UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES

Advanced Microeconometrics

Project 2

High-dimensional Linear Models and Convergence in Economic Growth

Mikala Ann de Wet Bræmhøj (vrz848), Caroline Bergholdt Hansen and Julie Cathrine Krabek Sørensen

5 November 2023

Contents

1 Introduction

A key question in growth theory concerns economic convergence, the idea that poorer countries tend to experience faster growth than richer countries. This paper investigates this so-called convergence hypothesis empirically by use of a high-dimensional (HD) data set. Due to limitations of OLS in such settings, we use a LASSO approach. Though the validity of the results can be questioned, they seem to confirm the convergence hypothesis.

2 Data

The data set included in the analysis is a collection of variables from 102 different countries across the globe. The data stems from the World Bank, as well as data collected by other researchers[1]. Following the work of prominent scholars within the field, we choose to include variables with geographic (cf. e.g. Diamond and Ordunio (1999)) and institutional (cf. e.g. Acemoglu, Johnson, et al. (2005)) character in our investigation.

3 Theory

Barro (1991) suggested examining the convergence hypothesis statistically by regressing a nation's annual GDP[2] growth, g_i, on its initial GDP, y_{i0}, and a set of control variables, $\mathbf{z}_i$:

$$g_i = \beta y_{i0} + \mathbf{z}_i \gamma + u_{it}, \mathbb{E}[u_i | y_{i0}, \mathbf{z}_i] = 0 \tag{1}$$

where u_{it} is the unobservable error term for country $i = 1, ..., n$. The parameter β is thus the partial effect of initial GDP on GDP growth, such that $\beta < 0$ signifies aforementioned economic convergence. Implicit in this way of specifying the model is that β is the object of interest, while the presence of the control variables is to avoid omitted variable bias rather than because we are particularly interested in the coefficient estimates for these variables. The model of interest, (1), is HD in that it involves a small number of countries (n) compared to the number of candidate regressors (p). This feature of the problem renders least squares (LS) estimation unoptimal, as is evident by the formula for the variance of its prediction error:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_i^{LS} - g_i^*\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i'\hat{\Psi}^{LS} - \mathbf{X}_i'\Psi\right)^2\right] = \frac{\sigma^2 p}{n} \tag{2}$$

[1] See the appendix for a full list of sources and variables.

[2] In this paper, GDP is always measured in per capita terms but shortened to just GDP.

where $g_i^* := \mathbb{E}[g_i|\mathbf{X}_i] = \mathbf{X}_i'\beta$ and $\mathbf{X}_i$ and Ψ are combined regressors $(y_{i0}, \mathbf{z}_i)$ with respective coefficients β and γ. Hence, the LS prediction error does not tend to zero as $\frac{p}{n} \nrightarrow 0$. LASSO, on the other hand, remains useful in HD settings as long as the underlying number of relevant regressors, s, is small relative to the number of observations, n. LASSO can be shown to outperform LS when $\frac{p}{s} \to \infty$.

3.1 The LASSO Estimator

LASSO encourages sparsity by introducing a penalty term to the LS minimization problem, thus penalizing the inclusion of additional regressors:

$$\hat{\beta}(\lambda) \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(g_i - \mathbf{X}_i' b \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\} \tag{3}$$

where $\mathbf{X}_i := [y_{i0}, \mathbf{z}_i]'$. The first term is the usual objective function of the LS estimator, while the second term introduces aforementioned penalty. This penalty depends on a penalty term, λ, and the sum of the absolute value of the coefficients. The LASSO estimator thus performs variable selection in the sense that a variable j is selected iff $\hat{\beta}_j(\lambda) \neq 0$. An implicit assumption of LASSO is that of sparsity, meaning that, from all the candidate regressors, p, there is an underlying set of relevant regressors (i.e., regressors with true coefficients different from zero), $s = \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0)$, which is small.

3.1.1 The Post-Double-LASSO Estimator

Post-Double-LASSO (PDL) estimation involves first estimating:

$$y_0 = \mathbf{z}'\psi + v, E[v|\mathbf{z}] = 0 \tag{4}$$

by single LASSO (SL) to obtain $\widehat{\psi}$ and then estimating (1) by SL to obtain $\widehat{\beta}$ and $\widehat{\gamma}$. (4) is known as the first stage and serves the purpose of extracting the variation in the treatment variable that is not explained by the control variables. Together, the assumptions $E[u|\mathbf{X}] = 0$ and $E[v|\mathbf{z}] = 0$ imply that:

$$E[(y_0 - \mathbf{z}'\psi)(g - \beta y_0 - \mathbf{z}'\gamma] = 0 \tag{5}$$

The added structure imposed by PDL implies the moment condition above, which in turn allows us to isolate the parameter of interest as:

$$\beta = \frac{E[(D - \mathbf{Z}'\psi_0)(Y - \mathbf{Z}'\gamma_0]}{E[(D - \mathbf{Z}'\psi_0)D]} \tag{6}$$

For a given sample, this estimate can, by the analogy principle, be computed as:

$$\check{\beta} = \frac{\sum_{i=1}^{n}(D_i - \mathbf{Z}_i'\widehat{\psi}_0)(Y_i - \mathbf{Z}_i'\widehat{\gamma}_0)}{\sum_{i=1}^{n}(D_i - \mathbf{Z}_i'\widehat{\psi}_0)D_i} \tag{7}$$

3.1.2 Penalty-Term Selection

Selecting the appropriate penalty term λ is key to the reliability of LASSO regression. Our focus will be directed towards utilizing the Bickel-Ritov-Tsybakov rule (BRT) and Belloni-Chen-Chernozhukov-Hansen rule (BCCH)[3].

The **BRT rule** relies on the assumption of conditional homoskedasticity, which implies that the residual term ε is independent of the predictor matrix X and further that the variance of ε is known. The penalty term is calculated:

$$\hat{\lambda}^{BRT} := \frac{2c\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\sqrt{\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n}X_{ij}^2} \tag{8}$$

Here, the econometrician must make choices for the parameters α, a significance level, and c, a scaling factor, which must be larger than 1 and is typically set to 1.1.

The **BCCH rule** is calculated similar to the BRT, but does not rely on the assumption of prior knowledge of the variance.

$$\hat{\lambda}^{BCCH} := \frac{2c}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p}\right)\max_{1\leq j\leq p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon_i}^2X_{ij}^2} \tag{9}$$

The epsilon is obtain from the auxillary regression $\hat{\varepsilon}_i = g_i - X_i'\hat{\beta}^{pilot}$, where $\hat{\beta}^{pilot}$ is an initial estimate derived with a pilot penalty term, which is calculated the same way as $\hat{\lambda}^{BCCH}$, but where $\hat{\varepsilon}_i$ is replaced with $g_i - \bar{g}$. It is noteworthy that the assumption of a known variance does not hold in most applications and so the BCCH is often preferred. The BCCH does however often yield a higher penalty term, and is thus more restrictive.

3.1.3 Inference

Inference based on SL estimation can be cumbersome due to the estimator not being analytically expressible, making it difficult to construct confidence intervals and hypothesis testing. We therefore resort to a different version of LASSO, namely PDL. While the asymptotic distribution of SL is generally unknown, it can be shown that, under certain sparsity conditions, PDL converges to a standard normal distribution as $n \to \infty$ and that $\sigma^2 = \frac{E[\varepsilon^2 v^2]}{(E[v^2])^2}$. By the analogy principle, we can estimate the variance as $\check{\sigma}^2 = \frac{n^{-1}\sum_i \hat{\varepsilon}_i^2 \hat{v}_i^2}{(n^{-1}\sum_i \hat{v}_i^2)^2}$. For our purposes, this is the key advantage of PDL compared to SL, as it allows us to

[3]Cross-validation is also a typical method, but where it is a valuable tool for out-of-sample prediction, it is not necessarily well suited for inference purposes. Therefore it is left out of the analysis.

compute asymptotically valid confidence intervals:

$$\widehat{CI}(1-\alpha) = \left[\check{\beta} \pm q_{1-\alpha/2}\frac{\check{\sigma}}{\sqrt{n}}\right] \tag{10}$$

where $\alpha \in (0,1)$ is the significance level and $q_\alpha := \Phi^{-1}(\alpha)$ is the standard normal quantile function. These estimates can be used for hypothesis testing even in settings of HD.

4 Analysis

We estimate eight different models. Table 1 summarises the results. Models (1)-(2) are estimated by OLS with (1) including the treatment variable as the only regressor, while (2) also includes the selected control variables specified in Section 2. Models (3)-(5) are estimated by PDL using the BRT rule for penalty-term selection and three distinct sets of control variables. Model (3) includes only the original selected control variables, whereas (4) also includes technical variables in the shape of interaction terms, and (5) further adds squares. Finally, models (6)-(8) are pairwise identical to (3)-(5) except the BCCH rule are used.

While the simple OLS regression (model (1)) does not yield a statistically significant estimate for the coefficient on the treatment variable, the resulting estimates from the model estimated by OLS including control variables (cf. column (2) in Table 1) seem to confirm the hypothesis that poorer countries tend to grow faster than richer countries, in that the coefficient estimate on y_{i0} is negative andstatistically significant. Similarly, the models estimated by PDL also yield negative as well as statistically significant estimates for the coefficient on log initial GDP. In summary, besides from model (1), which most likely suffers from omitted variable bias, the results all point in the same direction: confirming the economic divergence hypothesis.

5 Discussion and Conclusion

Even though most of the results confirm the economic divergence hypothesis, there are several reasons we might not have complete faith in their statistical validity. First of all, it is of concern that the models estimated by PDL lead to a very small set of regressors with non-zero coefficient estimates. In model (3), 'Africa' and 'Asia' are the only non-zero control variables out of 26, while the only non-zero control variable in model (4) is the interaction term between the absolute latitude and 'Asia', and in model (5) it is only the interaction term between geodesic centroid longitude and 'Oceania'. While the former case might be explained by Africa and Asia experiencing the highest GDP growth, it is not immediately clear why the two interaction terms in models (4) and (5) are the only control variables with a non-zero coefficient. It is not too surprising that PDL results in some variables being 'excluded', since the PDL performs variable selection. However, that so few variables have non-zero estimates could be an indicator that the penalties selected by the BRT and BCCH are 'too large', making it difficult to extract information and conclude upon the results. E.g., none of the models estimated by PDL using the BCCH result in any of the control variables having a non-zero coefficient estimate. Ultimately, although it is an advantage of PDL that it performs variable selection, it is not very useful if it barely selects *any* variables. It could, on the other hand, also be the case that none of the included control variables are relevant, and that this is why none of them are 'selected'. This seems a bit unlikely, however. Another concern is that of the suitability of OLS estimation for HD data. As already mentioned, OLS typically produces imprecise estimates in HD settings due to the variance of its prediction error depending on $\frac{p}{n}$. While adding the 26 control variables to model (2) does lead to substantially larger standard errors, we are nonetheless still able to establish statistical significance of the coefficient estimate of main interest. It is not certain that this would still be the case if we included even more control variables, in which case the problem would be 'truly' HD.

In conclusion, all the obtained estimates for β (except for model (1), whose validity is highly questionable), confirm the hypothesis of economic convergence, but the fact that OLS is typically not well-suited for this type of problem and PDL in this case leads to only very few variables being selected indicate that the results might not be completely trustworthy.

References

Acemoglu, Daron, Simon Johnson, and James A Robinson (2005). "Institutions as a fundamental cause of long-run growth". In: *Handbook of economic growth* 1, pp. 385–472.

Acemoglu, Daron, Suresh Naidu, et al. (2019). "Democracy does cause growth". In: *Journal of political economy* 127.1, pp. 47–100.

Ashraf, Quamrul and Oded Galor (Feb. 2013). "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development". In: *American Economic Review* 103.1, pp. 1–46. DOI: `10.1257/aer.103.1.1`. URL: `https://www.aeaweb.org/articles?id=10.1257/aer.103.1.1`.

Assenova, Valentina A and Matthew Regele (2017). "Revisiting the effect of colonial institutions on comparative economic development". In: *Plos one* 12.5, e0177100.

Barro, Robert J (1991). "Economic growth in a cross section of countries". In: *The quarterly journal of economics* 106.2, pp. 407–443.

Diamond, Jared M and Doug Ordunio (1999). *Guns, germs, and steel*. Vol. 521. Books on Tape New York.

Appendix

Table 1: Results

	OLS		PDL (BRT)			PDL (BCCH)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Initial log (gdp)	-0.087	-1.203	-0.131	-0.137	-0.137	-0.132	-0.132	-0.132
Standard error	0.106	0.183	0.015	0.015	0.015	0.016	0.016	0.016
λ^{dz}			0.549	0.672	0.675	0.916	1.110	1.115
λ^{yx}			0.568	0.693	0.696	1.180	2.582	2.594
CI low	-0.295	-1.561	-0.160	-0.167	-0.167	-0.163	-0.163	-0.163
CI high	0.121	-0.845	-0.102	-0.108	-0.107	-0.102	-0.102	-0.102
t-statistic	-0.821	-6.574	-8.733	-9.133	-9.133	-8.250	-8.250	-8.250
Observations	102	76	76	76	76	76	76	76
Controls	0	26	26	344	370	26	344	370
Controls post penalty	0	26	2	1	1	0	0	0

Table 2: Included variables

Variable	Source
Democracy measure by ANRR	ANRR
Average democracy in the region*initial regime (leaving own country out)	ANRR
Index of market reforms (1960)	ANRR
mean distance to coast or river	ANRR
mean distance to coast	ANRR
mean distance to river	ANRR
% land area in geographical tropics	ANRR
dummy =1 if landlocked	AR
Geodesic centroid longitude	QG
Total land area	QG
Arable land area	QG
Absolute latitude	QG
Land suitability for agriculture	QG
Land suitability Gini	QG
Mean elevation	QG
Standard deviation of elevation	QG
Terrain roughness	QG
Temperature	QG
Precipitation	QG
Percentage of population living in tropical zones	QG
Africa dummy	QG
Asia dummy	QG
Oceania dummy	QG
Americas dummy	QG
Capital formation (% of GDP per year, avg. of available years 1970-2020)	WB
GDP per capita in 1970 (log)	WB
Annual growth in GDP per capita, 1970-2020	WB
Annual growth in population, 1970-2020	WB

Sources:

WB: World Bank

AR: Acemoglu, Naidu, et al. (2019)

QG: Ashraf and Galor (2013)

ANRR: Assenova and Regele (2017)