these papers. Computing the principal components based on the covariance matrix leads in such a case to essentially fitting the 'large' variables, whereas the computation based on the correlation matrix corrects for scaling differences and leads to a scale-free computation of the principal components. To be precise, in this case a so-called weighted principal components problem is solved in which the function minimized is given by $\mathbb{E}(\nu' Q \nu) = tr(Q\Sigma_\nu)$ with $Q = diag(\sigma^{-2}_{x_2,1}, \ldots, \sigma^{-2}_{x_2,k_2})$, neglecting here for simplicity the separation of the variables in X_2 in quantitative and dummy variables.[6] This leads to $f = O_1' Q^{1/2} x_2$, $L = Q^{-1/2} O_1$ and $\nu = Q^{-1/2} O_2 O_2' Q^{1/2} x_2$, i.e. the auxiliary regressors are given by $\tilde{X}_2 = X_2 Q^{1/2} O_1$.

2.2 *Model Averaging of Principal Components Augmented Regressions*

For a chosen number of principal components, the PCAR (4) allows to estimate the conditional effects of the variables X_1 taking into account the relevant information contained in X_2 and summarized in $\tilde{X}_2$. As discussed in the introduction, one can also use (4) as a starting point to consider model averaging. By resorting to PCAR analysis, the number of regressions to be computed to estimate all sub-models is reduced from 2^k to 2^{k_1} if one computes all sub-models with respect to the focus variables. The number of regressions can be reduced further by partitioning the set of focus variables $X_1 = [X_{11} X_{12}]$, with $X_{11} \in \mathbb{R}^{N \times k_{11}}$ included in each regression and $X_{12} \in \mathbb{R}^{N \times k_{12}}$, where $k_1 = k_{11} + k_{12}$, containing the variables in- or excluded in the sub-models estimated. This further reduces the number of regressions to be computed to $2^{k_{12}}$ and makes it even more likely that all sub-models can be estimated. As already mentioned in the introduction, the small number

6 Performing the spectral decomposition on a correlation matrix allows for another simple descriptive criterion concerning the number of principal components. By construction the trace of a correlation matrix equals its dimension, i.e. is equal to k_2. Therefore, if all k_2 eigenvalues were equally large, they all would equal 1. This suggests to include as many principal components as there are eigenvalues larger than 1, i.e. to consider the eigenvalues larger than 1 as big and those smaller than 1 as small. The results correspond closely to those obtained with VPC_α with $\alpha = 0.2$.

of models has the advantage, for both frequentist and Bayesian approaches, that inference need not be based on estimation results obtained only on subsets of the model space containing mainly small models.[7]

We denote the sub-model regressions, based on the partitioning of (4) as

$$y = X_{11}\beta_{11}(j) + X_{12}(j)\beta_{12}(j) + \tilde{X}_2\beta_2(j) + \tilde{u}(j). \tag{6}$$

The sub-models $\mathcal{M}_j$ are indexed with $j = 1,\ldots,2^{k_{12}}$, where $X_{12}(j)$ denotes the j-th subset of X_{12}. The corresponding coefficient estimates are given by

$$\hat{\beta}(j) = [\hat{\beta}_{11}(j)'\hat{\beta}_{12}(j)'\hat{\beta}_2(j)']' \in \mathbb{R}^{k_{11}+k_{12}+r}.$$

Here, with some imprecision in notation we include in $\hat{\beta}_{12}(j) \in \mathbb{R}^{k_{12}}$ zero entries corresponding to all variables not included in model $\mathcal{M}_j$, whereas in (6) the dimension of $\beta_{12}(j)$ equals the number of variables of X_{12} included. We are confident that this does not lead to any confusion.[8] Furthermore, note already here that the regression including all explanatory variables, i.e. all variables in X_{12}, will be referred to as full model in the empirical application. Model average coefficients $\hat{\beta}^w$ are computed as weighted averages of the coefficient estimates of the sub-regressions, i.e.

$$\hat{\beta}^w = \sum_{j=1}^{2^{k_{12}}} w(j)\hat{\beta}(j), \tag{7}$$

$$\text{with } \ 0 \leq w(j) \leq 1 \ \text{ and } \ \sum_{j=1}^{2^{k_{12}}} w(j) = 1.$$

We consider four different weighting schemes: equal weights, smoothed AIC (S-AIC) and smoothed BIC (S-BIC) weights considered by BUCKLAND, BURNHAM, and AUGUSTIN (1997) and discussed in detail in CLAESKENS and HJORT

7 This statement has to be interpreted in the following way: Inference is based on a different type of subset of the model space, since all information contained in X_2 is summarized in $\tilde{X}_2$ and taken into account. This conditional model space, after purging the effects of $\tilde{X}_2$, however, can then be fully exhausted.

8 Note furthermore that we can, since $\tilde{X}_2$ is included in each regression, invoke the Frisch-Waugh theorem and entirely equivalently consider model averaging only for the regressions of y on X_{11} and the subsets of X_{12} by considering the residuals of the regressions of y, X_{11} and X_{12} on $\tilde{X}_2$. This equivalent interpretation highlights again that the inclusion of $\tilde{X}_2$ conditions on the 'relevant' information contained in X_2.

(2008) and MMA weights as considered in Hansen (2007). Equal weighting assigns weights $w(j) = 1/2^{k_{12}}$ to each of the models. By definition, this model averaging scheme does not allocate model weights according to any measure of quality of the individual models and thus serves more as a baseline averaging scheme. The other model averaging schemes base the model weights on different information criteria to give higher weights to models showing better performance in the 'metric' of the underlying information criterion. The S-AIC and S-BIC averaging schemes base their weights on the information criteria AIC and BIC, defined here as

$$AIC(j) = N \ln \hat{\sigma}_j^2 + 2 \dim(\mathcal{M}_j) \text{ and } BIC(j) = N \ln \hat{\sigma}_j^2 + \ln N \dim(\mathcal{M}_j),$$

where $\hat{\sigma}_j^2$ is the estimated residual variance of $\mathcal{M}_j$. Based on these the corresponding model weights are computed as

$$w(j) = \frac{\exp\{-\frac{1}{2} AIC(j)\}}{\sum_m \exp\{-\frac{1}{2} AIC(m)\}}$$

for S-AIC weights and as

$$w(j) = \frac{\exp\{-\frac{1}{2} BIC(j)\}}{\sum_m \exp\{-\frac{1}{2} BIC(m)\}}$$

for S-BIC weights. Hansen (2007), based on Li (1987), advocates the use of a Mallows criterion for model averaging that under certain assumptions results in optimal model averaging in terms of minimal squared error of the corresponding model average estimator amongst all model average estimators.[9] The MMA model weights are obtained by solving a quadratic optimization problem. Denote with

$$\hat{U} = [\hat{u}(1), \ldots, \hat{u}(2^{k_{12}})] \in \mathbb{R}^{N \times 2^{k_{12}}}$$

9 Liang et al. (2011) propose to select the weights by minimizing the trace of an unbiased estimator of the model average estimator MSE.

the collection of residual vectors of all models and with

$$M = [\dim(\mathcal{M}_1), \ldots, \dim(\mathcal{M}_{2^{k_{12}}})]' \in \mathbb{R}^{2^{k_{12}}}$$

the dimensions of all models. The dimension of $\mathcal{M}_j$ is given by $k_{11} + r$ plus the number of variables of X_{12} included in $\mathcal{M}_j$. Further, denote with $\hat{\sigma}_F^2$ the estimated residual variance from the full model including all variables of X_{12}. Then, the MMA weight vector is obtained by solving the following quadratic optimization problem, where

$$w = [w(1), \ldots, w(2^{k_{12}})]' \in \mathbb{R}^{2^{k_{12}}}$$

is the vector of weights corresponding to all models.

$$\min_w \left\{ w' \hat{U}' \hat{U} w + 2\hat{\sigma}_F^2 w' M \right\} \text{ subject to: } w \geq 0, \ \sum_{j=1}^{2^{k_{12}}} w(j) = 1. \tag{8}$$

Each of the variables in X_{12} is included in exactly half of the models considered. The model average coefficient corresponding to each of the variables $X_{12,i}$, $i = 1, \ldots, k_{12}$ can be written as

$$\begin{aligned} \hat{\beta}_{12,i}^w &= \sum_{j=1}^{2^{k_{12}}} w(j) \hat{\beta}_{12,i}(j) \\ &= \sum_{j: X_{12,i} \notin M_j} w(j) 0 + \sum_{j: X_{12,i} \in M_j} w(j) \hat{\beta}_{12,i}(j). \end{aligned} \tag{9}$$

The above equation (9) shows the shrinkage character of model averaging. This is most clearly seen for equal weighting, for which the inclusion weight of variable i, i.e.

$$\sum_{j: X_{12,i} \in M_j} w(j),$$

is exactly $1/2$ for all variables $X_{12,i}$. Hence for equal weighting the average coefficient is given by $1/2^{k_{12}}$ times the sum of all coefficient estimates over only $2^{k_{12}-1}$ (i.e. half of the) models. More generally, for any given weighting scheme the inclusion weight of variable i indicates the importance of this particular variable, in the 'metric' of the chosen weighting scheme. Thus, the inclusion weight is in a certain sense the frequentist alternative to Bayesian posterior inclusion probabilities. If the inclusion weight of a certain variable is high (e.g. higher than 0.5), this means that the 50% of the models in which this variable is included have a high explanatory power or good performance with respect to e.g. AIC or BIC. Thus, the assessment concerning the importance of variables can be based on their inclusion weights. Inclusion weights can also be computed for sets of variables together, which then allows to assess the *joint* explanatory power of a certain group of variables. One can then e.g. also compare the joint inclusion weight of two variables with the individual inclusion weights of the two variables when considered separately to assess the joint importance of two variables, compare also LEY and STEEL (2007) or DOPPELHOFER and WEEKS (2009) who study the joint effects of growth determinants in a Bayesian framework.[10]

If one does not want to resort to the inclusion weights, or wants to have additional tools at hand, the variables' importance can, of course, be assessed also via significance testing. Proper frequentist inference concerning model average coefficients has to take into account that model average estimators are (random) mixtures of correlated estimators. Frequentist (or classical) inference taking these aspects into account has been developed in HJORT and CLAESKENS (2003) and is discussed in detail in CLAESKENS and HJORT (2008, Section 7.5). We use their two-stage simulation approach for the computation of *conservative* confidence intervals based on approximating the limiting distributions of model average estimators, compare Sections 7.5.3 and 7.5.4 of CLAESKENS and HJORT (2008). A description is available in supplementary material.

2.3 A Comparison with WALS

Let us now compare our approach with that of MAGNUS, POWELL, and PRÜFER (2010), which they label weighted average least squares (WALS). Using our notation and setup, the WALS approach can be described as follows.[11] Considering again the sub-model based regression as in (6)

$$\begin{aligned} y &= X_{11}\beta_{11}(j) + X_{12}(j)\beta_{12}(j) + \tilde{X}_2\beta_2(j) + \tilde{u}(j) \\ &= \underbrace{[X_{11}\ \tilde{X}_2]}_{Z} b_1(j) + X_{12}(j)\beta_{12}(j) + \tilde{u}(j), \end{aligned} \tag{10}$$

10 One can also compute the distribution of inclusion weights over model sizes to see *how many* variables are necessary to explain growth well.

11 Clearly, in the WALS approach no principal components are included, but we keep them here for comparability.

with $b_1(j)=[\beta_{11}(j)',\beta_2(j)']'$, and where the variables $X_{12}(j)$, over which model averaging is performed, are orthogonalized with respect to the variables always included, i.e. the following regressions are considered

$$y = Zb_1(j) + X_{12}(j)^{\perp}\beta_{12}^{\perp}(j) + \tilde{u}(j) \tag{11}$$

where $X_{12}^{\perp}$ is X_{12} ortho-normalized with respect to Z such that

$$X_{12}^{\perp\prime}\left(I_n - Z(Z'Z)^{-1}Z'\right)X_{12}^{\perp} = I_{k_{12}} \tag{12}$$

The OLS estimator of β in the full model – with X_{12} transformed as described – is referred to as $\hat{\beta} = [\hat{b}_1', \hat{\beta}_{12}^{\perp\prime}]'$ and the OLS estimator of b_1 from the model including only Z is denoted by $\hat{b}_{1,r} = (Z'Z)^{-1}Z'y$.

Due to the specific orthogonalization chosen it holds that

$$\hat{b}_1(j) = \hat{b}_{1,r} - QW(j)\hat{\beta}_{12}^{\perp} \tag{13}$$

$$\hat{\beta}_{12}^{\perp}(j) = W(j)\hat{\beta}_{12}^{\perp}, \tag{14}$$

with $Q = Z(Z'Z)^{-1}Z'X_{12}^{\perp}$ and $W(j)$ a $k_{12} \times k_{12}$ diagonal matrix with ones and zeros on the diagonal such that the i-th diagonal element is 0 if and only if the i-th component of $\beta_{12}^{\perp}$ is restricted to be 0, i.e. if the corresponding variable is excluded.

Considering model averaging at the moment only for b_1 it holds, by construction, that

$$\begin{aligned}\hat{b}_1^{WALS} &= \sum_{j=1}^{2^{k_{12}}} w(j)\hat{b}_1(j) \\ &= \hat{b}_{1,r} - Q\sum_{j=1}^{2^{k_{12}}} w(j)W(j)\hat{\beta}_{12}^{\perp} \\ &= \hat{b}_{1,r} - QW\hat{\beta}_{12}^{\perp} \\ &= \hat{b}_{1,r} - Q\hat{\beta}_{12}^{\perp,WALS},\end{aligned} \tag{15}$$

since of course

$$\hat{\beta}_{12}^{\perp,WALS} = \sum_{j=1}^{2^{k_{12}}} w(j)W(j)\hat{\beta}_{12}^{\perp} = W\hat{\beta}_{12}^{\perp}.$$

Next, note that due to the orthogonalization, the different components of $\hat{\beta}_{12}^{\perp}$ are (under the normality assumption conditionally) independent of each other.

Furthermore, it follows from the orthogonalization that W is diagonal. Therefore, calculation of $\hat{\beta}_{12}^{\perp,WALS}$ separates into k_{12} 1-dimensional problems.[12] The specific weights chosen depend upon the prior assumption. Magnus, Powell, and Prüfer (2010) choose the Laplace prior and estimator, which leads to a particularly simple calculation of the WALS estimator.

The description shows the similarities and differences to our approach: Both approaches are geared towards complexity reduction with respect to model averaging, where the WALS approach is more efficient in reducing the complexity due to the orthogonality of the variables that are in- respectively excluded. This clearly, however, limits the applicability to 'full rank' regression problems and thus in particular the number of variables has to be smaller than (or equal to) the number of observations, which is not the case for our approach. The distinction in focus and auxiliary variables has a slightly different interpretation. In our case, the focus variables are all variables that are not input in the calculation of principal components. Model averaging is then performed over (a subset of) the focus variables. Magnus, Powell, and Prüfer (2010) consider model averaging over their auxiliary, orthogonalized variables only. In a sense their approach is thus conceptually comparable to a setup in which we would model average (also) over the principal components rather than a subset of the focus variables. Model averaging also over principal components is considered in Wagner and Zeugner (2012).

3. Empirical Results

In the empirical analysis we use three data sets, with two of them well-known and widely studied. These are the data sets of Sala-i-Martin, Doppelhofer, and Miller (2004) and Fernandez, Ley, and Steel (2001). Sala-i-Martin, Doppelhofer, and Miller (2004) consider data for 88 countries and 67 explanatory variables with the dependent variable being the average growth rate of per capita GDP over the period 1960–1996.[13] The data set used in Fernandez, Ley, and Steel (2001) is based on the data set used in Sala-i-Martin (1997b). In

12 The diagonality of W implies that not all $2^{k_{12}}$ weights $w(j)$ have to be computed, but only k_{12} linear combinations thereof.

13 A detailed description of this data set is given in Sala-i-Martin, Doppelhofer, and Miller (2004) in Tables 1 (variable list) and A.1 (country list).

particular a subset of the Sala-i-Martin data containing the 25 variables singled out as important by Sala-i-Martin (1997a) are used. These variables are available for 72 countries. Fernandez, Ley, and Steel (2001) add 16 further variables that are also available for these 72 countries, which gives a total 41 explanatory variables. The dependent variable is the average annual growth rate of real per capita GDP over the period 1960–1992. The third data set is an (in terms of variables) extended and updated version of the Sala-i-Martin, Doppelhofer, and Miller (2004) data set, see the description in Appendix A. The data set contains 86 countries (compared to the 88 countries used in Sala-i-Martin, Doppelhofer, and Miller, 2004, Liberia and Taiwan are missing) and 85 explanatory variables.[14] When possible the variables have been updated (for details see the table with the variables description in Appendix A), and the dependent variable is the average growth rate of per capita GDP over the period 1960–2004. Compared to the original data set new variables that are included are in particular related to information technology and climate (respectively climate related disasters). For the extended and updated data set it is clear that with an equal number of countries and variables regression based methods cannot be meaningfully directly applied. For the extended data set we partition the explanatory variables in 12 groups (see Table 8 in Appendix A) and also study the importance of the variables within these groups. The final analysis is then the assessment of the relative importance of the 16 variables that are significant from the 12 groups jointly. For the Sala-i-Martin, Doppelhofer, and Miller (2004) and Fernandez, Ley, and Steel (2001) data sets we only study the effect of individual variables on GDP growth as this allows most directly for a comparison to the underlying studies in which the variables' effects on growth are also considered individually. However, when considering the conditional effect of individual variables we consider both *growth* and *convergence* equations. The latter include in addition to the explanatory variable under study also initial GDP as an explanatory variable, and of course the principal components. Note that in the former, however, in our approach the effect of initial GDP is partly included since initial GDP is in that case throughout included in the sets of control variables from which the principal components are computed.

Clearly, endogeneity may be a problem that plagues growth and convergence analysis (Durlauf, Kourtellos, and Tan, 2008; Koop and Strachan, 2012).

14 Since all regressions considered in this paper also include an intercept, this then leads to the fact that for this data set the total number of explanatory variables is equal to the number of countries.

If instruments are available, PCAR analysis can be applied using IV rather than OLS estimation. This stems from the fact that the principal components are linear combinations of the original variables. In case that only focus variables are endogenous, standard IV estimation in the PCAR is straightforward. In case that (also) some auxiliary variables are endogenous, one can calculate the principal components based on the fitted values of a regression of the auxiliary variables on the instruments (including the exogenous auxiliary variables). We do not pursue this approach here, as we intend to benchmark our method against standard results available in the literature.

3.1 The Sala-i-Martin, Doppelhofer, and Miller (2004) Data

We start by considering the conditional effects of all the variables individually on GDP growth. The results for the Sala-i-Martin, Doppelhofer, and Miller (2004) data are given in Table 1. The table reads as follows: In the column 'Reg. w/o PC' the results of the regression of GDP growth on the respective variable only are reported, the column labeled 'PCAR w/o GDP0' shows the results of the growth regressions of GDP growth on the indicated variable when in addition principal components (calculated from the rest of the variables) are included in the regression. The column 'PCAR GDP0' shows the results of the convergence regressions of GDP growth on inital GDP, the indicated variable and the principal components. The column 'SDM04' displays the unconditional posterior means of the coefficient estimates computed from Sala-i-Martin, Doppelhofer, and Miller (2004, Table 3, pp. 828–829) and Sala-i-Martin, Doppelhofer, and Miller (2004, Table 4, p. 830) for mean prior model size 7. The column 'PIP' displays the posterior inclusion probabilities from Sala-i-Martin, Doppelhofer, and Miller (2004, Table 3, pp. 828–829) and the corresponding ranks are displayed in the column labeled 'Rank'. Throughout the tables in the paper numbers in bold indicate significance at the 5% level and numbers in italic indicate significance at the 10% level.

As expected, the results from the first column differ most strongly from the results in the other columns. These simple regressions, of course, suffer heavily from omitted variables biases and we include them to illustrate the effect that not controlling for other variables has on the results. These effects are twofold: First, many variables (altogether 42 out of 67) appear to be significant and second, as also expected, the coefficient estimates differ from the PCAR results. The results of the growth and convergence equations augmented by principal components are very similar for this data set, both in terms of significance of variables as well as with respect to the coefficient estimates.

We start by discussing the convergence equation results (from column 'PCAR GDP0'). In addition to initial GDP (GDP0, negative, 4)[15] 8 variables are found to be significant at the 5% level and 6 at the 10% level using White (1980) heteroskedasticity robust standard errors. The variables, other than initial GDP, that are significant at the 5% level are in alphabetical order of the variable name: the fraction of Buddhists in the population in 1960 (BUDDHA, positive, 16); the fraction of Confucian in the population in 1960 (CONFUC, positive, 9); the population density in 1960 (DENS, positive, 19); the dummy for East Asia (EAST, positive, 1); the logarithm of hydro carbon deposits in 1993 (LHCPC, positive, 43); life expectancy in 1960 (LIFE, positive, 8); primary schooling enrolment in 1960 (P, positive, 2); and real exchange rate distortions (RERD, negative, 21). Clearly, the significance of the three variables BUDDHA, CONFUC and EAST is related to the same thing, namely the exceptional growth performance of the South East Asian economies (over the sample period).

The variables significant at the 10% level are: the average share of public investment as fraction of GDP over 1960–1965 (GGCFD, negative, 28); the average share of nominal government spending to nominal GDP between 1960–1990 (GOVNOM1, negative, 33); the average investment price level between 1960–1964 on PPP basis (IPRICE, negative, 3); the fraction of GDP in mining (MINING, positive, 12); the number of revolutions and coups (REVCOUP, negative, 41); and the number of years open between 1950–1994 (YRSOPEN, positive, 14). The negative coefficient signs for GGCFD and GOVNOM1 are in line with Wagner's law, according to which public expenditure is increasing with increasing income. Thus, negative coefficients to these variables reflect the typically observed lower growth in richer countries (with larger public shares).

The results are quite similar for the growth and the convergence equations. In the growth equations 15 variables are significant at least at the 10% level and also, with few exceptions, the same variables as found for the convergence equations are found to be significant. Thus, including initial GDP as either a separate regressor or including it in the set of variables from which the principal components are computed does not imply huge differences in terms of results (which is not the case for all data sets considered).

How do these results compare to those of Sala-i-Martin, Doppelhofer, and Miller (2004)? First, the conditional convergence speed found with our approach is much higher than that found by Sala-i-Martin, Doppelhofer, and

15 Positive respectively negative in the brackets indicates whether the estimated coefficients are positive or negative and the number gives the posterior inclusion probability rank of the variable as computed in Sala-i-Martin, Doppelhofer, and Miller (2004).

Miller (2004). Second, given that we find 15 variables significant at the 10% level in the convergence equations, let us compare the set of significant variables with the top 15 ranked variables, according to posterior inclusion probability (PIP), of Sala-i-Martin, Doppelhofer, and Miller (2004). Only about half of the top PIP ranked variables is found to be significant and in particular we find the following of their top ranked variables to be not significant: the population density in coastal areas in the 1960's (DENSC, 6); the regional dummies for Latin America (LAAM, 11), sub-Saharan Africa (SAFRICA, 10) and Spanish colony (SPAIN, 13); the fraction of tropical area (TROPICAR, 5); the fraction of Muslims in the population in 1960 (MUSLIM, 15); and an index of Malaria prevalence in 1966 (MALFAL, 7).

In our results, compared to these *non-economic* variables more core economic variables are found to be significant. Therefore, conditioning on all available information when assessing an individual variable's contribution to economic growth does make a substantial difference for the Sala-i-Martin, Doppelhofer, and Miller (2004) data. One reason for this difference may well be that the results of Sala-i-Martin, Doppelhofer, and Miller (2004) are based on models with expected prior model size equal to only 5 or 7, which is a substantially smaller model size than what we consider by including also principal components computed from the other variables in all regressions.

Table 1: Results for the Sala-i-Martin, doppelhofer, and miller (2004) Data

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	SDM04	PIP	Rank
GDP0	**0.006400**	**–0.011095**	**–0.011095**	–0.005849	0.685	4
ABSLATIT	**0.000438**	–0.000070	–0.000100	0.000004	0.033	34
AIRDIST	**–0.000003**	0.000001	0.000001	0.000000	0.039	32
AVELF	**–0.026867**	–0.005313	–0.003442	–0.001185	0.105	17
BRIT	0.002770	0.001727	0.000448	0.000099	0.027	42
BUDDHA	**0.050549**	**0.019939**	**0.018931**	0.002340	0.108	16
CATH	0.002411	0.001972	–0.001572	–0.000278	0.033	35
CIV	**0.012316**	–0.008508	–0.006847	–0.000209	0.029	39
COLONY	**–0.015655**	–0.002018	–0.000828	–0.000145	0.029	40
CONFUC	**0.113650**	**0.032199**	**0.030335**	0.011212	0.206	9
DENS	–0.000011	*0.000017*	**0.000019**	0.000001	0.086	19
DENSC	**0.000016**	–0.000001	9.8×10^{-8}	0.000004	0.428	6
DENSI	*0.000026*	–0.000009	–0.000016	0.000000	0.015	64
DPOP	**–0.730844**	0.179474	0.281626	0.000396	0.019	55
EAST	**0.031652**	**0.015169**	**0.013641**	0.017946	0.823	1
ECORG	**0.003412**	0.000016	0.000083	–0.000003	0.015	65
ENGFRAC	0.003453	–0.007457	–0.003862	–0.000073	0.02	50
EUROPE	**0.011110**	0.005989	0.000595	–0.000068	0.03	37
FERT	**–0.019411**	–0.007400	–0.013488	–0.000233	0.031	36

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	SDM04	PIP	Rank
GDE	*0.214804*	–0.000992	0.004636	0.000952	0.021	45
GEEREC	**0.476699**	**0.322562**	0.256021	0.002720	0.021	46
GGCFD	0.023626	*–0.090675*	*–0.080975*	–0.002954	0.048	28
GOVNOM1	–0.023395	*–0.065118*	*–0.056796*	–0.001211	0.036	33
GOVSH61	**–0.090321**	–0.008593	–0.012222	–0.002197	0.063	24
GVR61	**–0.113312**	–0.019279	–0.017019	–0.004594	0.104	18
H	**0.111234**	–0.036607	–0.004912	–0.004251	0.061	25
HERF	*–0.0197022*	0.000146	–0.000436	–0.000095	0.02	51
HINDU	0.009485	0.005662	0.007608	0.000790	0.045	30
IPRICE	**–0.000156**	*–0.000064*	*–0.000068*	–0.000065	0.774	3
LAAM	–0.005085	–0.008372	–0.009401	–0.001901	0.149	11
LANDAREA	0.000000	0.000000	0.000000	0.000000	0.016	60
LANDLOCK	**–0.011634**	0.000595	–0.000439	–0.000044	0.021	47
LHCPC	0.000285	0.000438	**0.000726**	0.000008	0.025	43
LIFE	**0.000842**	*0.000632*	**0.000796**	0.000169	0.209	8
LT100CR	**0.019717**	–0.007775	–0.005493	–0.000049	0.019	56
MALFAL	**–0.024310**	–0.007681	–0.006690	–0.003957	0.252	7
MINING	–0.012027	0.049197	*0.043134*	0.004814	0.124	12
MUSLIM	0.000897	–0.000276	0.002243	0.001440	0.114	15
NEWSTATE	–0.003111	0.002427	0.002242	0.000022	0.019	57

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	SDM04	PIP	Rank
OIL	–0.001489	0.001472	0.002148	0.000092	0.019	58
OPEN	**0.015025**	0.004423	0.004736	0.000673	0.076	22
ORTH	0.013594	0.004509	–0.002212	0.000085	0.015	66
OTHFRAC	**0.011954**	0.002035	0.003393	0.000001	0.086	20
P	**0.036750**	**0.023802**	**0.024177**	0.021374	0.796	2
PI	–0.000120	–0.000100	–0.000126	–0.000001	0.02	52
SQPI	–0.000001	–0.000002	–0.000002	0.000000	0.018	59
PRIGHTS	**–0.003778**	–0.000503	–0.000859	–0.000122	0.066	23
POP15	**–0.056553**	0.058964	0.027975	0.001843	0.041	31
POP	5.1×10^{-8}	5.6×10^{-8}	2.8×10^{-8}	0.000000	0.021	48
POP65	**0.148299**	0.098562	0.104597	0.000426	0.022	44
PRIEXP	**–0.033106**	0.002324	–0.000395	–0.000601	0.053	27
PROT	0.004790	*–0.013002*	–0.009576	–0.000546	0.046	29
RERD	**–0.000233**	*–0.000084*	**–0.000090**	–0.000006	0.082	21
REVCOUP	**–0.017251**	–0.007036	*–0.008016*	–0.000205	0.029	41
SAFRICA	**–0.023227**	–0.005591	–0.004190	–0.002265	0.154	10
SCOUT	0.001302	–0.002663	–0.002587	–0.000099	0.03	38
SIZE	**0.003427**	–0.002729	–0.001319	–0.000010	0.02	53
SOCIALIST	–0.007213	0.003169	0.007415	0.000080	0.02	54
SPAIN	*–0.005358*	–0.000369	0.002187	–0.001319	0.123	13

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	SDM04	PIP	Rank
TOT1DEC	0.010210	0.019852	0.001728	0.000685	0.021	49
TOTIND	0.006335	0.004392	0.007170	–0.000060	0.016	61
TROPICAR	**–0.016914**	–0.009268	–0.008735	–0.008308	0.563	5
TROPPOP	**–0.028968**	–0.002226	–0.000501	–0.000623	0.058	26
WARTIME	**–0.016530**	–0.008367	–0.011393	–0.000023	0.016	62
WARTORN	–0.004908	–0.003217	–0.003438	–0.000011	0.015	67
YRSOPEN	**0.0341981**	**0.013225**	*0.009378*	0.001453	0.119	14
ZTROPICS	–0.016845	–0.000532	–0.000011	–0.000033	0.016	63

Note: See the text for a description of the table.

3.2 *The Fernandez, Ley, and Steel (2001) Data*

The results for the Fernandez, Ley, and Steel (2001) data are presented in Table 2, where the first three columns are as in Table 1. The column 'FLS01' displays the unconditional posterior means of the coefficient estimates and the columns 'PIP' and 'Rank' the posterior inclusion probabilities and the corresponding ranks. These results have been obtained from Mark Steel and differ slightly from the results in the published paper due to Monte Carlo simulation variation. We find, again considering the convergence equation, ten variables to be significant at the 5% level and three at the 10% level. The variables significant at the 5% level are: initial GDP (GDP0, negative, 1)[16]; the fraction of population Confucian (Confucius, positive, 2); the degree of capitalism (EcoOrg, positive, 10); equipment investment (EqipInv, positive, 4); life expectancy (LifeExp, positive, 3); the fraction of GDP in mining (Mining, positive, 9); the share of population Muslim (Muslim, positive, 6); non-equipment investment (NEquipInv, positive, 12); the fraction of population Protestant (Protestants, negative, 11); and a dummy for sub-Saharan Africa (SubSahara, negative, 5). Significant at the 10% level are: the fraction of population Buddhist (Buddha, positive, 15); the fraction of population Jewish (Jewish, positive, 38); and a dummy for Latin America (LatAmerica, negative, 13).

As for the Sala-i-Martin, Doppelhofer, and Miller (2004) data, the differences between the growth (11 variables significant at least at the 10% level) and convergence equation results are not very big. The variables Jewish, Muslim and NEquipInv that are significant in the convergence equations are not significant in the growth regressions, in which instead the standard deviation of the black market premium (stdBMP, negative, 31) is significant. Note also that in the simple bivariate regressions 25 variables are found to be significant at the 5% level and 3 at the 10% level. This shows again that not conditioning on the other explanatory variables leads to a very different and misleading picture concerning the determinants of long-run economic growth.

Compared to the Sala-i-Martin, Doppelhofer, and Miller (2004) BACE approach and data, for this data set our results are much closer to the BMA findings of Fernandez, Ley, and Steel (2001). The only differences (considering the top 13 ranked variables in terms of PIP for comparison) are that the rule of law (RuleofLaw, 8) and the number of years open (YrsOpen, 7) are not significant, whereas the religion related variables Buddha and Jewish are significant at the

16 Positive respectively negative in the brackets indicates whether the estimated coefficients are positive or negative and the number gives the posterior inclusion probability of the variables.

10% level. Not only are the significant variables found with our approach very similar to those with high posterior inclusion probabilities in Fernandez, Ley, and Steel (2001), also the estimated coefficients are quite close to the posterior means of the coefficient estimates of Fernandez, Ley, and Steel (2001), at a negligible fraction of computational cost. Thus, the application to this data set already forcefully demonstrates the virtues of our simple approach to single out the determinants of economic growth.

Table 2: Results for the FERNANDEZ, LEY, and STEEL (2001) Data

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	FLS01	PIP	Rank
GDP0	**0.004364**	**–0.015781**	**–0.015781**	–0.014130	0.999	1
AbsLat	**0.000356**	0.000025	0.000128	0.001560	0.044	33
Age	–0.000029	–0.000054	–0.000037	–0.015020	0.086	22
Area	–0.000001	–0.000001	0.000000	–0.000880	0.030	40
BlMktPm	–0.009316	–0.003803	–0.004394	–0.041150	0.185	16
BritCol	0.002460	–0.001780	–0.001900	–0.003311	0.040	36
Buddha	**0.044672**	**0.012043**	*0.009555*	0.047970	0.200	15
Catholic	**–0.011212**	0.000971	–0.003024	–0.010160	0.133	17
Confucius	**0.106994**	**0.065072**	**0.046786**	0.490000	0.989	2
CivlLib	**–0.003668**	–0.000801	–0.001041	–0.049120	0.129	18
EcoOrg	**0.005338**	*0.002316*	**0.002151**	0.149200	0.457	10
English	0.001093	–0.001927	–0.001868	–0.010600	0.069	26
EquipInv	**0.336561**	**0.142618**	**0.101684**	0.548300	0.923	4
EthnoLFrac	**–0.017271**	0.004348	0.004424	0.009789	0.057	28
Foreign	0.005576	–0.003101	0.000696	0.011630	0.066	27
FrenchCol	*–0.010966*	–0.005571	–0.002048	0.006639	0.051	30
HighEnroll	*0.072704*	–0.028504	0.005634	–0.007921	0.044	34
Hindu	–0.008578	–0.025181	–0.025377	–0.035160	0.128	19
Jewish	**0.012135**	0.003429	*0.010248*	–0.002343	0.037	38
LabForce	1.9×10^{-8}	2.7×10^{-8}	-5.7×10^{-9}	0.019020	0.080	24

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0	FLS01	PIP	Rank
LatAmerica	**–0.012609**	**–0.010774**	*–0.007387*	–0.080840	0.214	13
LifeExp	**0.000783**	**0.000502**	**0.000773**	0.951300	0.930	3
Mining	0.003311	*0.042383*	**0.029581**	0.146000	0.459	9
Muslim	–0.002330	0.002337	**0.009979**	0.257600	0.638	6
NEquipInv	**0.147691**	0.045010	**0.061612**	0.133900	0.429	12
OutwarOr	0.004104	*–0.004873*	–0.002851	–0.003731	0.041	35
PolRights	**–0.003520**	–0.001171	–0.001392	–0.028630	0.097	21
Popg	**–0.633515**	–0.009578	–0.073212	0.005344	0.039	37
Protestants	–0.008623	**–0.022914**	**–0.017806**	–0.142700	0.455	11
PrScEnroll	**0.0333256**	0.000808	0.012432	0.107100	0.213	14
PrExports	**–0.031102**	–0.007146	–0.008447	–0.029360	0.100	20
PublEdu	*0.403301*	0.071368	0.039498	0.000709	0.029	41
RevCoup	**–0.0211534**	0.003320	0.002584	–0.000238	0.032	39
RuleofLaw	**0.027364**	0.007073	0.007229	0.242400	0.490	8
RFEXDist	**–0.000161**	–0.000008	–0.000009	–0.017060	0.084	23
SpanishCol	**–0.011573**	–0.005269	0.000674	0.008980	0.055	29
SubSahara	**–0.019390**	–0.010804	**–0.009984**	–0.465900	0.730	5
stdBMP	**–0.000066**	**–0.000020**	–0.000003	–0.005951	0.048	31
WarDummy	**–0.010645**	–0.001792	–0.003217	–0.014500	0.078	25
WorkPop	–0.012874	–0.009276	–0.009770	–0.005795	0.045	32
YrsOpen	**0.030957**	0.003074	0.005011	0.257500	0.512	7

Note: See the text for a description of the table.

3.3 The Extended and Updated Data Set

We now turn to a more detailed analysis of the extended and updated data set. Here we consider, as mentioned before, three different set-ups. As for the other two data sets we consider the individual variables' conditional effects in both growth and convergence equations and contrast them with the findings from the simple regressions. Second, we partition the explanatory variables into 12 *thematic* groups and we consider the conditional effects of the variables in these groups by using principal components augmentation combined with frequentist model averaging. This experiment is intended to shed light on which variables are important within the thematic groups of variables (if any). The 'final' analysis then collects all 16 variables that are significant in the analysis of the 12 groups and considers these variables' joint (conditional) effects on GDP growth.

3.3.1 The Individual Variables' Conditional Effects

For the extended data set, see the results in Table 3, 7 variables are significant at the 5% level and 8 at the 10% level in the convergence equations. The variables significant at the 5% level are the labor force participation rate of males over 1980–1990 (ACTIVM, negative)[17], the fraction of Confucian in the population in 1960 (CONFUC, positive), initial GDP (GDP0, negative), the average share of real government spending to real GDP between 1960–1990 (GOVSH61, negative), the average investment price level between 1960–1970 (IPRICE, negative), the share of mining in GDP (MINING, positive) and mobile phone subscribers per person over 1960–2000 (TELCELL, positive). The variables significant at the 10% level are: the population growth rate over 1960–1990 (DPOP, negative), the fraction of Orthodox in the population 1960 (ORTH, negative), the fraction of the population over 65 in 1960 (POP65, positive), the political rights index (PRIGHTS, negative), real exchange rate distortions (RERD, negative), the dummy for outward orientation (SCOUT, negative), telecommunication revenues as a fraction of GDP over 1960–2000 (TELREVEN, negative), the fraction of the population living in tropical area (TROPPOP, negative).

In contrast to the other two data sets the results now differ markedly between the convergence and the growth equations. Only 3 of the variables that are significant in the convergence equations are significant also in the growth equations (GDP0, CONFUC and MINING). Thus, with 13 variables signifcant in

17 Positive respectively negative in the brackets again indicates whether the estimated coefficients are positive or negative.

the growth equations there are 10 different variables significant in the growth equations compared to the convergence equations. The differing significant variables at the 5% level are: the absolute latitude (ABSLATIT, positive), the average losses caused by climatic disasters as a percentage of GDP between 1960–1990 (CLIMDISTR, negative), the population growth rate over 1960–1990 (DPOP, negative), a dummy for Europe (EUROPE, positive), GDP per capita relative to the US (GAP, negative), the rural population as fraction of total population (POPRURAL, positive) and the fraction of a country's population living in the tropics (TROPPOP, negative). Thus, for this data set the 'robustness' of results, with respect to growth or convergence equations, observed for the other 2 data sets is not prevalent. Not controlling for additional variables leads in this data set to 53 variables significant at the 5% level and 6 more significant at the 10% level in the simple regressions. This again forcefully demonstrates that the omitted variables biases occurring in regressions in which important variables are missing – as is almost certainly by construction the case in these simple regressions – generally leads to severely distorted assessments concerning the importance of variables for long-run economic growth.

It is also interesting to note that the results are quite different from those obtained for the original Sala-i-Martin, Doppelhofer, and Miller (2004) data, which this data set is an extension and update of. Only 5 of the variables significant in the convergence equations for the original data are still significant in the convergence equations for the updated and extended data set. These are initial GDP (GDP0), the fraction of Confucian (CONFUC), the investment price (IPRICE), the fraction of mining in GDP (MINING) and real exchange rate distortions (RERD). Thus, the results are very sensitive with respect to changes in the data, a finding also obtained in Ciccone and Jarocinski (2010), who report a large sensitivity of growth determinants to updates in the data. A major difference between the data sets is that in the updated data set the growth rates are computed over the period 1960–2004 compared to the period 1960–1996 in the original data set. Thus, the Asian crisis is now included in the sample period and thus over the longer period the average growth rates of the South East Asian economies are not as impressive as when computed only over the period 1960–1996. Nevertheless CONFUC is still significant, whereas the dummy for South East Asia (EAST) is not anymore.

Table 3: Results for the Extended and Updated Data Set

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0
GDP0	**0.002946**	**-0.018072**	**-0.018072**
ABSLATIT	**0.000341**	**0.000421**	0.000260
ACTIVF	*-0.032920*	-0.006782	-0.007401
ACTIVM	**-0.118958**	-0.033812	**-0.073717**
AIRDIST	**-0.000002**	-3.7×10^{-7}	0.000000
AVELF	**-0.018378**	-0.004524	0.002884
BRIT	0.005780	*0.008345*	0.003576
BUDDHA	**0.038795**	0.011740	0.003036
CATH	0.001774	0.005225	0.004275
CIV	**0.012204**	0.002134	0.000778
CLIMDSTR	**-0.008766**	**-0.006602**	-0.004737
COLONY	**-0.012738**	-0.003816	-0.000441
CONFUC	**0.081867**	**0.052880**	**0.037371**
DENS	-0.000005	0.000014	0.000011
DENSC	**0.000010**	-9.9×10^{-6}	-0.000005
DENSI	**0.000031**	0.000001	-0.000002
DPOP	**-0.646445**	**-0.814880**	*-0.448580*
EAST	**0.023083**	0.006497	0.003976
ENGFRAC	0.004618	-0.006957	0.000101
EUROPE	**0.009753**	**0.018175**	0.011725
ECORG	**0.002590**	-0.000496	0.000734
FDININ	0.077991	0.137978	0.157212
FERT	**-0.014395**	0.005632	0.013909
GAP	**0.000187**	**-0.000560**	0.000099
GDE	0.077991	-0.161289	-0.064709
GFCF	**0.125346**	0.034063	0.017283
GGCFD	0.012386	-0.035751	-0.014333
GEEREC	**0.464655**	0.285134	0.132441
GEOLDSTR	-0.004993	-0.006328	-0.004537
GOVNOM1	0.001256	0.004155	-0.007057

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0
GOVSH61	-0.025308	-0.029997	**-0.029827**
GVR61	**-0.061832**	-0.010507	-0.005228
H	**0.080524**	-0.021960	-0.009840
HERF	*-0.016323*	-0.005046	-0.005285
HINDU	0.013231	-0.011788	-0.017382
HOSPBED	**1.779390**	-1.163317	-0.913626
INTERNET	**0.136928**	0.032600	0.042900
IPRICE	**-0.000053**	-0.000023	**-0.000023**
ITM	**0.034287**	0.005811	0.010875
ITX	**0.027878**	-0.007367	-0.004261
LAAM	*-0.005167*	0.000887	0.002414
LANDAREA	0.000000	0.000000	0.000000
LANDLOCK	-0.007467	0.002159	0.001598
LHCPC	0.000193	0.000009	0.000421
LIFE	**0.000609**	0.000312	0.000424
LT100CR	**0.016414**	0.001248	0.003786
MALFAL	**-0.015965**	-0.002725	0.001684
MINING	0.007966	**0.082149**	**0.062771**
MORTF	**-0.057917**	-0.017428	0.027286
MORTM	**-0.057561**	-0.011229	0.016193
MUSLIM	-0.001359	0.000932	0.001595
NEWSTATE	*-0.003043*	-0.001238	-0.001648
OIL	-0.007671	-0.007831	-0.003253
OPEN	**0.011085**	-0.010321	-0.002577
ORTH	**0.013217**	-0.003720	*-0.009953*
OTHFRAC	0.004081	-0.001363	0.000659
P	**0.022589**	-0.006016	0.004695
PI	*-0.000157*	0.000164	-0.000025
POP	4.7×10^{-8}	3.9×10^{-8}	8.5×10^{-9}
POP15	**-0.0456781**	0.053682	0.046885
POP1564	**0.127041**	0.022832	-0.052233
POP65	**0.126666**	0.097261	*0.151341*

Variable	Reg. w/o PC	PCAR w/o GDP0	PCAR GDP0
POPRURAL	**-0.018441**	**0.026843**	0.010023
PRIEXP	**-0.023358**	0.002068	-0.003719
PRIGHTS	**-0.003546**	-0.002199	*-0.001845*
PROT	0.004709	*-0.013868*	-0.009146
RERD	**-0.000171**	-0.000042	*-0.000067*
REVCOUP	**-0.013995**	0.001116	-0.004948
SAFRICA	**-0.016775**	-0.012147	-0.006795
SCOUT	-0.000851	-0.005144	*-0.004130*
SIZE	**0.002426**	-0.000718	0.001095
SOCIALIST	**-0.008357**	-0.001058	-0.001625
SPAIN	**-0.006303**	-0.000096	0.002486
SQPI	*-0.000002*	0.000001	1.9×10^{-7}
TELCELL	**0.175188**	0.053793	**0.070519**
TELFIX	**0.036958**	-0.022918	-0.018522
TELREVEN	0.008365	-0.011827	*-0.0167470*
TOT1DEC	-0.035726	-0.009931	-0.057109
TOTIND	0.006804	-0.000393	-0.000061
TROPICAR	**-0.012020**	-0.008620	-0.004056
TROPPOP	**-0.022984**	**-0.018384**	*-0.008977*
WARTIME	**-0.013377**	0.000602	0.006174
WARTORN	-0.004799	-0.003814	-0.001201
YRSOPEN	**0.024889**	*0.011181*	0.003953
ZTROPICS	-0.008683	0.002856	0.003289

Note: See the text for a description of the table.

3.3.2 The Conditional Effects of Groups of Variables

The outlined approach of combining PCAR with model averaging is now applied to identify the significant variables within groups of variables. In order to do so the variables in the extended data set listed in Table 8 in Appendix A are partitioned into 12 *thematic* groups. These groups are: domestic macroeconomic variables (with 9 variables), external macroeconomic variables (6), information technology variables (6), structural economic variables (9), population age variables

(8), population geography variables (6), education variables (5), health variables (5), historical and political variables (10), religion variables (8), climate variables (4) and location variables (9). In the notation of Section 2 all variables from these groups are considered as focus variables that are in- respectively excluded when averaging over all possible models. Principal components are computed from all the variables from all other groups and are included in all regressions.

The full set of results for all 12 groups is available in Appendix B, downloadable from the website of the Journal (www.sjes.ch). Here we only illustrate the results in detail for the domestic macroeconomic variables and very briefly mention the findings for the other groups. In Table 4 the estimated coefficients are reported and in Table 5 the inclusion weights for the 3 used data dependent weighting schemes are reported. The rows of Table 4 have to be read as follows. The row labeled 'w/o PC' reports the results of the OLS regression including all 9 variables but no principal components (to assess again the effect of not controlling for the other available variables), the column labeled 'Full' corresponds to the OLS regression including all 9 variables and the principal components. The remaining 4 rows show the model averaging results using the 4 described model averaging schemes. The number of principal components is 15 computed from quantitative variables and 7 computed from dummy variables.

As expected and as before the inclusion or exclusion of the principal components – i.e. the controlling or not controlling for the information contained in the other variables – does affect the findings. Two of the variables that are significant when no principal components are included, gross fixed capital formation (GFCF) and the size of the economy in 1960 (SIZE), are not significant anymore once the principal components are included.[18] The results are quite similar between the full regression on the one hand and the results for the different model averaging schemes on the other. No differences occur with respect to significance of coefficients and also the coefficient estimates are quite similar, with the exception of the results for the equal weighting scheme, which leads to coefficient values relatively dissimilar to the other 4 estimates based on regressions including principal components. Conditioning on the principal components leads to only initial GDP (GDP0, negative), the average share of real government spending to real GDP between 1960–1990 (GOVSH61, negative) and the average investment price level between 1960–1970 (IPRICE, negative) to be significant (all at the 5% level). In this respect it is interesting to note that these

18 On the other hand, the average share of real government spending to real GDP between 1960–1990 (GOVSH61) is not significant when no principal components are included but becomes significant once the principal components are included.

3 variables have also been significant (with very similar coefficients) in the convergence equations considered in the previous subsection.

Table 4: Coefficient Estimates for the Domestic Macroeconomic Variables for the Extended Data Set

	GDP0	GFCF	GOVNOM1
w/o PC	**–0.004602**	**0.120515**	–0.004154
Full	**–0.017706**	0.023281	0.001619
Equal	**–0.009193**	0.014238	0.003320
S-AIC	**–0.018293**	0.005305	0.000338
S-BIC	**–0.018366**	0.001771	0.000146
MMA	**–0.017872**	0.001214	0.000000
	GOVSH61	**GVR61**	**IPRICE**
w/o PC	–0.014091	–0.029872	**–0.000036**
Full	**–0.032945**	–0.002905	**–0.000028**
Equal	**–0.015591**	–0.006263	**–0.000014**
S-AIC	**–0.027560**	–0.001024	**–0.000026**
S-BIC	**–0.023000**	–0.000537	**–0.000025**
MMA	**–0.024213**	0.000000	**–0.000023**
	PI	**SIZE**	**SQPI**
w/o PC	0.000081	**0.002662**	–0.000003
Full	0.000131	–0.000572	–0.000001
Equal	0.000020	–0.000169	-3.6×10^{-7}
S-AIC	0.000005	–0.000094	1.6×10^{-8}
S-BIC	0.000001	–0.000020	6.2×10^{-9}
MMA	0.000000	0.000000	0.000000

With respect to the inclusion weights reported in Table 5 two things have to be noted. First, the 3 significant variables have the highest inclusion weights. The inclusion weights for these 3 variables are much larger (the smallest one being 0.782 using S-BIC for GOVSH61) than for all the other variables (the largest one for the other variables being 0.228 using S-AIC for GFCF). Thus, the inclusion weight and significance results perfectly coincide. Second, the inclusion weights do not differ markedly between the three data dependent model averaging schemes used.

Table 5: Inclusion Weights for the Variables that are In- Respectively Excluded in Model Averaging for the Three Data Dependent Model Averaging Schemes for the Domestic Macroeconomic Variables for the Extended and Updated Data Set

	GDP0	GFCF	GOVNOM1	GOVSH61	GVR61
S-AIC	1.000 (1)	0.228 (4)	0.155 (9)	0.922 (3)	0.164 (6)
S-BIC	1.000 (1)	0.079 (4)	0.052 (8)	0.782 (3)	0.060 (5)
MMA	0.972 (1)	0.028 (4)	0.000 (8)	0.827 (3)	0.000 (9)

	IPRICE	PI	SIZE	SQPI
S-AIC	0.988 (2)	0.162 (7)	0.172 (5)	0.156 (8)
S-BIC	0.955 (2)	0.053 (7)	0.059 (6)	0.051 (9)
MMA	0.880 (2)	0.000 (7)	0.000 (5)	0.000 (6)

Qualitatively similar results prevail for the other groups, e.g. with respect to differences occurring whether principal components are included or not and with only minor differences across the data dependent model averaging schemes. The results discussed here are very robust with respect to the weighting scheme used and again there is a high degree of similarity of the results between significance according to *t*-statistics and large inclusion weights. From the external macroeconomic variables only one variable is significant at the 5% level: GDP per capita relative to the US (GAP, negative). The following structural economic variables are significant: the average share of public expenditures on defence as fraction of GDP between 1960–1965 (GDE, negative), the average share of expenditures on public investment as fraction of GDP between 1960–1965 (GGCFD, negative), the fraction of mining in GDP (MINING, positive) and the dummy for outward orientation (SCOUT, negative). From the population age data 4 variables are significant: the labor force participation rate of males over 1980–1990 (ACTIVM, negative), the population growth rate over 1960–1990 (DPOP, negative), fertility in 1960 (FERT, positive), and the fraction of the population younger than 15 in 1960 (POP15, positive). In addition to the group of external macroeconomic variables there are four other groups of variables with one significant variable. For the population geography variables this is the fraction of the population living in tropical area (TROPPOP, negative), for the religion variables it is the fraction of population Confucian (CONFUC, positive), for the climate variables it is the average loss of climatic disasters as a percentage of GDP (CLIMDSTR, negative) and for the location variables it is the absolute latitude (ABSLATIT, positive). In the remaining four groups, information technology, education, health and historical and political data, no variable is significant.

Note that as for the domestic macroeconomic variables and as already in the analysis of the variables' individual (conditional) effects throughout, when considering groups of variables jointly, more variables appear to be significant when no principal components are included. Thus, again controlling for the available information via principal components augmentation leads to different conclusions that are less prone to be contaminated by omitted variables biases.

3.3.3 The Conditional Effects of the Significant Variables of the 12 Groups

We now combine the variables significant from the 12 groups to determine their importance when considered jointly. Considering significance at the 10% level there are 16 variables significant in the twelve groups. These 16 variables are separated further in two groups, 8 variables are included in all regressions (in the terminology of Section 2 these comprise the group X_{11}) and the other 8 are in- respectively excluded in model averaging. The variables always included are those that are significant at the 5% level when considering their individual (conditional) effect as calculated in Section 3.3.1: ACTIVM, CONFUC, DPOP, GDP0, GOVSH61, IPRICE, MINING, TROPPOP. The variables averaged over (i.e. the variables in X_{12} in our terminology from Section 2) are ABSLATIT, CLIMDSTR, FERT, GAP, GDE, GGCFD, POP15 and SCOUT.[19]

Not including the principal components leads to 9 significant variables, 8 of them significant at the 5% level and as before the number of significant variables decreases once we consider the effects conditional on the principal components.[20] The variables significant at the five percent level when including the principal components (for at least one weighting scheme) are: the population growth rate

19 Of course it is not a problem to average over 2^{16} rather than 2^8 models. We perform this separation mainly since the quadratic optimization problem to compute the MMA weights is computationally intensive compared to the other computations and to illustrate that partitioning the focus variables in two sub-groups is a way of further reducing the model space.

20 The number of principal components is 15 from the quantitative variables and 6 from the dummy variables.

Table 6: Coefficient Estimates for the Extended Data Set Using the Variables that Are Significant when Considering the Different Groups of Variables

	ACTIVM	CONFUC	DPOP	GDP0
w/o PC	**–0.058937**	**0.033689**	–0.487856	**–0.015061**
Full	*–0.047096*	**0.035666**	**–0.834946**	**–0.015728**
Equal	–0.038024	*0.033470*	**–0.510727**	**–0.015495**
S-AIC	–0.046248	*0.034521*	**–0.784470**	**–0.015005**
S-BIC	–0.04663	0.032428	**–0.711345**	**–0.014461**
MMA	–0.043117	*0.033075*	**–0.564446**	**–0.015202**

	GOVSH61	IPRICE	MINING	TROPPOP
w/o PC	*–0.018081*	**–0.000031**	0.043711	**–0.019215**
Full	–0.012919	*–0.000019*	**0.053421**	*–0.008781*
Equal	–0.019555	*–0.000021*	**0.055692**	–0.009308
S-AIC	–0.017558	*–0.000020*	**0.052939**	–0.010862
S-BIC	–0.019666	–0.000021	**0.052523**	–0.011843
MMA	–0.01866	*–0.000022*	**0.054624**	–0.009998

	ABSLATIT	CLIMDSTR	FERT	GAP
w/o PC	0.000053	**–0.006136**	–0.01645	**0.000316**
Full	0.000115	**–0.008201**	**0.028920**	0.000293
Equal	0.000033	**–0.003868**	**0.012387**	0.000142
S-AIC	0.000015	**–0.008066**	0.029492	0.000164
S-BIC	0.000005	**–0.008051**	0.026146	0.000088
MMA	0.000043	**–0.005875**	**0.015883**	0.000144

	GDE	GGCFD	POP15	SCOUT
w/o PC	–0.00131	–0.021028	**0.117394**	–0.000927
Full	–0.014888	–0.022141	0.025653	–0.003281
Equal	–0.006213	–0.008795	0.021728	–0.001782
S-AIC	–0.004086	**–0.005137**	*0.004975*	–0.002108
S-BIC	–0.001574	**–0.001796**	**0.001833**	–0.001237
MMA	–0.010317	–0.00961	0.01757	–0.001815

(DPOP, negative), initial GDP (GDP0, negative), the share of mining in GDP (MINING, positive), the losses due to climate disasters (CLIMDSTR, negative), fertility (FERT, positive) which surprisingly has a positive coefficient estimate, the average share of expenditure on public investment (GGCFD, negative) and the share of the population younger than fifteen in 1960 (POP15, positive). At the 10% level in addition the share of Confucian (CONFUC, positive) and the relative price of investment goods (IPRICE, negative) are significant.

Again the weighting scheme itself does not strongly influence the coefficient estimates or the coefficients' significance (with the exception of equal weighting). Amongst the 8 variables in- respectively excluded in model averaging, the significant ones (CLIMDSTR, FERT, GGCFD and POP15) have, with the exception of POP15, high inclusion weights. Outward orientation (SCOUT) ranks third and GAP ranks fourth in terms of inclusion weights but the coefficient estimates are not significant. Apart from these exceptions again the inclusion weights and significance results coincide well.

Table 7: Inclusion Weights for the Variables that are In- Respectively Excluded in Model Averaging for the Three Data Dependent Model Averaging Schemes for the Extended Data Set Using the Variables that Are Significant when Considering the Different Groups of Variables

	ABSLATIT	CLIMDSTR	FERT	GAP
S-AIC	0.175 (7)	1.000 (1)	0.974 (2)	0.566 (4)
S-BIC	0.057 (7)	0.997 (1)	0.894 (2)	0.297 (4)
MMA	0.466 (7)	0.752 (1)	0.591 (2)	0.517 (4)

	GDE	GGCFD	POP15	SCOUT
S-AIC	0.161 (8)	0.232 (5)	0.191 (6)	0.610 (3)
S-BIC	0.055 (8)	0.081 (5)	0.067 (6)	0.347 (3)
MMA	0.459 (8)	0.467 (6)	0.475 (5)	0.518 (3)

Note that the results differ from the results obtained when considering the conditional effects of the individual variables separately (compare Table 3). The variables FERT, GGCFD and POP15 are not significant when considered individually. This indicates that it is indeed important to study the joint effects of variables on GDP growth (LEY and STEEL, 2007; DOPPELHOFER and WEEKS, 2009). Our analysis studying thematic groups of variables and now the subset of the significant variables from the 12 groups are of course just a few examples of analyzing variables' joint effects. One could e.g. rather than consider the variables individually consider the effects of all pairs or triples of variables jointly (again conditioning also on the information contained in all other variables via principal components augmentation).

Taking the evidence from this and the previous two subsections together we find the following variables to be (in our sense) 'robustly' related to long-run economic growth: the population growth rate, initial GDP, the share of mining in GDP, the losses due to climate disasters, the relative investment price and the share of Confucian in the population. All these variables have a clear interpretation and all coefficients have the 'correct' signs and 'sensible' magnitudes (e.g. the convergence coefficient for initial GDP). Losses due to climate disasters have been included in neither the Sala-i-Martin, Doppelhofer, and Miller (2004) nor the Fernandez, Ley, and Steel (2001) data and could therefore obviously not be identified in these papers. Looking at the data one can identify a set of poor countries in which the average losses due to natural disasters are above 0.8 percentage points of GDP. These are Bolivia, Honduras, Haiti, Jamaica, Madagascar, Nepal and Zimbabwe.

The findings confirm that our approach is a valuable tool for disentangling the variables relevant for economic growth in a setting designed to control for estimator bias as well as estimator variance in a computationally very cheap framework that can be easily extended in many ways (e.g. Wagner and Zeugner, 2012, extend PCAR to a Bayesian model averaging framework and consider as an extension a model dependent determination of the number of principal components to be included in the regressions).

4. Summary and Conclusions

The empirical analysis of economic growth is complicated by (i) the uncertainty about which variables are relevant and which are not and (ii) the availability of data sets where the number of variables is of the same magnitude as the number of observations. These features render regression analysis difficult. Including only few variables can imply substantial omitted variables bias whereas including many or all variables will reduce bias but increase estimator variance. We propose to tackle these problems by using principal components augmented regressions (PCARs), which has several advantages: First, PCAR analysis leads to the estimation of well-behaved regressions that efficiently include the information contained in all available variables, which implies that only minimal omitted variables bias will be present, whilst keeping estimator variance low due to the mutual orthogonality of the principal components. Thus, well defined estimates that take the theory open endedness of economic growth into account by conditioning on a large information set are obtained.

PCAR analysis is combined with frequentist model averaging to exemplify the usefulness of the approach for model averaging applications. If preferred, PCAR can of course also be combined with Bayesian model averaging (Wagner and Zeugner, 2012). Inference for frequentist model average coefficients is based on the results of Claeskens and Hjort (2008). We also compute the frequentist counterpart to the Bayesian posterior inclusion probability, the inclusion weights (which can be also computed for groups of variables, to study the jointness of growth determinants, or over different model sizes). Model averaging in conjunction with PCAR analysis becomes computationally very cheap since the model space is substantially reduced by the partitioning of the variables in focus and auxiliary variables with the latter being purged in the principal components.

The application to three data sets clearly demonstrates the usefulness and virtues of the proposed approach. For the Sala-i-Martin, Doppelhofer, and Miller (2004) data we find economically meaningful variables to be significantly related to economic growth. For the Fernandez, Ley, and Steel (2001) data we find very similar results as the authors themselves. The computations for both data sets take seconds, which compares very favorably with the computationally intensive approximate approaches pursued in the original contributions. A more detailed analysis of an extended and updated version of the Sala-i-Martin, Doppelhofer, and Miller (2004) data, where we do not only investigate all variables' conditional individual effects on economic growth but also consider thematic groups of variables jointly, also leads to very interesting results. We find the following variables to be significantly related to economic growth: initial GDP, the population growth rate, the share of mining in GDP, the losses due to climate disasters, the relative investment price and the share of Confucian in the population.

The findings in the paper show that the proposed approach is a valuable computationally and conceptually very simple addition to the toolkit of the empirical growth research community that is suited for being applied to many other questions and that is easily extendible.

Appendix A: Description of Extended and Updated Data Set

Table 8: Description of Variables in the Extended and Updated Data Set.

	SDM04	Updated	Description for extended and updated data set
Macroeconomic domestic data			
DGDP	X	X	Average growth rate of GDP per capita over 1960–2004.
GDP0	X	X	Logarithm of GDP per capita in 1960.
GFCF			Gross fixed capital formation (fraction of GDP), 1960–1990.
GOVNOM1	X	X	Average share of nominal government spending to nominal GDP over 1960–1990.
GOVSH61	X	X	Average share of real government spending to real GDP over 1960–1990 (2000 constant prices).
GVR61	X		Share of expenditures on government consumption of GDP in 1961.
IPRICE	X	X	Average investment price level over 1960–1970 on PPP basis.
PI	X		Average inflation over 1960–1990.
SIZE	X		Size of economy, logarithm of aggregate GDP in 1960.
SQPI	X		Square of average inflation over 1960–1990.
Macroeconomic international data			
FDININ			Foreign direct investment, net inflows (fraction of GDP), 1960–1990.
GAP			GDP per capita relative to the US, US=100 (at PPP in international $, current prices), average over 1960–1970.
OPEN	X		Average of ratio of exports plus imports to GDP over 1965–1974.
RERD	X		Real exchange rate distortions.
TOT1DEC	X		Growth of terms of trade in 1960.
TOTIND	X		Terms of trade ranking.
Information technology data			
INTERNET			Internet users (per person), 1990–2000.
ITM			Computer, communications and other services (as fraction of commercial service imports), 1960–1990. The value for HKG was based on the period 1998–2004 and the value for ZAR is set to zero.

	SDM04	Updated	Description for extended and updated data set
ITX			Computer, communications and other services (as fraction of commercial service exports), 1960–1990. The value for HKG was based on the period 1998–2004 and the value for ZAR is set to zero.
TELCELL			Mobile phone subscribers (per person), 1960–2000.
TELFIX			Telephone mainlines (per person), 1975–1990.
TELREVEN			Telecommunications revenue (fraction of GDP), 1960–2000. The values for DOM and ZAR are based on the period 2003–2005.
Structural economic data			
ECORG	X		Index of extent to which economies favor a capitalist form of production.
GDE	X		Average share of public expenditures on defence as fraction of GDP over 1960–1965.
GGCFD	X		Average share of expenditures on public investment as fraction of GDP over 1960–1965.
LHCPC	X		Logarithm of hydrocarbon deposits in 1993.
MINING	X		Fraction of GDP in mining.
OIL	X		Dummy for oil producing country.
PRIEXP	X		Fraction of primary exports in total exports in 1970.
SCOUT	X		Dummy for outward orientation.
YRSOPEN	X		Number of years economy has been open between 1950 and 1994.
Population – age data			
ACTIVF			Labor force participation rate, female (fraction of male population aged 15–64), 1980–1990.
ACTIVM			Labor force participation rate, male (fraction of male population aged 15–64), 1980–1990.
DPOP	X	X	Population growth, 1960–1990.
FERT	X		Fertility in 1960.
POP15	X		Fraction of population younger than 15 in 1960.
POP1564			Population aged 15–64 (fraction of total), 1960–1990.
POP	X		Population in 1960.

	SDM04	Updated	Description for extended and updated data set
Population – geography data			
AVELF	X		Index of ethnoliguistic fractionalization: Probability that 2 randomly chosen people do not speak the same language.
DENS	X		Population per area in 1960.
DENSC	X		Population per costal area in 1965.
DENSI	X		Interior (more than 100km from coastline) population per area in 1965.
POPRURAL			Rural population (fraction of total population).
TROPPOP	X		Fraction of country's population living in geographical tropics.
Education data			
ENGFRAC	X		Fraction of population speaking English.
GEEREC	X		Average share of public expenditures on education as fraction of GDP over 1960–1965.
H	X		Higher education enrollment rate, 1960.
OTHFRAC	X		Fraction of population speaking foreign languages.
P	X		Primary school enrollment rate in 1960.
Health data			
HOSPBED			Hospital beds (per person), 1960–2000.
LIFE	X		Life expectancy in 1960.
MALFAL	X		Index of malaria prevalence in 1966.
MORTF			Mortality rate, adult-female (fraction of female adults), 1960–1990. The values for TUR and ZAR are based on the period 2002–2005.
MORTM			Mortality rate, adult-male (fraction of male adults), 1960–1990. The values for TUR and ZAR are based on the period 2002–2005.
Historical and political data			
BRIT	X		Dummy for British colony after 1776.
CIV	X		Index of civil liberties in 1972.
COLONY	X		Dummy for former colony.
NEWSTATE	X		Timing of national independence measure: 0 if before 1914, 1 if between 1914 and 1945, 2 if between 1945 and 1989 and 3 if after 1989.
PRIGHTS	X		Political rights index.

	SDM04	Updated	Description for extended and updated data set
REVCOUP	X		Number of revolutions and military coups.
SOCIALIST	X		Dummy for countries under socialist rule for considerable time during 1950–1995.
SPAIN	X		Dummy for former Spanish colonies.
WARTIME	X		Fraction of time spent in war between 1960–1990.
WARTORN	X		Indicator for countries that participated in external war between 1960–1990.
Religion data			
BUDDHA	X		Fraction of population Buddhist in 1960.
CATH	X		Fraction of population Catholic in 1960.
CONFUC	X		Fraction of population Confucian in 1960.
HERF	X		Religious intensity: Religion measure.
HINDU	X		Fraction of population Hindu in 1960.
MUSLIM	X		Fraction of population Muslim in 1960.
ORTH	X		Fraction of population Orthodox in 1960.
PROT	X		Fraction of population Protestant in 1960.
Climate data			
CLIMDSTR			Average loss caused by climatic disasters as a percentage of GDP over 1960–1990 (current USD). Climatic disasters: floods, cyclones, hurricanes, ice storms, snow storms, tornados, typhoons, storms, wild fire, droughts, cold wave.
GEOLDSTR			Average loss caused by geological disasters as a percentage of GDP over 1960–1990 (current USD). Geological disasters: volcanic eruptions, natural explosions, avalanches, land slides, earthquakes, wave/surge.
TROPICAR	X		Fraction of country's land area within geographical tropics.
ZTROPICS	X		Fraction of tropical climate zone.
Location data			
ABSLATIT	X		Absolute latitude.
AIRDIST	X		Minimal distance in km from New York, Rotterdam or Tokyo.
EAST	X		Dummy for East Asia.
EUROPE	X		Dummy for Europe.
LAAM	X		Dummy for Latin American countries.

	SDM04	Updated	Description for extended and updated data set
LANDAREA	X		Land area in km^2.
LANDLOCK	X		Dummy for landlocked country.
LT100CR	X		Fraction of land area near navigable water (within 100 km of ocean or ocean navigable river).
SAFRICA	X		Dummy for Sub-Saharan African countries.

Note: The second column of the table (SDM04) indicates whether this or a similar variable was included in the Sala-i-Martin, Doppelhofer, and Miller (2004) data and the third column indicates whether the variable has been updated compared Sala-i-Martin, Doppelhofer, and Miller (2004).

Table 9: Data Sources for Extended and Updated Data set

Source	Variables
Barro (1999)	ABSLATIT, BRIT, BUDDHA, CATH, CIV, COLONY, CONFUC, HERF, HINDU, MUSLIM, OIL, OPEN, ORTH, POP, PRIGHTS, PROT, SPAIN, TOTIND.
Barro and Lee (1993)	DENS, GDE, GEEREC, GGCFD, GVR61, H, LANDAREA, LIFE, P, POP15, POP65, REVCOUP, TOT1DEC, WARTIME, WARTORN.
Barro and Sala-i-Martin (1995)	FERT
Easterly and Levine (1997)	AVELF
EM-DAT: OFDA/CRED	CLIMDSTR, GEOLDSTR
Gallup et al. (2001)	AIRDIST, DENSC, DENSI, LHCPC, LT100CR, MALFAL, NEWSTATE, SOCIALIST, TROPICAR, TROPPOP, ZTROPICS.
Hall and Jones (1999)	ECORG, ENGFRAC, MINING, OTHFRAC.
Levine and Renelt (1992)	PI, RERD, SCOUT, SQPI.
Heston et al. (2006)	GAP, GDP, GOVNOM1, GOVSH61, GR, IPRICE.
Sachs and Warner (1997)	PRIEXP, YRSOPEN.
UNCTAD	The value of FDININ for HKG is based on available values for HKG and SGP.
WDI (2007)	ACTIVF, ACTIVM, DPOP, FDININ, GFCF, HOSPBED, INTERNET, ITM, ITX, MORTF, MORTM, POP1564, POPRURAL, TELCELL, TELFIX, TELREVEN.
www.nationsencyclopedia.com	The value of HOSPBED in 1989 for ZAF.

References

BARRO, R.J. (1999), "Determinants of Democracy", *Journal of Political Economy*, 107.

BARRO, R.J., and J.-W. LEE (1993), "International Comparisons of Educational Attainment", *Journal of Monetary Economics*, 32.

BARRO, R.J., and X. SALA-I-MARTIN (1995), *Economic Growth*, New York: McGraw-Hill.

BROCK, W.A., and S.N. DURLAUF (2001), "Growth Empirics and Reality", *World Bank Economic Review*, 15.

BUCKLAND, S.T., K.P. BURNHAM, and N.H. AUGUSTIN (1997), "Model Selection: An Integral Part of Inference", *Biometrics*, 53.

CICCONE, A., and M. JAROCINSKI (2010), "Determinants of Economic Growth: Will Data Tell?", *American Economic Journal: Macroeconomics*, 2.

CLAESKENS, G., and N.L. HJORT (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

DOPPELHOFER, G., and M. WEEKS (2009), "Jointness of Growth Determinants", *Journal of Applied Econometrics*, 24.

DURLAUF, S.N., P.A. JOHNSON, and J.R.W. TEMPLE (2005), *Growth Econometrics*, vol. 1 of *Handbook of Economic Growth*.

DURLAUF, S.N., A. KOURTELLOS, and C.M. TAN (2008), "Are any Growth Theories Robust?", *Economic Journal*, 118.

EASTERLY, W. and R. LEVINE (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, 112.

FABER, K., and B.R. KOWALSKI (1997), "Modification of Malinowski's F-Test for Abstract Factor Analysis Applied to the Quail Root II Data Set", *Journal of Chemometrics*, 11.

FERNANDEZ, C., E. LEY, and M.F.J. STEEL (2001), "Model Uncertainty in Cross-Country Growth Regressions", *Journal of Applied Econometrics*, 16.

GALLUP, J.L., A. MELLINGER, and J.D. SACHS (2001), "Geography Datasets", Center for International Development at Harvard University (CID), 2001. Data Web site: http://www2.cid.harvard.edu/ciddata/geographydata.htm

HALL, R.E., and C.I. JONES (1999), "Why Do Some Countries Produce so Much More Output per Worker than Others?", *Quarterly Journal of Economics*, 114.

HANSEN, B. (2007), "Least Squares Model Averaging", *Econometrica*, 75.

HENDRY, D.F., and H.-M. KROLZIG (2004), "We Ran One Regression", *Oxford Bulletin of Economics and Statistics*, 66.

Heston, A., R. Summers, and B. Aten (2006), *Penn World Table Version 6.2*, Center for International Comparisons at the University of Pennsylvania (CICUP).

Hjort, N.L., and G. Claeskens (2003), "Frequentist Model Average Estimators", *Journal of the American Statistical Association*, 98.

Hoover, K.D., and S.J. Perez (2004), "Truth and Robustness in Cross-Country Growth Regressions", *Oxford Bulletin of Economics and Statistics*, 66.

Koop, G., and S. Potter (2004), "Forecasting in Dynamic Factor Models using Bayesian Model Averaging", *The Econometrics Journal*, 7.

Koop, G., R. Leon-Gonzalez and R. Strachan (2012), "Bayesian Model Averaging in the Instrumental Variable Regression Model", *Journal of Econometrics*, 171.

Kritchman, S., and B. Nadler (2008), "Determining the Number of Components in a Factor Model from Limited Noisy Data", *Chemometrics and Intelligent Laboratoryc Systems*, 94.

Lawley, D.N., and A.E. Maxwell (1963), *Factor Analysis as a Statistical Method*, London: Butterworths.

Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: Wiley.

Leamer, E.E. (1985), "Sensitivity Analyses Would Help", *American Economic Review*, 75.

Levine, R., and D. Renelt (1992), "Sensitivity Analysis of Cross-Country Growth Regressions", *American Economic Review*, 82.

Ley, E., and M.F.J. Steel (2007), "Jointness in Bayesian Variable Selection with an Application to Growth Regressions", *Journal of Macroeconomics*, 29.

Li, K.-C. (1987), "Asymptotic Optimality of C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set", *Annals of Statistics*, 15.

Liang, H., G. Zou, A.T.K. Wan, and X. Zhang (2011), "Optimal Weight Choice for Frequentist Model Average Estimators", *Journal of the American Statistical Association*, 106.

Magnus, J.R., O. Powell, and P. Prüfer (2010), "A Comparison of two Averaging Techniques with an Application to Growth Empirics", *Journal of Econometrics*, 154.

Malinowski, E.R. (1989), "Statistical F-Tests for Abstract Factor Analysis and Target Testing", *Journal of Chemometrics*, 3.

Raftery, A.E. (1995), "Bayesian Model Selection in Social Research", *Sociological Methodology*, 25.

Sachs, J., and A.M. Warner (1997), "Fundamental Sources of Long-Run Growth", *American Economic Review* (Papers and Proceedings), 77.

SALA-I-MARTIN, X. (1997a), "I Just Ran 2 Million Regressions", *American Economic Review*.

SALA-I-MARTIN, X. (1997b), "I Just Ran 4 Million Regressions", *NBER Working Paper*.

SALA-I-MARTIN, X., G. DOPPELHOFER, and R.I. MILLER (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach", *American Economic Review*, 94.

SCHNEIDER, U., and M. WAGNER (2012), "Catching Growth Determinants with the Adaptive LASSO", *German Economic Review*, 13.

SCHOTT, J.R. (2006), "A High-Dimensional Test for the Equality of the Smallest Eigenvalues of a Covariance Matrix", *Journal of Multivariate Analysis*, 97.

WAGNER, M., and J. HLOUSKOVA (2012), "Growth Regressions, Principal Components and Frequentist Model Averaging", *mimeo*.

WAGNER, M., and S. ZEUGNER (2012), "Principal Components Augmented Model Averaging", *mimeo*.

WHITE, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48.

SUMMARY

In this paper we use principal components augmented regressions (PCARs), partly in conjunction with model averaging, to determine the variables relevant for economic growth. The use of PCARs allows to effectively tackle two major problems that the empirical growth literature faces: (i) the uncertainty about the relevance of variables and (ii) the availability of data sets with the number of variables of the same order as the number of observations. The use of PCARs furthermore implies that the computational cost is, compared to standard approaches used in the literature, negligible. The proposed methodology is applied to three data sets, including the SALA-I-MARTIN, DOPPELHOFER, and MILLER (2004) and FERNANDEZ, LEY, and STEEL (2001) data as well as an extended version of the former. Key economic variables are found to be significantly related to economic growth, which demonstrates the relevance of the proposed methodology for empirical growth research.