We just explained convergence factors with machine learning tools

Piotr Wójcik and Bartłomiej Wieczorek
pwojcik@wne.uw.edu.pl

Monthly joint seminar of QFRG and DSLab
21th January 2020

UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Section 1

Introduction

Objective of research

- analysis of **factors of economic growth and real economic convergence** is one of the important topics of research in macroeconomics
- it also has important implications for economic policy
- the aim of the research is to **identify relevant factors of economic growth** between countries in a (potentially) non-linear approach
- main focus is put on **comparing the conclusions of previous studies** on the importance of growth factors and conditional convergence with the results of nonlinear machine learning models
- this is **work in progress** – all comments and suggestions are welcome
- research financed by Polish National Science Center, project no. 2016/21/B/HS4/00670.

Objective of research – cont'd

- we assume that due to non-linear relationships between growth and its factors using linear models might lead to **incorrect conclusions**
- non-linear models may **better reflect true relationships**, but **the shape on relationship is not known in advance**
- that is why we use selected **machine learning tools** that can **flexibly adjust** to data and **uncover unknown real relationships**
- therefore we also assume that **machine learning tools** explain cross-country growth rates with higher accuracy than linear models
- all algorithms are applied on two empirical datasets used in previous empirical studies

Beta convergence

- *beta* convergence concentrates on the relationship between the average growth rate and initial income:

$$\frac{1}{T} \cdot \ln\left(\frac{y_{i,T}}{y_{i,t_0}}\right) = \alpha - \left(\frac{1 - e^{-\beta T}}{T}\right) \cdot \ln(y_{i,t_0}) + \gamma X_i + u_{i,t},$$

where y_{i,t_0} is the initial *per capita* income of country i, X_i is a vector of **structural characteristics conditioning convergence**, T is the number of periods and $u_{i,t}$ is the random disturbance

- positive estimate of β means that initially poorer countries grow faster than the richer (i.e. *beta* convergence)

Conditional *beta* convergence – difficulties

- Durlauf et al. (2009) indicate difficulties in verifying conditional *beta* convergence
- choice of control variables has a key impact on the inference about its occurrence
- there is **no consensus** on what their set is the best
- conclusions regarding the significance of individual factors **may contradict** each other
- different empirical approaches were used fo far

Section 2

Different empirical approaches used so far

(1) millions of regressions

- Sala-i-Martin (1997) considered almost all possible combinations of 62 variables and estimated **two million regressions**

$$y = \beta_C * C + \beta_Z * Z + \beta_X * X + \epsilon$$

where C is a vector of variables **included in every regression**, Z is a **variable of interest** and X is a subset of remaining variables of a given size

- significance of individual variables was measured by **weighted statistics based on all regressions in which this variable was the variable of interest**

(2) general-to-specific

- previous approach was criticized by Hendry and Krolzig (2004)
- they indicated that it is sufficient to estimate **one regression** and apply the **general to the specific** approach
- elimination of irrelevant variables is based on their **statistical significance**

(3) bayesian model averaging (BMA)

- alternatively one can use **bayesian model averaging** – applied to growth regressions by Sala-i-Martin et al. (2004)
- they called this specific approach **Bayesian Averaging of Classical Estimates** (BACE)
- it also requires estimating many regreesions but the weighting scheme is different

(3) bayesian model averaging – cont'd

- Bayes' rule in densities: $g(\beta|y) = \frac{f(y|\beta)g(\beta)}{f(y)}$
- In the equation above, $g(\beta)$ is the **prior density** of a parameter vector β.
- model averaging is a **special case of Bayes' rule**
- suppose we divide the parameter space into two regions (M_0 and M_1) – there regions we call hypotheses (e.g. $\beta > 0$ vs $\beta \leq 0$)
- given 2 regions, Bayes' rule implies

$$g(\beta|y) = P(M_0)\frac{f(y|\beta)g(\beta|M_0)}{f(y)} + P(M_1)\frac{f(y|\beta)g(\beta|M_1)}{f(y)}$$

(3) bayesian model averaging – cont'd

- if a researcher is incapable or unwilling to specify prior belief, a standard remedy is to apply **diffuse priors**
- however, it introduces a problem when regression models contain **different sets of variables**
- Sala-i-Martin (2004) shows that the ratio of posterior probabilities (**posterior odds ratio**) for two di fferent sets of variables, X and Z, for models M_0 and M_1 respectively, converges to:

$$\frac{P(M_0|y)}{P(M_1|y)} = \frac{P(M_0)}{P(M_1)} T^{(k_1-k_0)/2} \left(\frac{SSE_0}{SSE_1}\right)^{-T/2}$$

- here, T is a sample size and k_i is a number of included regressors in model M_i

(3) bayesian model averaging – cont'd

- in order to get weights for different models, one needs **posterior probabilities of each model**, not the odds ratio:

$$P(M_j|y) = \frac{P(M_j)T^{-k_j/2}SSE_j^{-T/2}}{\sum_{i=1}^{2^K} P(M_i)T^{-k_i/2}SSE_i^{-T/2}}$$

- once the model weights are calculated, the Bayes' rule says that the posterior density of a parameter is the **average of the posterior densities conditional on the models**
- a **posterior mean** is defined as the **expected value of a posterior distribution**
- taking expectations with respect to β, with 2^K terms instead of two, gives:

$$E(\beta|y) = \sum_{j=1}^{2^K} P(M_j|y)\hat{\beta}_{OLS}$$

(4) LASSO (recently)

- BMA was questioned by Ciccone, Jarociński (2010) **both on theoretical and empirical grounds**
- they conclude that BMA results are **sensitive to minor errors in measurement**
- for example the PWT 6.2 revision of the PWT 6.1 1960-1996 data leads to **substantial changes** regarding the role of government, international trade, demography, and geography
- Hofmarcher et al. (2015) show that LASSO **eliminates the collinearity** problem of explanatory variables and **gives growth predictions better than BMA**

LASSO

- linear regression (OLS) formula is expanded in the following way:

$$\min_{\hat{\beta}}[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|] = \min_{\hat{\beta}}[RSS + \lambda \sum_{j=1}^{p} |\beta_j|]$$

- adding a **penalty** in the optimization results in searching for parameters that **fit the data well**, but **are as small as possible** (nearest 0)
- parameters at less important variables will shink towards 0, their impact on the model **will be limited**
- in LASSO some parameters will be **exactly equal to zero**
- therefore LASSO will select **the most important explanatory variables** by eliminating non-important features from the model

LASSO – cont'd

- for $\lambda = 0$ the model simplifies to a regular linear regression (OLS)
- different values of λ will result in **different model parameters** β
- the optimal value of *lambda* should be found with the use of **cross validation**
- at the expense of a **certain bias** of the parameter estimates, LASSO often allows to obtain **more precise forecasts** on the test sample
- last but not least, LASSO can be used even when the initial **number of variables exceeds the number of observations**

Different approaches – summary

- all above mentioned approaches assume a **linear relationship** between growth and explanatory variables
- theory predicts **non-linear relationships** (Levine and Renelt, 1992),
- this is **confirmed by empirical studies** – growth process may have non-linear nature (Henderson et al. 2012)
- He and Xu (2019) show that identifying a variable as a statistically relevant factor of growth might result from **inappropriate linear model specification**
- a **correct non-linear specification does not confirm** its correlation with growth

Section 3

Methodology used in the analysis

Approach in this paper

- This presentation shows our **novel approach**
- Its main purpose is to identify **empirically relevant determinants of growth** with the use of machine learning algorithms
- we use tools that **allow for non-linearity** and **efficiently indicate important variables**
- apart from **LASSO** (already used before), support vector regression (**SVR**) random forests (**RF**), Adaptive Boosting (**AdaBoost**) and Extreme Gradient Boosting (**XGBoost**) are used
- all of those methods **deal successfully with linear and nonlinear relationships**
- in addition the **interpretable machine learning** (Murdoch, 2019) tools are applied to **understand the relationships**

SVM/SVR

- support vector machine (SVM, Vapnik, 1992) is a supervised learning model that **fits a hyperplane** that is positioned as close to all observations as possible
- it can be used for both classification (SVM) and regression (SVR) problems
- SVR fits the hyperplane to the data, but ignores the observations within a given, ϵ distance from the hyperplane
- the use of **kernel functions** ($\phi(x)$) allows to model **nonlinear relationships between variables**
- it is a function of two variables $k(x, z)$, such that $k(x, z) > 0$ and $k(x, z) = k(z, x)$
- kernels allow to indirectly map the data to a **more dimensional space** in which a linear relationship fits the data well
- this allows to achieve **an analogous effect**, like extending the set of features, but is **much less computationally intensive**

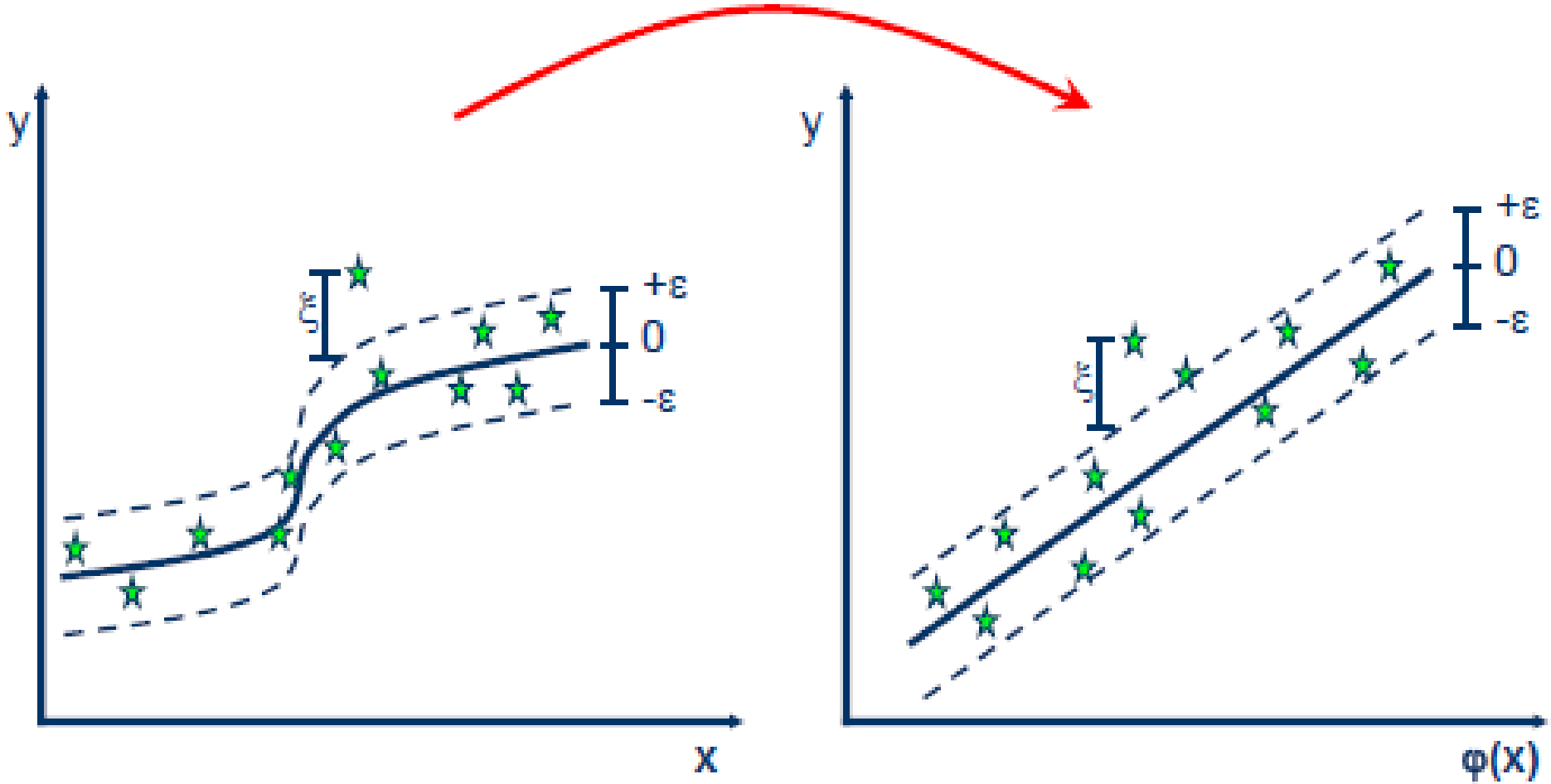
SVM/SVR – kernel trick example

- suppose one has two X variables, and the kernel function is simply a quadratic function
- then one can write:

$$\begin{aligned} k(x,\beta) &= (x'\beta)^2 \\ &= x_1^2\beta_1^2 + 2x_1\beta_2 x_2\beta_2 + x_2^2\beta_2^2 \\ &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)(\beta_1^2, \sqrt{2}\beta_1\beta_2, \beta_2^2)' \\ &= \phi(x)'\phi(\beta) = \langle \phi(x), \phi(\beta) \rangle \end{aligned}$$

- the effect is the same as if one used the **squares of both variables** and their **interaction** into the model

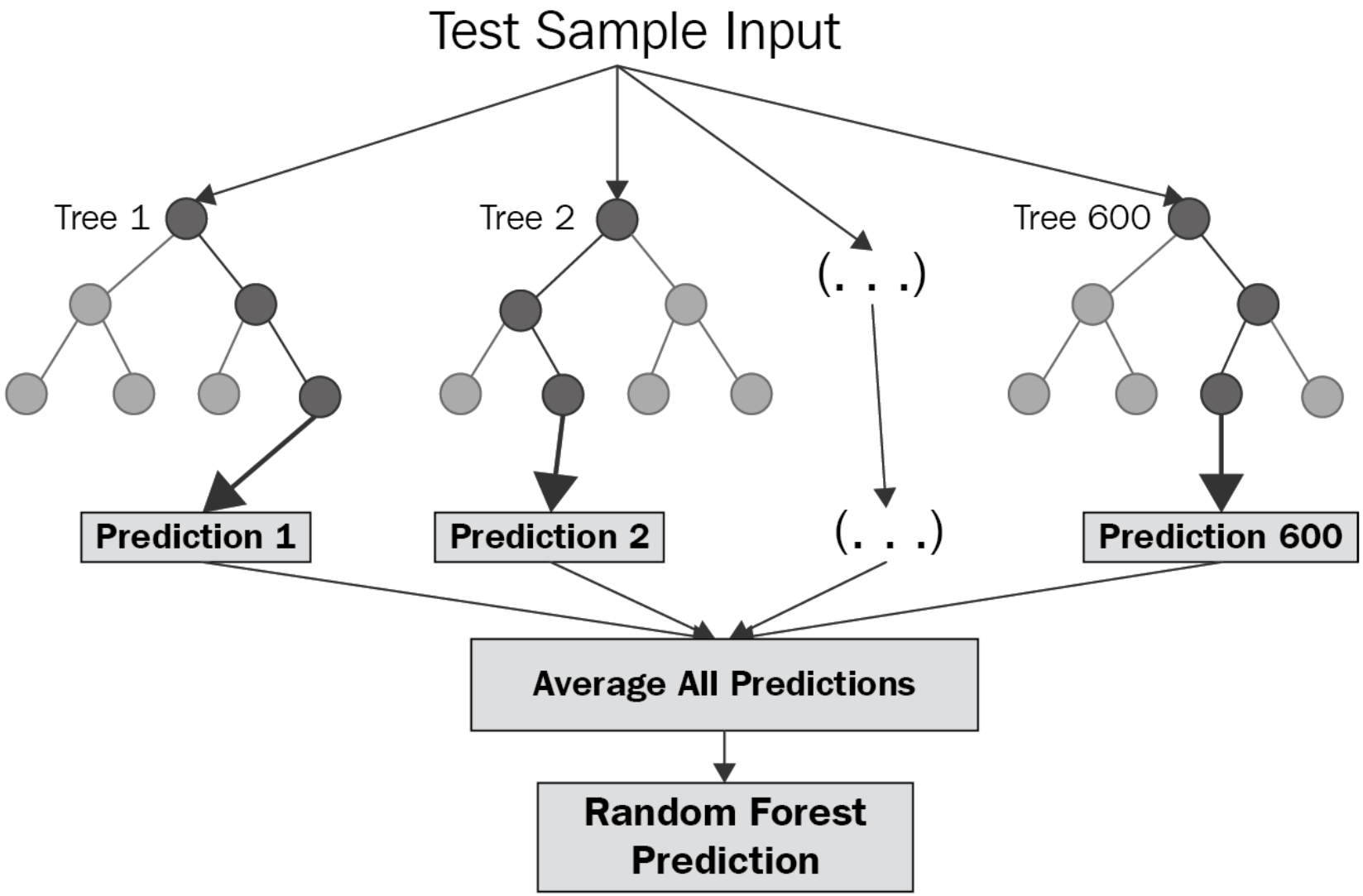
SVM/SVR – visualization



Random forests

- random forests introduced by Breiman (2001) are a **combination of tree predictors**
- each tree is trained on a **different bootstrap sample** of original data
- in addition, at each division in each tree **only a random subset of all predictors is considered**
- to calculate the final prediction **forecasts from all trees are averaged**
- random forests are **robust to the problem of multicollinearity** and **indifferent to non-linear interlinkages** between variables

Random forests – visualization



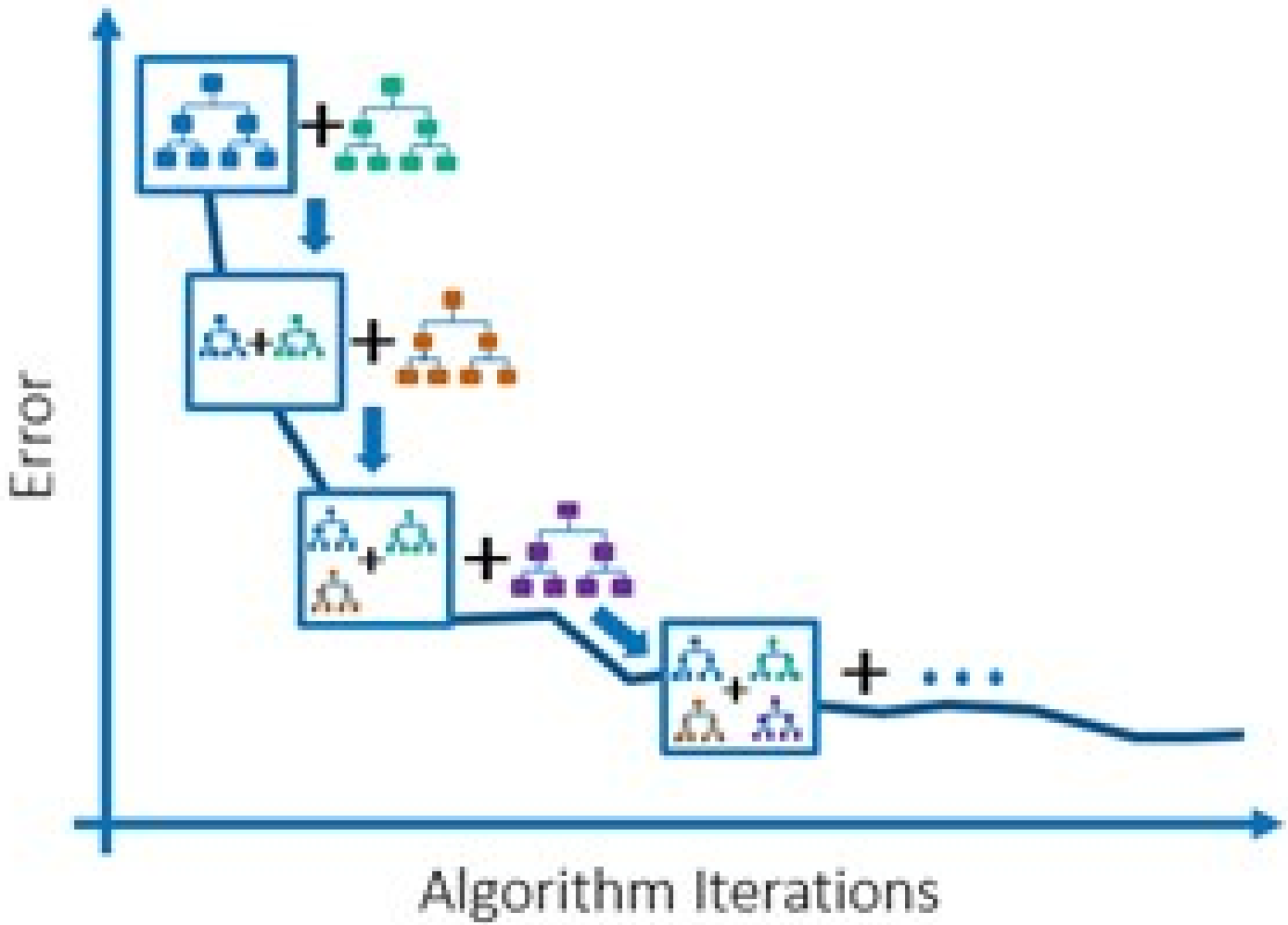
Adaptive Boosting

- **boosting** is a group of algorithms that convert a set of weak learners (classifier which accuracy is just slightly higher than random classifier) into a strong learner
- new learners (models) are added **sequentially**
- **AdaBoost** (from **Ada**ptive **Boost**ing) is a simple iterative boosting algorithm
- it starts by training / testing a weak learner (usually a stump – decision/regression tree of the depth of 1) on the data, **weighting each observation equally**
- observations which are **misclassified get their weights increased** for the next rounds, while correct ones get their weights decreased

Extreme Gradient Boosting

- gradient boosting: boosting method; based on the **gradient descent algorithm** for the minimization of the loss function
- in simple: in the next iteration a **model is applied on the residuals from the previous step**
- **XGBoost** – variant of gradient boosting, but additionally equipped with several features, like randomization parameter (which reduces correlation between trees) and approximation methods to find more effective path to reach the minimum of loss function

Boosting – visualization



Leave One Out Cross Validation – LOOCV

- the optimal values of hyperparameters are found with the use of **leave-one-out cross validation** (**LOOCV**)
- each model is estimated n times on the sample without the $1st, 2nd, 3rd, \ldots$ observation respectively
- each time the single observation left aside is used as the **test sample** – for prediction
- prediction is compared with the real value and **prediction error** is calculated
- in this case one obtains n prediction errors
- the estimate of the prediction error is the **average of these values** (MAE)
- the procedure is repeated **for every combination of hyperparameters** considered
- finally we select the combination that **minimizes the prediction error**

Interpretable machine learning (**IML**) – variable importance

- we use a **Permutation Importance** (or **Mean Decrease Accuracy**) method
- it describes how much the model's performance relies on different covariates

IML – variable importance – cont'd

- Consider a set of n observations for a set of p explanatory variables
- For each explanatory variable X_j :
  - replace vector x_j of observed values of X_j by vector x^*_{-j} of permuted values.
  - Calculate model predictions $\widetilde{y}_{-j}$ for the modified data
  - Calculate the value of loss function for modified data :

$$L^{*,-i} = Loss(y, \widetilde{y}_{-j})$$

  - Variable importance is calculated as

$$VI(x^j) = L^{*,-j}/L$$

where L is the value of the loss function for the original data

IML – understanding the relationship

- **ceteris paribus profiles** (CPPs) are plotted
- they show how a conditional **expectation of the outcome changes** with a particular predictor **keeping all other variables constant**
- averaging CPPs over all observations shows the **expected model response** as a function of a selected feature
- it is known as a **Partial Dependence Profile** (Friedman, 2000)

Section 4

Empirical analysis

Datasets

Fernández, Ley and Steel (2001) dataset – refered to hereafter as **FLS (2001)**:

- includes 41 explanatory variables for 72 countries
- analysed by: Fernández et al. (2001), Hendry & Krolzig (2004), Ley & Steel (2007)

Sala-i-Martin, Doppelhofer and Miller dataset – refered to hereafter as **SDM (2004)**:

- includes 67 explanatory variables for 88 countries
- analysed by: Sala-i-Martin et al. (2004), Doppelhofer & Weeks (2011)

Variable importance on FLS data

	rank of the variable importance									
variable	FLS (2001)	S (1997)	OLS (HK, 2004)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)	RF	AdaB	XGB
GDP level in 1960	1	1	1	1	1	1	1	10	10	24
Fraction Confucian	2	1	11	6	6	5	5	4	9	12
Life expectancy	3	7	2	2	2	3	2	6	7	20
Equipment investment	4	1	9	8	10	9	8	1	2	2
Sub-Saharan dummy	5	10	4	4	5	4	3	32	30	38
Fraction Muslim	6	1	-	17	20	27	26	23	27	15
Rule of law	7	1	10	12	13	10	10	27	37	35
Number of Years open economy	8	1	-	33	29	36	25	5	4	17
Degree of Capitalism	9	17	-	18	24	19	16	29	36	30
Fraction Protestant	10	22	-	16	22	11	11	16	11	18
Fraction GDP in mining	11	13	16	13	16	15	13	25	22	13
Non-Equipment Investment	12	19	-	15	19	16	12	3	3	6
Latin American dummy	13	8	7	9	8	8	9	26	20	26
Primary School Enrollment, 1960	14	15	12	11	12	12	17	21	23	9
Fraction Buddhist	15	23	-	22	25	18	14	2	1	1
Black Market Premium	16	30	-	19	21	20	19	28	13	34
Fraction Catholic	17	24	-	-	-	37	39	12	24	10
Civil Liberties	18	10	-	20	14	28	33	24	31	14
Fraction Hindu	19	35	3	3	3	2	4	35	40	38
Primary exports, 1970	20	16	-	30	27	30	27	13	5	3

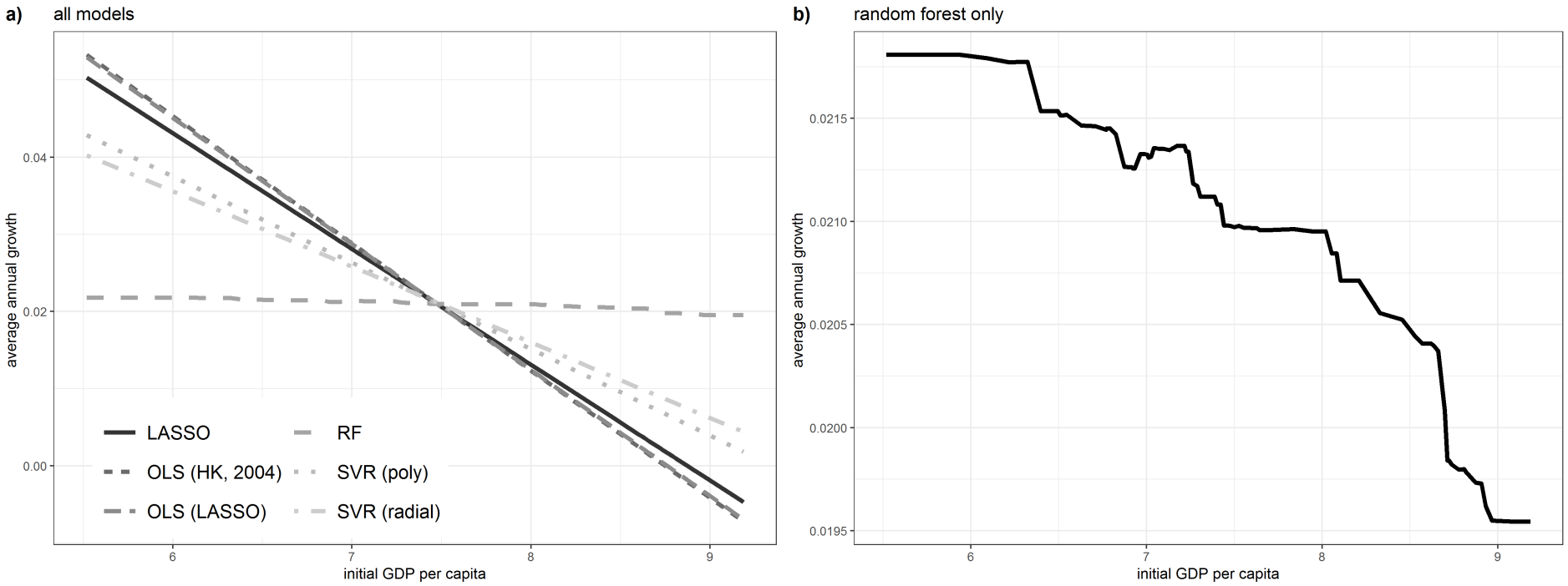
Variable importance on FLS data – cont'd

	rank of the variable importance - cont'd									
variable	FLS (2001)	S (1997)	OLS (HK, 2004)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)	RF	AdaB	XGB
Political Rights	21	8	-	23	36	26	31	14	28	16
Exchange rate distortions	22	21	-	31	30	38	36	18	19	27
Age	23	27	-	29	32	21	21	30	16	29
War dummy	24	18	-	28	28	31	32	40	41	31
Size labor force	25	28	5	5	4	6	7	9	14	4
Fraction speaking foreign language	26	29	-	37	33	29	34	31	26	28
Fraction of Population Speaking English	27	26	-	26	23	24	28	38	29	32
Ethnolinguistic fractionalization	28	36	14	10	11	13	15	19	18	23
Spanish Colony dummy	29	25	8	14	9	14	20	34	34	38
stdev of black-market premium	30	14	-	35	35	25	24	7	15	5
French Colony dummy	31	34	15	25	17	32	37	37	39	36
Absolute latitude	32	20	-	-	-	35	29	8	8	11
Ratio workers to population	33	32	-	32	37	33	30	20	12	7
Higher education enrollment	34	39	6	7	7	7	6	15	21	8
Population Growth	35	31	-	-	-	34	35	17	25	22
British Colony dummy	36	39	13	24	15	17	23	39	33	37
Outward Orientation	37	37	-	21	18	23	22	41	38	38
Fraction Jewish	38	33	-	36	34	41	41	36	35	33
Revolutions and coups	39	12	-	34	31	39	38	33	32	25
Public Education Share	40	38	-	27	26	22	18	11	6	19
Area (Scale Efiect)	41	41	-	-	-	40	40	22	17	21

Variable importance on FLS data – comments

- all models indicate conditional convergence – importance of **GDP level in 1960**
- all models agree on the importance of **fraction Confucian**, **life expectancy** and **equipment investment**
- however, machine learning tools **do not confirm** 5-10 out of 20 most important variables indicated by FLS
- LASSO and SVR are consistent with **HK** in terms of selected variables
- RF and boosting algorithms indicate the importance of openess, primary exports and non-equipment investment
- there are important growth determinants missed by BMA and confirmed by all other methods: **size of labour force**, **ethnolinguistic fractionalization** and **higher education enrollment**

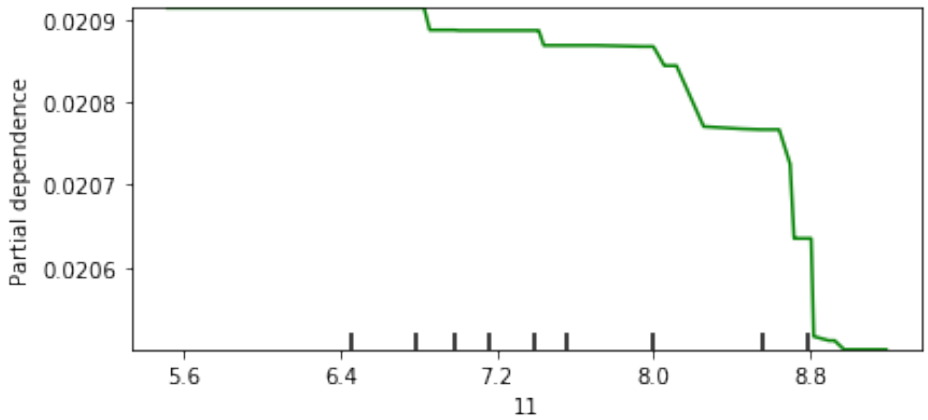
Partial Dependence Plots for initial income on FLS data



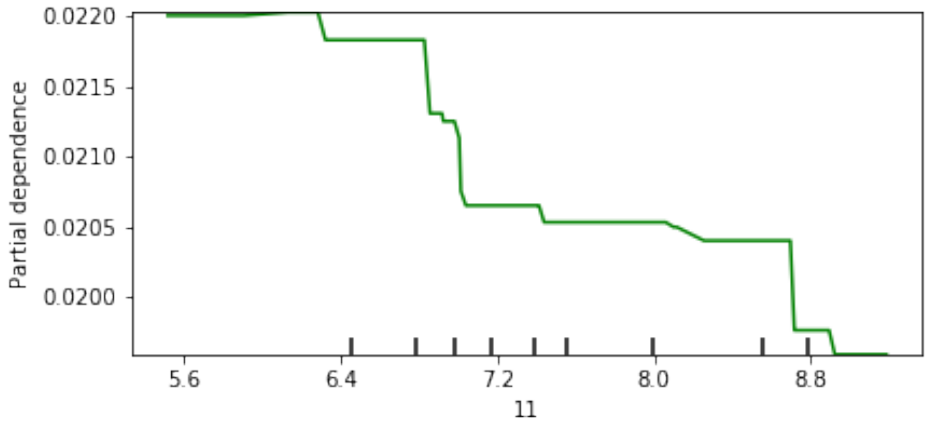
- linear and negative relationship between the growth rate and initial GDP in almost all cases
- for random forest the slope changes for different intervals of initial income (convergence clubs)

PDPs for initial income on FLS data (boosting)

adaboost



xbgoost



Measures of model fit for FLS data

model	RMSE	MAE	MAPE	R2
OLS (HK, 2004)	0.0055	0.0041	0.3259	0.9072
LASSO	0.0040	0.0031	0.2849	0.9521
OLS / LASSO	0.0037	0.0029	0.3704	0.9590
SVR (radial)	0.0035	0.0025	0.4364	0.9621
SVR (poly)	0.0033	0.0024	0.4348	0.9672
RF	0.0045	0.0033	0.3231	0.9378
XGBoost	0.0074	0.0074	0.1463	0.9945
AdaBoost	0.0083	0.0083	0.0375	0.9988

- each of the estimated models explains more than 90% of the variability of growth
- almost anything is **better than a linear model**
- **nonlinear SVR models** explain the relationships better than LASSO and **have the lowest prediction errors**

Variable importance on SDM data

	rank of the variable importance								
variable	SDM (2004)	DW (2011)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)	RF	AdaB	XGB
East Asian dummy	1	1	1	2	1	3	1	2	2
Primary schooling in 1960	2	2	3	5	4	4	3	5	3
Investment price	3	3	4	3	3	9	4	7	12
GDP in 1960 (log)	4	4	-	-	27	38	32	20	23
Fraction of tropical area	5	5	5	1	14	21	20	65	17
Population density coastal in 1960s	6	6	11	7	11	13	12	24	16
Malaria prevalence in 1960s	7	7	9	10	9	1	10	1	1
Life expectancy in 1960	8	8	-	-	10	2	9	8	5
Fraction Confucian	9	9	2	4	2	10	2	11	15
African dummy	10	10	-	-	8	15	8	66	59
Latin American dummy	11	11	-	-	48	51	25	56	32
Fraction GDP in mining	12	12	-	-	23	41	23	45	28
Spanish colony	13	13	-	-	28	49	39	64	54
Years Open 1950-94	14	14	6	11	7	7	6	6	4
Fraction Muslim	15	15	-	-	57	47	47	50	31
Fraction Buddhist	16	16	7	6	5	5	7	4	19
Ethnolinguistic fractionalization	17	17	-	-	19	26	14	12	8
Government consumption share in 1960s	18	18	10	8	15	35	15	52	43
Population density 1960	19	19	-	-	46	12	43	67	9
Real exchange rate distortions	20	20	12	9	6	8	5	43	29
Fraction speaking foreign language	21	21	-	-	16	52	18	53	38

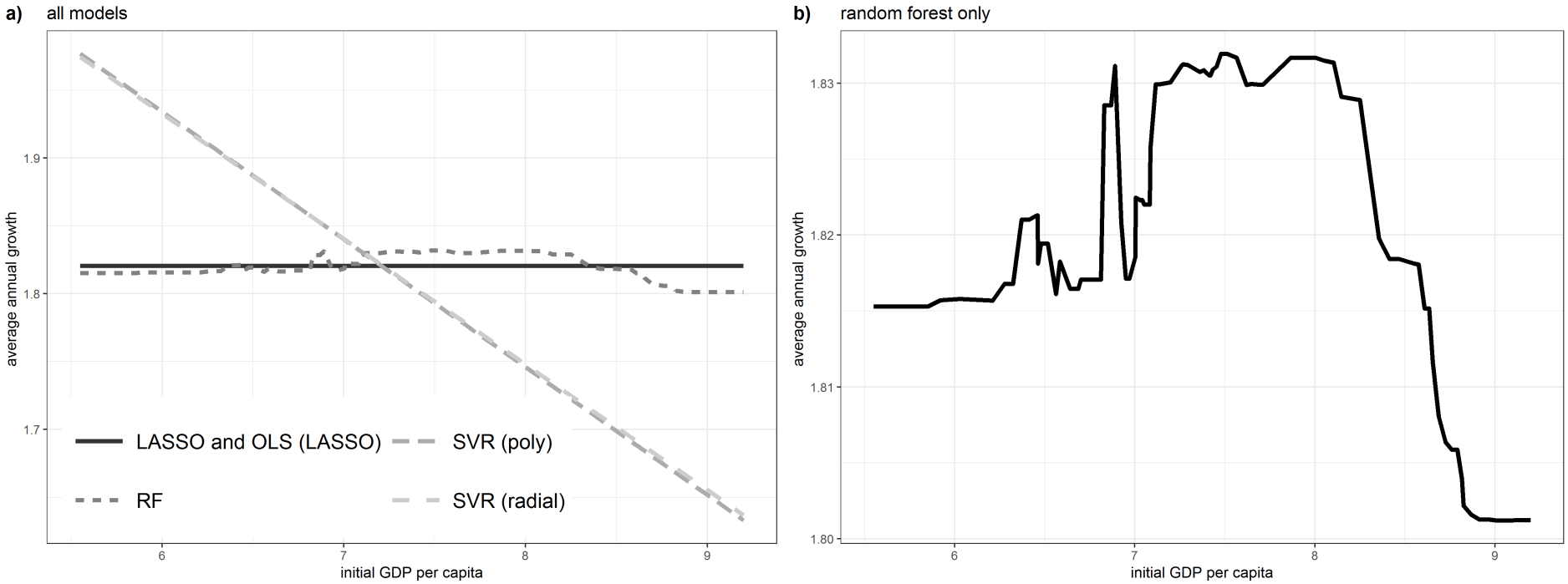
Variable importance on SDM data – cont'd

	rank of the variable importance - c.d.								
variable	SDM (2004)	DW (2011)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)	RF	AdaB	XGB
(Imports exports)/GDP	22	22	-	-	13	40	16	32	10
Political rights	23	23	-	-	63	19	63	29	13
Government share of GDP	24	24	-	-	18	44	17	30	36
Higher education in 1960	25	25	-	-	39	14	33	23	34
Fraction population in tropics	26	26	8	12	12	17	11	14	42
Primary exports in 1970	27	27	-	-	17	30	13	31	46
Public investment share	28	28	-	-	64	50	60	33	27
Fraction Protestant	29	29	-	-	24	45	21	28	24
Fraction Hindu	30	30	-	-	42	59	44	56	35
Fraction population less than 15	31	31	-	-	37	33	40	22	52
Air distance to big cities	32	32	-	-	58	23	57	40	14
Government consumption share deflated with GDP prices	33	33	10	8	15	35	15	52	43
Absolute latitude	34	34	-	-	43	6	42	9	11
Fraction Catholic	35	35	-	-	67	34	67	13	53
Fertility in 1960's	36	36	-	-	30	20	36	38	7
European dummy	37	37	-	-	62	63	51	56	59
Outward orientation	38	38	-	-	45	62	52	42	59
Colony dummy	39	39	-	-	26	65	29	56	58
Civil liberties	40	40	-	-	25	53	19	19	26
Revolutions and coups	41	41	-	-	29	43	22	35	33
British colony	42	42	-	-	40	61	24	53	44
Hydrocarbon deposits in 1993	43	43	-	-	31	28	45	15	37
Fraction population over 65	44	44	-	-	54	27	53	16	30

Variable importance on SDM data – comments

- DW (2011) gives exatcly the **same ranking** as SDM (2004)
- **three most important variables** from SDM (2004) and DW (2011) **confirmed by all other models**
- GDP in 1960 (log) is **excluded in LASSO** and in **SVR, RF and boosting its rank is quite high**
- this suggests lack of conditional convergence, unlike in the BMA approach
- therefore in a **correct non-linear specification** initial GDP is **not** (strongly) correlated with growth
- several variables seem to have **stronger impact on growth** according to machine learning algorithms than in BMA: *fraction Confucian*, *years open 1950-94*, *fraction Buddhist* or *government consumption share in 1960s* (negative impact)

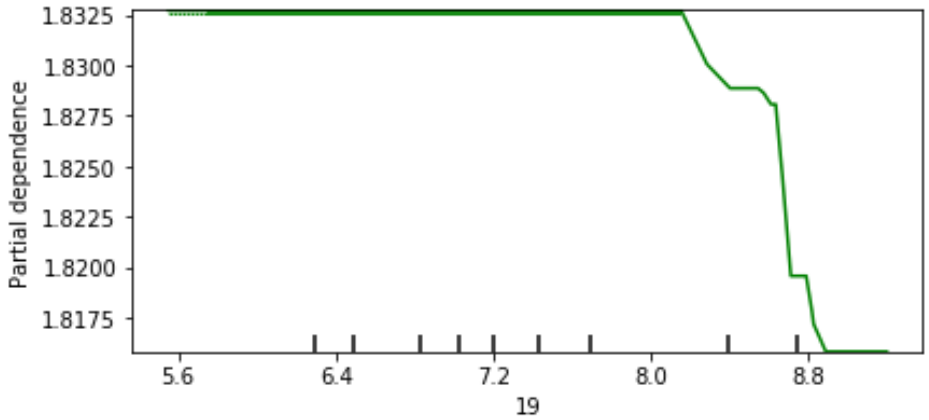
Partial Dependence Plots for initial income on SDM data



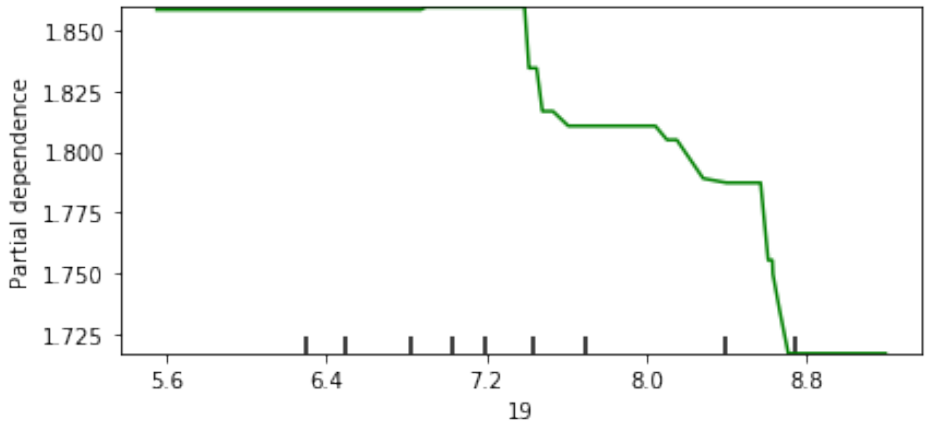
- only SVR models show a linear and negative relationship between growth and initial GDP
- for LASSO the line is flat
- results of random forest show that only countries that are initially richest converge (convergence clubs)

PDPs for initial income on SDM data (boosting)

adaboost



xbgoost



Measures of model fit for SDM data

model	RMSE	MAE	R2
LASSO	1.0600	0.8245	0.6859
OLS / LASSO	0.9867	0.7613	0.7278
SVR (radial)	0.9312	0.6573	0.7576
SVR (poly)	0.9313	0.6576	0.7575
RF	0.5052	0.3690	0.9287
XGBoost	0.8677	0.8869	0.9959
Adaboost	0.9455	0.9455	0.9999

- random forest and boosting algorithms explain more than 90% of the variability of growth
- generally non-linear machine learning algorithms (RF and SVR) are better than the linear LASSO or OLS

Section 5

Conclusions

Conclusions

- using **linear approach** to modelling growth may **lead to incorrect conclusions** regarding the importance of its factors
- within a (correct?) non-linear specification **conditional cross-country convergence was not observed**
- instead, one could only identify (weak) convergence of clubs – for the initially richest countries
- machine learning algorithms are useful tools to model relationships when its shape is not known
- they explain more variability of growth than the traditional linear models or BMA, often questioned for their subjectiveness

Questions and discussion

References

Cuaresma, J. C., Doppelhofer, G., & Feldkircher, M. (2014). The determinants of economic growth in European regions. *Regional Studies*, *48*(1), 44–67. https://doi.org/10.1080/00343404.2012.678824

Doppelhofer, G., & Weeks, M. (2011). *Robust growth determinants* (Discussion Paper Nos. 3). NHH Dept. of Economics.

Durlauf, S. N., Johnson, P. A., & Temple, J. R. W. (2009). The econometrics of convergence. In T. C. Mills & K. Patterson (Eds.), *Palgrave handbook of econometrics: Volume 2: Applied econometrics* (pp. 1087–1118). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244405_23

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563–576. https://doi.org/10.1002/jae.623

Hendry, D. F., & Krolzig, H.-M. (2004). We ran one regression. *Oxford*