Sufficient reductions in regression with mixed predictors

Efstathia Bura

ASTAT
Institute of Statistics and Mathematical Methods in Economics

joint with Liliana Forzani, Rodrigo García Arancibia, Pamela Llop and Diego Tomassi at Universidad Nacional del Litoral.

Contents

Contents

The problem

- Most data sets comprise of measurements on continuous and categorical variables.

The problem

- Most data sets comprise of measurements on continuous and categorical variables.
  - Facebook is collecting data on *media mix*, i.e., factors that may have influence over sales, both continuous and discrete, and is using them to quantify the weight of each factor to create a model to predict marketing results for future strategy.

The problem

- Most data sets comprise of measurements on continuous and categorical variables.
  - Facebook is collecting data on *media mix*, i.e., factors that may have influence over sales, both continuous and discrete, and is using them to quantify the weight of each factor to create a model to predict marketing results for future strategy.
- Little work has been previously done, mostly for the unsupervised problem (*location model* of Olkin and Tate 1961, Gaussian Graphical Model (GGM))

The problem

- Most data sets comprise of measurements on continuous and categorical variables.
  - Facebook is collecting data on *media mix*, i.e., factors that may have influence over sales, both continuous and discrete, and is using them to quantify the weight of each factor to create a model to predict marketing results for future strategy.
- Little work has been previously done, mostly for the unsupervised problem (*location model* of Olkin and Tate 1961, Gaussian Graphical Model (GGM))

Contents

Our approach

Suppose the response Y is either continuous or categorical. We consider the conditional distribution

$$Y \mid (\mathbf{X}, \mathbf{H}), \tag{1}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is a vector of p continuous, and $\mathbf{H} = (H_1, H_2, \ldots, H_q)^T$ is a vector of q binary predictor variables.

- We want to find a lower dimensional function $\mathbf{R} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ of the mixed predictor vector $\mathbf{Z} = (\mathbf{X}^T, \mathbf{H}^T)^T$ such that

$$F(Y \mid \mathbf{Z}) = F(Y \mid \mathbf{R}(\mathbf{Z})),$$

  where $F(\cdot|\cdot)$ denotes the conditional cumulative distribution function of the response given the predictors.
- $\mathbf{R}$ is called a sufficient reduction of the regression of Y on $\mathbf{Z}$.

Contents

A Minimal Sufficient Reduction

for the regression of Y on $(\mathbf{X}, \mathbf{H})$ is given by

$$\mathbf{R}(\mathbf{X}, \mathbf{H}) = \boldsymbol{\alpha}_{\mathbf{b}}^T \left(\mathbf{t}\left(\mathbf{X}, \mathbf{H}\right) - \mathrm{E}\left(\mathbf{t}(\mathbf{X}, \mathbf{H})\right)\right)$$

where

$$\mathbf{t}(\mathbf{X}, \mathbf{H}) \quad = \quad \left(\mathbf{X}^T, \mathbf{H}^T, \left(\mathbf{J}_q \, \mathrm{vech}(\mathbf{H}\mathbf{H}^T)\right)^T\right)^T$$

and $\boldsymbol{\alpha}_{\mathbf{b}}$ is a basis for $\mathcal{S}_{\mathbf{b}} = \mathrm{span}\{\mathbf{b}\}$ with

$$\mathbf{b} \quad = \quad \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix} \quad : \text{to be estimated}$$

What we do

1. We assume a random sample $(y_i, \mathbf{x}_i, \mathbf{h}_i)$, $i = 1, \ldots, n$, is drawn from the joint distribution of $(Y, \mathbf{X}, \mathbf{H})$

What we do

1. We assume a random sample $(y_i, \mathbf{x}_i, \mathbf{h}_i)$, $i = 1, \ldots, n$, is drawn from the joint distribution of $(Y, \mathbf{X}, \mathbf{H})$
2. We derive
   - the MLEs of the parameters $\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \beta, \boldsymbol{\tau}_0, \boldsymbol{\tau}$, and the MLE of the reduction,

What we do

1. We assume a random sample $(y_i, \mathbf{x}_i, \mathbf{h}_i)$, $i = 1, \ldots, n$, is drawn from the joint distribution of $(Y, \mathbf{X}, \mathbf{H})$
2. We derive
   - the MLEs of the parameters $\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \beta, \boldsymbol{\tau}_0, \boldsymbol{\tau}$, and the MLE of the reduction,
   - the asymptotic normality of $\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}$ and asymptotic tests for dimension, and

What we do

1. We assume a random sample $(y_i, \mathbf{x}_i, \mathbf{h}_i)$, $i = 1, \ldots, n$, is drawn from the joint distribution of $(Y, \mathbf{X}, \mathbf{H})$
2. We derive
   - the MLEs of the parameters $\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\tau}_0, \boldsymbol{\tau}$, and the MLE of the reduction,
   - the asymptotic normality of $\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}$ and asymptotic tests for dimension, and
   - a regularized estimator for simultaneous variable (feature) selection and dimension reduction (feature extraction).
3. Details in Bura et al. 2022 and code at https://github.com/lforzani/SDR_mixed_predictions

Contents

Governance index application

We want to build a *Composite Governance index* (CG),

$$\mathbf{R}(\mathbf{Z}) = \boldsymbol{\omega}^T \mathbf{Z} \in \mathbb{R},$$

to predict Y = the logarithm of per capita *Gross Domestic Product* (GDP), measured in 2010 US dollars, over the period 1996 to 2018.

- The World Bank considers the following six aggregate indicators of governance that combine the views of a large number of enterprise, citizen and expert survey respondents:
  - control of corruption (X_1);
  - rule of law (X_2);
  - regulatory quality (X_3);
  - government effectiveness (X_4);
  - political stability (X_5);
  - voice and accountability (X_6).

No country effect

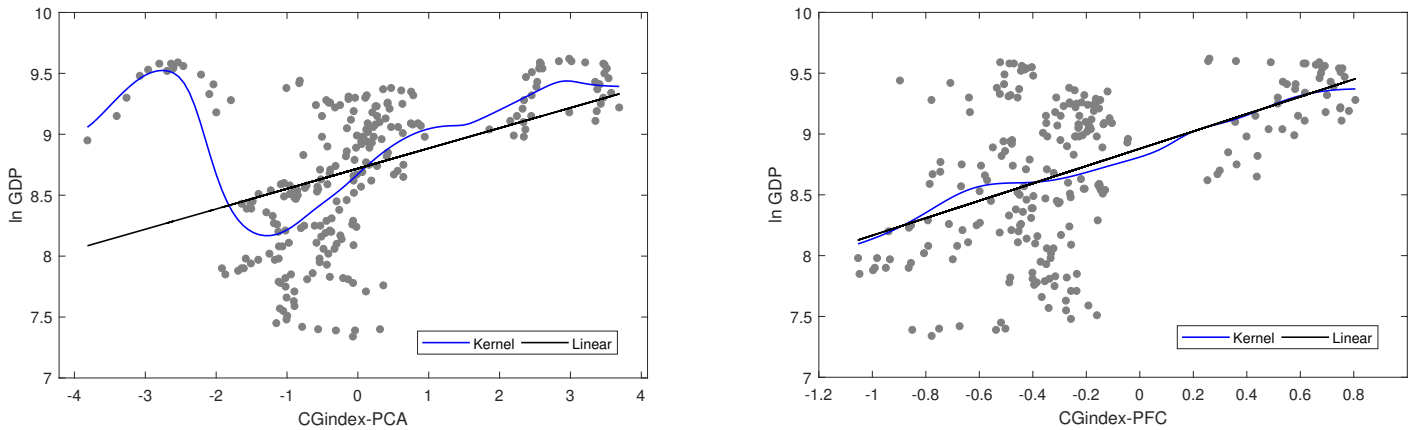

Figure: Log of per capita GDP versus `PCA` and `PFC` (Y-targeted PCA) based composite governance indexes.

Country effect

We add country effect by introducing eleven binary variables **H**.
We plot the log of GDP versus the CG index constructed by PCA for mixed variables (PCAMIX) in the left panel and by our mixed OPTIMAL SDRapproach in the right panel.

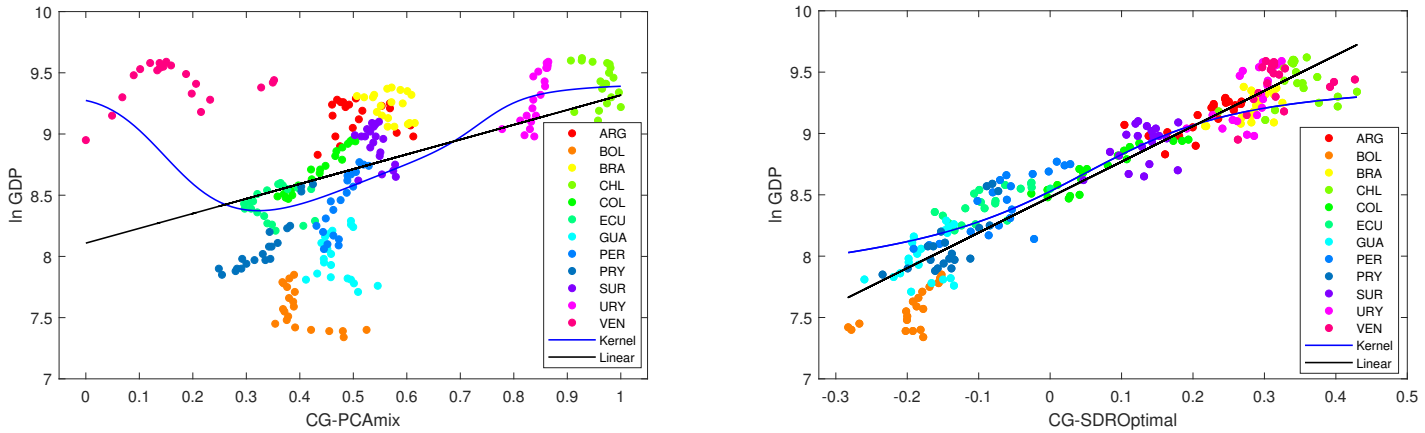

Figure: Log of per capita GDP versus Composite Governance index with country effect.

Contents

References I

Bura, E., L. Forzani, R. G. Arancibia, P. Llop, and D. Tomassi (2022). "Sufficient reductions in regression with mixed predictors". In: *Journal of Machine Learning Research* 23.102, pp. 1–47.

Olkin, I. and R. Tate (1961). "Multivariate Correlation Models with Mixed Discrete and Continuous Variables". In: *Ann. Math. Statist.* 32.2, pp. 448–465.

Thanks!