STABLE ESTIMATION OF GRANGER-CAUSAL FACTORS OF COUNTRY-LEVEL INNOVATION

Aurélie C. Lozano, Prasanna Sattigeri, Aleksandra Mojsilović, and Kush R. Varshney

Mathematical Sciences Department
IBM Thomas J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598 USA

ABSTRACT

Increasing the innovativeness of a country is a known way to uplift its people, but how to increase the innovativeness of a country is not well understood. In this paper, we develop such an understanding by analyzing time series of global competitiveness index data and of development indicators data. Specifically, we estimate causative factors of innovation that countries can invest in or otherwise pursue through policy decisions. We take a sparse multivariate regression approach to find Granger-causal development indicators for an innovation index, which is based on group orthogonal matching pursuit. Due to high correlation between various indicators, small perturbations can cause the support of the sparse regression solution to change drastically with nonzeros shifting among correlated sets of indicators. Such behavior does not detract from predictive accuracy, but is undesirable for interpretation and decision making. To address this issue, we use randomization and stability selection techniques. We show favorable empirical results.

***Index Terms*—** Granger causality, orthogonal matching pursuit, social good, sparse regression, stability selection

1. INTRODUCTION

We often think of increasing access to clean water, food, health care, education, and energy as key activities of international development, but science, technology, and especially innovation are also keys to sustainable and inclusive development. As countries become more innovative, access to appropriate technologies improves the living conditions of the citizens and allows them to be more productive and have more income [1]. In fact, among the 17 sustainable development goals ratified by the member states of the United Nations in September 2015, one is focused specifically on supporting domestic technology development, research, and innovation in developing countries [2].

To foster innovation, first, one must be able to measure it, and second, understand the levers to increase it. On the first point, measuring innovation is an ongoing effort in the international community: currently there are many innovation indices, surveys, and reports. A comprehensive recapitulation of these reports observes a significant diversity in how they approach the topic of innovation [3]: some have updated their innovation index yearly to provide a basis for comparisons over time, while others have only published their innovation index once. Similarly, the geographical scope of innovation covered varies as well. The study highlights several other issues with the existing approaches that we discuss further in [4], but the main shortcoming from the perspective of the current work is the lack of causal factors necessary for policy-making.

Associations, correlations, and predictive models have many uses, but are not sufficient to drive policy because they do not indicate causal relationships that would permit one to change inputs and expect desired changes in the output of interest. Granger causality is one form of causal inference based on time series analysis [5]. In its original form, Granger causality is an operational statement between two time series. If past values of one time series improve the prediction of a second time series above the predictability offered by the past values of that second time series itself, then the first time series Granger-causes the second. This notion has been extended to multiple time series in [6], with the time series prediction performed using sparse multivariate regression [7].

In this work, we perform Granger causality analysis to understand which country-level development indicators, such as 'business costs of crime and violence,' 'internet access in schools,' and 'buyer sophistication' have a causal relationship to an overall innovation index of a country. In the empirical section, the innovation index we use as a response variable comes from the World Economic Forum's (WEF's) Global Competitiveness Report and the country-level development indicators come from the same source as well as the World Bank's World Development Indicators.

The reason for performing causal analysis is to provide insight and understanding to policymakers, but such insight is difficult in the presence of highly correlated variables, which we have in our problem because many development indicators are highly correlated with each other. In sparse regression, it is well know that there is instability in solutions when there

978-1-5090-4545-7/16/$31.00 ©2016 IEEE

are highly correlated variables because one variable with a nonzero coefficient can be switched out for another that is highly correlated with it and have little change to the prediction [8, 9]. In this work, we address this issue by performing stability selection, a methodology in which randomization is used to create multiple solutions and only variables that appear frequently in many solutions are retained [10].

This work attempts to provide a step forward towards understanding the key drivers of innovation, which would enable the measurement of innovation capabilities in an ongoing, dynamic, regional and action-oriented way. We utilize a data-driven approach to identify measurable drivers of innovation, based on causal analysis between development metrics, innovation indicators, and perceived innovation levels. Our hope is that this work will contribute to better understanding of what makes a country innovative, that it will offer actionable guidance on improving innovation outcomes at global and country levels, and eventually lead toward the construction of an open innovation index.

This paper is organized as follows. In Section 2 we provide details of the causality analysis and stability selection. In Section 3 we provide an overview of datasets used and describe the indicators and innovation scores considered. Section 4 presents illustrative empirical results. Conclusions and next steps are found in Section 5.

2. CAUSAL ANALYSIS

Our country level analysis considers each country separately, and attempts to identify factors that are causally related to its innovation output. This analysis will shed light on a lever a particular country might be applying (either favorably or adversely) that is causally related to its level of innovation. Note, however, that if for a given country there is no activity in a particular metric (for example, no changes in R & D investment, or growth in Internet users), despite its potential relevance to innovation, this metric will not be identified as a causal indicator for the given country.

To identify indicators that are causally related with innovation measurements, we formulate our approach in terms of the notion of Granger causality [5]. Granger causality is an operational notion of causality that has been employed in statistics, econometrics, machine learning and data mining. The main argument is that time series a is a potential cause of time series b, if a significantly helps improve the prediction of the future values of b. Our approach combines two main components: (i) generalized Granger causal modeling via sparse regression to determine causal relationships between multiple time series simultaneously; (ii) randomization within sparse regression to alleviate the issue of correlated predictors. We describe below the specifics of these two components.

2.1. Multivariate Granger Causality via Sparse Regression

In our analysis, we test if the past development metrics are predictive of the future innovation index score. Let us consider a country C with N development metrics. Each metric is represented as a time-series with T time-points. Each country also has a target time-series, which corresponds to the innovation score. We adopt a generalized Granger Causality test described in [6], where we jointly assess the potential causal effect of multiple time series on innovation rather than preforming separate pair-wise tests between a potential cause and innovation. To do so, we predict the innovation metric value at time t using its past d values at time $t-1$ to $t-d$ and the past d values of all the development metrics. The prediction problem is solved using sparse linear regression. Sparse linear regression approaches jointly perform variable selection and parameter estimation. The selected variables are considered as causal drivers of the innovation score.

In our experiments we employed (Group)-OMP [7] but could have also used (Group) Lasso [11], among other alternatives. The stopping point for variable selection was tuned using approximate C_p criterion [11] to maximize the regression performance. Due to the limited time span of the available data, we only consider a lag of $d = 1$ in Section 4. Considering larger lags is essential. To do so, as future work we plan to assemble a dataset with longer history.

2.2. Handling Correlated Time Series using Stability Selection

A key challenge is that of the correlation among time series, which is exarcerbated by the small historical time span vs. the large number of time series under consideration. To tackle this issue, we enhance our causal modeling approach to introduce randomization in the Group OMP procedure, following the randomized OMP approach for stability selection [10]. At each iteration, instead of picking the predictor which leads to the largest reduction of the residual error, the randomized Group OMP will pick a predictor at random within a set that sufficiently reduced the error. The insight behind this technique is that, given multiple causal modeling run, true associations of covariates to innovation will be selected at high frequency because true association signals are likely to be less sensitive to the randomization. Such a scheme is expected to help distinguishing between true and spurious causes. In our experiments for each coutry we perform 100 randomized runs and count how many times each time series is selected as cause. We report this count which we call *stability score.*

3. DATASETS

Our analysis seeks to discover input/output relationships between historical data on numerous country-level metric (in-

Macroeconomic environment	Market Size
Higher education and training	Infrastructure
Goods market efficiency	Health and education
Labor market efficiency	Financial markets
Technological readiness	Institution
Business sophistication	Innovation

Table 1. The 12 pillars of the Global Competitiveness Index.

put) and perceived levels of innovation (output). To do so, we onsider data from the Global Competitiveness Report (GCR) [12], and country level metrics in the World Development Indicators (WDI) [13].

3.1. Global Competitiveness Report

To capture information on the perceived level of innovation (output) we use the Global Competitiveness Report. GCR is a yearly report published by the WEF since 2004 [14]. The report ranks countries based on the Global Competitiveness Index, (GCI), which assesses the ability of countries to provide high levels of prosperity to their citizens. This in turn depends on how productively a country uses available resources. Therefore, the Global Competitiveness Index measures the set of institutions, policies, and factors that set the sustainable current and medium-term levels of economic prosperity" [15]. Over 110 variables contribute to the index; two thirds of them come from the Executive Opinion Survey, and one third comes from publicly available sources, such as the UN. This survey contains the responses of roughly 14000 business leaders from 142 economies. The GCI variables are organized into twelve pillars (see Table 1), with each pillar representing an area considered as an important determinant of competitiveness. Each of the pillars is further divided into several sub-components, which help measure that pillar. Of particular interest to this analysis is the 12th pillar: Innovation. We will use this pillar score from the years 2007–2014 as the ground truth for a country's innovation score.

3.2. World Development Indicators

World Development Indicators form the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates. This statistical reference includes over 1500 indicators covering more than 150 economies. The annual publication is released in April of each year, and the online database is updated three times a year. The World Banks Open Data site provides access to the WDI database free of charge to all users. A selection of the WDI data is featured at data.worldbank.org. We will use the statistics provided by these indicators as inputs to our analyses.

4. EMPIRICAL RESULTS

Example of causal relationships uncovered between the non-innovation related GCI metrics and overall innovation score are presented in Table 2 for Kenya, Table 3 for Sweden, and Table 4 for Mauritius ranked by their stability score. An exhaustive report for all countries and for both GCI and WDI metrics is the object of future work. To shed light on how specific a given cause might be to a particular country, we also report a so-called *global score*. This score equals the selection frequency of this relationship over all countries divided by the maximum selection frequency over all relationships and all countries. The higher the score, the less "country specific" the relationship; in other words the more commonly encountered the relationship is across countries.

As can be seen from the tables, metrics concerning education are commonly selected as drivers of innovation. It is noteworthy to also notice that certain time series characterizing the overall "state" of a country are identified (e.g. Business impact of HIV AIDS, quality of electric supply). Though we do not think of these are obvious direct causes of innovation, they are certainly reflecting serious impediments.

5. CONCLUSION

In this work, we proposed an analysis that would contribute to better understanding of innovation and how to drive actionable insights for diverse countries and diverse innovation conditions around the world. Our approach is data-driven, and aims to produce results that are repeatable, systematic and objective that can lead to dynamic country-level benchmarks. This would allow policymakers and organizations like WEF to more efficiently shape policy and interventions in underdeveloped countries in order to increase their developmental activity and capacity for innovation.

As future work, we plan enrich our dataset by incorporating more indicators and longer history and also further explore enhancements to our causal models to handle correlated indicators. In addition to the country level analysis, we plan to conduct group level analyses on groups of similar countries (either in terms of geography, level of economic development, or WDI neighborhoods). Such multitask learning analysis sheds insight into actions or levers a majority of the countries from the group are applying. These extensions will improve the robustness of the models and make them more informative and reliable. We also plan to build visualizations that would allow us to communicate actionable insights and country-level benchmarks in an easily consumable way.

Table 2. Top non-innovation GCI metrics (ranked by their stability score) with causal relationship to the overall innovation score for Kenya

CAUSAL DRIVER	STABILITY SCORE (%)	GLOBAL SCORE (%)
Availability of latest technologies	100	77
Degree of customer orientation	99	58
Quality of math and science education	84	94
Business costs of crime and violence	66	61
Control of international distribution	55	73
Business impact of HIV AIDS	51	62

Table 3. Top non-innovation GCI metrics (ranked by their stability score) with causal relationship to the overall innovation score for Sweden

CAUSAL DRIVER	STABILITY SCORE (%)	GLOBAL SCORE (%)
Quality of overall infrastucture	100	63.7
Quality of math and science education	52	94
Quality of management schools	48	70.9

Table 4. Top non-innovation GCI metrics (ranked by their stability score) with causal relationship to the overall innovation score for Mauritius

CAUSAL DRIVER	STABILITY SCORE (%)	GLOBAL SCORE (%)
Value chain breadth	55	62
Efficacy of corporate boards	40	71
Quality of electricity supply	40	42
Quality of management schools	40	71
Internet access in schools	32	71
Buyer sophistication	31	83
Production process sophistication	31	52

6. REFERENCES

[1] "Science, technology and innovation and intellectual property rights: The vision for development," UN System Task Team on the Post-2015 UN Development Agenda, May 2012.

[2] "Sustainable development goals," `http://www.un.org/sustainabledevelopment/sustainable-development-goals`, United Nations, Sept. 2015.

[3] "The World Economic Forum Economics of Innovation Global Agenda Council evaluates leading indicators of innovation," World Economic Forum, 2015.

[4] P. Sattigeri, A. C. Lozano, A. Mojsilović, K. R. Varshney, and M. Naghshineh, "Understanding innovation to drive sustainable development," in *Proc. ICML Workshop #Data4Good: Machine Learning in Social Good Applications*, June 2016, pp. 21–25.

[5] C. W. J. Granger, "Some recent development in a concept of causality," *Journal of Econometrics*, vol. 39, no. 1, pp. 199–211, 1988.

[6] A. C. Lozano and V. Sindhwani, "Block variable selection in multivariate regression and high-dimensional causal inference," in *Advances in Neural Information Processing Systems*, 2010, pp. 1486–1494.

[7] A. C. Lozano, G. Swirszcz, and N. Abe, "Grouped orthogonal matching pursuit for variable selection and prediction," in *Advances in Neural Information Processing Systems*, 2009, pp. 1150–1158.

[8] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang, "Correlated variables in regression: Clustering and sparse estimation," *Journal of Statistical Planning and Inference*, vol. 143, no. 11, pp. 1835–1858, 2013.

[9] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.

[10] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pp. 49–67, 2006.

[12] K Schwab and X. Sala-i Martín, "The global competitiveness report 2013–2014: Full data edition," in *World Economic Forum*, 2013, p. 551.

[13] *World Development Indicators (WDI) 2014*, World Bank Publications, 2014.

[14] "Global competitiveness reports (GCRs)," `http://reports.weforum.org/global-competitiveness-report-2015-2016`.

[15] "Global competitiveness index (GCI) - methodology," `http://reports.weforum.org/global-competitiveness-report-2014-2015/methodology`.