Clustered Covariate Regression

Abdul-Nasah Soale & Emmanuel Selorm Tsyawo

Temple University

Presenter: Emmanuel Selorm Tsyawo (Job Market Candidate)
estsyawo@temple.edu

Oct. 11, 2019

Introduction

- Identification is necessary for consistency
- A crucial aspect to identification in single-index models is the non-singularity of the Gram matrix $E(\mathbf{x}'\mathbf{x})$
  - Singularity of $E(\mathbf{x}'\mathbf{x}) \implies$ non-identification in e.g. linear and probit regressions
- Singularity of the Gram matrix caused by
  - High dimensionality - many/more covariates relative to observations
  - Multicollinearity
  - Both

Motivation

- Singularities of $E(\mathbf{x}'\mathbf{x})$ in the limit theory engenders
  - inconsistency
  - differing rates of convergence

  (Phillips 2016)

- Estimation impossible without adjustments when non-singularity holds in finite samples

Contribution

A novel approach to identification, inference, and estimation, namely, *Clustered Covariate Regression* (CCR)

- in the presence of rank deficiency, weak identification due to
  - high-dimensionality
  - multicollinearity
  - both

How does the CCR work?

- Project an $n \times p$ rank-deficient $\mathbf{x}$ using a $p \times k$ projection matrix $\boldsymbol{m}$ such that $\boldsymbol{m}'\mathbf{x}'\mathbf{x}\boldsymbol{m}$ is non-singular.
- $p > k$, $n >> k$

Outline

Strands of related literature

- Variable selection, e.g.,
  - the Lasso (Belloni, Chernozhukov, and Hansen (2014b))
  - sequential hypothesis tests (Chudik, Kapetanios, and Pesaran (2018))
- Dimension reduction (using pre-constructed projection matrices) e.g., Wold, Esbensen, and Geladi, 1987
  - principal component regression (PCR)
  - partial least squares (PLS)

What is gained by the CCR?

- The CCR spans the class of single-index models, e.g., linear, logit, and quantile regressions
- CCR obviates sparsity assumption in the high-dimensional parameter $\boldsymbol{\beta}$
  - but assumes (approximate) reducibility of a high-dimensional $\boldsymbol{\beta} \in \mathbb{R}^p$ vector to a smaller identifiable $\boldsymbol{\delta} \in \mathbb{R}^k$, $p > k$
- CCR obviates the choice of tuning parameter and penalty function
  - but requires a criterion (e.g., BIC) for model selection
- The CCR projection matrix $\boldsymbol{m}$ is model, outcome, covariate dependent; it is not pre-constructed
- Immune to multicollinearity via projection

The conditional functional

Suppose a conditional functional (e.g., conditional expectation, conditional quantile)

$$\nu(y_i|\mathbf{x}_i) = g(\mathbf{x}_i\boldsymbol{\beta}) = g(\mathbf{x}_i\boldsymbol{m}\boldsymbol{\delta})$$

- $\boldsymbol{\beta} = \boldsymbol{m}\boldsymbol{\delta}$ (or approximately)
- unknown parameters $\boldsymbol{\beta}$ is $p \times 1$, $\boldsymbol{\delta}$ is $k \times 1$
- $\boldsymbol{m}$ is an unknown $p \times k$ (clustering) projection matrix

The projection matrix I

Characteristics of the projection matrix $\boldsymbol{m}$:

- $\boldsymbol{m}$ belongs to a set $\mathcal{M}$ of $p \times k$ matrices
- has exactly p non-zero elements
- the columns of $\boldsymbol{m}$ correspond to clusters
- each row comprises only one non-zero element
- the vector of non-zero elements in $\boldsymbol{m}$ are researcher-specified, e.g. standard deviations or a vector of ones
- Cluster assignments unknown a priori

The projection matrix II

An example

- $\boldsymbol{\delta} = [\delta_1, \delta_2]'$, $\mathbf{x} = [x_1, x_2, x_3, x_4]$, and $\boldsymbol{m}' = \begin{bmatrix} \eta_1 & 0 & 0 & \eta_4 \\ 0 & \eta_2 & \eta_3 & 0 \end{bmatrix}$
- The linear predictor function
  $\mathbf{x}\boldsymbol{m}\boldsymbol{\delta} = x_1\eta_1\delta_1 + x_2\eta_2\delta_2 + x_3\eta_3\delta_2 + x_4\eta_4\delta_1 =$
  $x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 = \mathbf{x}\boldsymbol{\beta}$
- factorises $\boldsymbol{\beta}$ into $\boldsymbol{m}\boldsymbol{\delta}$
- if scaling, η_j specified by researcher else $\eta_j = 1$ for all $j = 1, \ldots, p$

The CCR estimation problem

- An objective, e.g. linear regression
  $Q_n(\mathbf{x}\boldsymbol{m}\boldsymbol{\delta}) \equiv n^{-1}\sum_{i=1}^{n} q(\mathbf{x}_i\boldsymbol{m}\boldsymbol{\delta})$, $q(\mathbf{x}_i\boldsymbol{m}\boldsymbol{\delta}) = (y_i - \mathbf{x}_i\boldsymbol{m}\boldsymbol{\delta})^2$
- $Q_n(\boldsymbol{\delta}) \equiv \frac{1}{|\mathcal{A}_n|}\int_{\mathcal{M}} Q_n(\mathbf{x}\boldsymbol{m}\boldsymbol{\delta})\mathbf{1}_{\mathcal{A}_n}(\boldsymbol{m})d\mu(\boldsymbol{m}) = \frac{1}{|\mathcal{A}_n|}\sum_{\boldsymbol{m}\in\mathcal{A}_n} Q_n(\mathbf{x}\boldsymbol{m}\boldsymbol{\delta})$
- μ is the counting measure
- set $\mathcal{A}_n \equiv \{\boldsymbol{m} \in \mathcal{M} : Q_n(\mathbf{x}\boldsymbol{m}\boldsymbol{\delta}) \leq Q_n(\mathbf{x}\boldsymbol{s}\boldsymbol{\delta})\ \forall \boldsymbol{s} \in \mathcal{M} \setminus \boldsymbol{m}\}$
- $\mathbf{1}_{\mathcal{A}_n}(\boldsymbol{m}) \equiv \mathbf{1}\{\boldsymbol{m} \in \mathcal{A}_n\}$ indicator for $\boldsymbol{m} \in \mathcal{A}_n$
- $|\mathcal{A}_n| \equiv \mu(\mathcal{A}_n)$ denotes the cardinality of $\mathcal{A}_n$

The Estimation Problem

Minimisation

$\hat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} Q_n(\boldsymbol{\delta})$

- If an $\boldsymbol{m} \in \mathcal{A}_n$ is known, $\boldsymbol{\delta}$ can be estimated
- But, $\boldsymbol{m} \in \mathcal{A}_n$ is unknown, $\implies \boldsymbol{\delta}$ is unknown
- Approach: use a sequential scheme to estimate $\boldsymbol{m} \in \mathcal{A}_n$ and $\boldsymbol{\delta}$

The Sequential CCR Algorithm

fix k (number of clusters)

1. Initialise counter $l = 0$, parameter vector $\boldsymbol{\beta}^{(l)} = \boldsymbol{m}^{(l)}\boldsymbol{\delta}^{(l)}$
2. Update $l \leftarrow l + 1$, for each $j = 1, \ldots, p$,
   - update $\hat{\beta}_j^{(l)} = \arg\min_{\beta_j} Q_n(\mathbf{x}_{-j}\boldsymbol{\beta}_{-j}^{(l)} + x_j\beta_j)$
   - assign $\hat{\beta}_j^{(l)}$ to a cluster and update $\boldsymbol{m}^{(l)}$
   - update $\boldsymbol{\delta}^{(l)} = \arg\min_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} Q_n(\mathbf{x}\boldsymbol{m}^{(l)}\boldsymbol{\delta})$
   - update $\boldsymbol{\beta}^{(l)} \leftarrow \boldsymbol{m}^{(l)}\boldsymbol{\delta}^{(l)}$
3. Check convergence for $Q_n(\mathbf{x}\boldsymbol{m}^{(l-1)}\boldsymbol{\delta}^{(l-1)}) - Q_n(\mathbf{x}\boldsymbol{m}^{(l)}\boldsymbol{\delta}^{(l)}) < \epsilon$ else return to step 2

Without clustering, the algorithm is similar to the (block)-coordinate descent algorithm (used for e.g. Lasso)

Optimal k is determined using a model selection criterion, e.g., BIC

Assumptions

1. $\mu(\mathcal{A}_n) \to 1$ as $n \to \infty$, i.e., $\mathcal{A}_n \to \{\boldsymbol{m}_o\}$ as $n \to \infty$
2. $\sqrt{p/n} \to 0$ as $p, n \to \infty$

Theorem - Consistency

- $\hat{\boldsymbol{\delta}}_n \xrightarrow{p} \boldsymbol{\delta}_o$ under standard assumptions
- $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_o$ where $\hat{\boldsymbol{\beta}}_n \equiv \boldsymbol{m}_n \hat{\boldsymbol{\delta}}_n$ and $\boldsymbol{\beta}_o \equiv \boldsymbol{m}_o \boldsymbol{\delta}_o$

Theorem - Asymptotic Normality

$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_o) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_o)$ where $\mathbf{V}_o = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$,
$\mathbf{A}_o \equiv E[\mathbf{H}(\mathbf{x}_i \boldsymbol{m}_o \boldsymbol{\delta}_o)]$, and $\mathbf{B}_o \equiv E[\mathbf{s}(\mathbf{x}_i \boldsymbol{m}_o \boldsymbol{\delta}_o)\mathbf{s}(\mathbf{x}_i \boldsymbol{m}_o \boldsymbol{\delta}_o)']$

CCR in a baseline specification

Table: Baseline model: $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$, $\boldsymbol{\delta} = [-2, -1, 0, 1, 2, 3]'$

n	p	CCR	OLS	LASSO	PCR	PLS
30	12	0.201(0.086)	0.199(**0.054**)	**0.197**(0.055)	0.395(0.215)	0.325(0.112)
	24	**0.365**(0.136)	0.375(0.145)	0.372(0.140)	0.515(0.173)	0.400(**0.119**)
	36	0.821(0.181)	-	1.400(6.244)	0.769(0.144)	**0.766**(**0.133**)
90	12	**0.069**(0.023)	0.091(**0.021**)	0.091(0.022)	0.357(0.231)	0.188(0.055)
	24	**0.060**(0.023)	0.010(**0.018**)	0.099(0.018)	1.061(0.158)	0.353(0.078)
	36	**0.065**(0.027)	0.110(**0.018**)	0.109(0.018)	1.225(0.117)	0.490(0.083)

Note: (1) Results: average $d_{\beta} = \|\hat{\beta} - \beta_o\|_1 / p$. (2) 1000 simulations each (3) $\sigma(d_{\beta})$ in parentheses. (4) Optimal k - BIC. (5) Two-step Lasso, PCR, and PLS - 10-fold CV (6) $k^* = 6$.

CCR under Multicollinearity

Table: Multicollinearity: $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \mathbf{\Sigma})$, $\mathbf{\Sigma}_{jj'} = 0.5^{|j-j'|}$

n	p	CCR	OLS	LASSO	PCR	PLS
30	12	0.283(0.106)	0.250(**0.075**)	**0.246**(0.076)	0.336(0.143)	0.554(0.126)
	24	0.468(0.168)	0.481(0.194)	**0.448**(0.164)	0.479(0.137)	0.533(**0.118**)
	36	1.118(0.241)	-	1.197(2.346)	**0.714**(0.132)	0.757(**0.118**)
90	12	**0.094**(0.041)	0.116(**0.030**)	0.114(0.031)	0.250(0.127)	0.389(0.067)
	24	**0.076**(0.036)	0.128(**0.026**)	0.126(0.026)	0.300(0.118)	0.285(0.061)
	36	**0.083**(0.043)	0.142(**0.026**)	0.140(0.026)	0.327(0.111)	0.389(0.065)

Note: (1) Results: average $d_\beta = \|\hat{\beta} - \beta_o\|_1/p$. (2) 1000 simulations each (3) $\sigma(d_\beta)$ in parentheses. (4) Optimal k - BIC. (5) Two-step Lasso, PCR, and PLS - 10-fold CV (6) $k^* = 6$.

The empirical model

Estimating private and spillover effects of R&D on productivity

$$E[y_{it}|\boldsymbol{w}_{it}, \boldsymbol{x}_t] = \alpha_0 + \boldsymbol{w}_{it}\boldsymbol{\theta} + x_{it}\gamma_{ii} + \sum_{j \neq i} x_{jt}\gamma_{ij} + \alpha_t + \delta_i$$

- $k_w + T + N + N^2$ parameters, NT observations
- e.g. $T = 27, N = 50, p = 2577$ parameters, $n = NT = 1350$ firm-year observations
- $p > n$

Conclusion

In a nutshell, we propose Clustered Covariate Regression. It is a novel approach to

- handling rank-deficiency in single-index models
  - multicollinearity, high-dimensionality, or both
- offers advantages: e.g. obviates sparsity in β and increases precision
- interesting extensions (left for future work)
  - high-dimensional causal inference
  - multicollinearity in non-linear models
  - estimating latent network structures from panel data