Machine Learning techniques for Panel Data

Federico Nutarelli

IMT Lucca for Advanced Studies

21-11-2019

Introduction

Two papers analysed: (i) ensemble methods; (ii) bias correction in Fixed Effects and Arellano Bond.

1. Main problems: (i) predicting and imputing counterfactual values of outcomes for treated units, had they not received the treatment. (ii) bias in Fixed Effects and Arellano Bond estimates

Introduction

Two papers analysed: (i) ensemble methods; (ii) bias correction in Fixed Effects and Arellano Bond.

1. Main problems: (i) predicting and imputing counterfactual values of outcomes for treated units, had they not received the treatment. (ii) bias in Fixed Effects and Arellano Bond estimates

2. Literature: (i) regression models, synthetic control methods and matrix completion methods. (ii) bias estimation

Introduction

Two papers analysed: (i) ensemble methods; (ii) bias correction in Fixed Effects and Arellano Bond.

1. Main problems: (i) predicting and imputing counterfactual values of outcomes for treated units, had they not received the treatment. (ii) bias in Fixed Effects and Arellano Bond estimates

2. Literature: (i) regression models, synthetic control methods and matrix completion methods. (ii) bias estimation

3. Current work: it is considered an ensemble method and it is shown that it performs better then single methods. (ii) sample splitting methods

Simple example (i)

Suppose $(Y_i, X_{i1}, X_{i2}), \quad i = 1, \dots, N$

Let $\hat{Y}_{1i} = \hat{\beta}_{1,0} + \hat{\beta}_{1,1} X_{i1}$ and $\hat{Y}_{2i} = \hat{\beta}_{2,0} + \hat{\beta}_{2,2} X_{i2}$

Three ways to combine them together:

Nest	Mixture	Ensemble
$\mathbb{E}[Y_i \mid X_{i1}, X_{i2}] =$	$Y_i =$	$\hat{Y}_i =$
$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$	$D_i Y_1 + (1 - D_i) Y_2$	$\theta_0 + \theta_k \hat{Y}_{ki}$

for $k = \{1, 2\}$, $Y_1 = \beta_{1,0} + \beta_{1,1} X_{i1} + \epsilon_{1,i}$, $Y_2 = \beta_{2,0} + \beta_{2,2} X_{i2} + \epsilon_{2,i}$ and D_i having a binomial distribution.

Which combination?

- Nest and mixture model the data generating process (d.g.p.); ensembling deals with fitted values
- Ensemble $\longrightarrow$ *choose* $\theta_0, \theta_1, \theta_2$ to optimize out-of-sample fit. Risk: overfitting
- Mixture. Risk: misspecification. E.g. if $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \sigma^2)$, d.g.p is as in nested. Hence: use nested or ensembling. Avoid mixture

Which combination?

- Nest and mixture model the data generating process (d.g.p.); ensembling deals with fitted values
- Ensemble $\longrightarrow$ *choose* $\theta_0, \theta_1, \theta_2$ to optimize out-of-sample fit. Risk: overfitting
- Mixture. Risk: misspecification. E.g. if $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \sigma^2)$, d.g.p is as in nested. Hence: use nested or ensembling. Avoid mixture

Since all combinations have pros and cons. Why ensembling should be preferred? 1. Computational feasibility 2. building general models that nest simpler ones can be challenging,

Setting

Assume we observe outcomes for N units in T time periods. Only unit N^{th} is exposed to the treatment in T (treatment period).

We observe $Y_{NT}(1)$ and $Y_{it}(0) \quad \forall i \neq N$. To estimate the causal effect $\tau = Y_{NT}(1) - Y_{NT}(0)$ we wish to impute the missing $Y_{NT}(0)$ based on the $NT1$ observations on $Y_{it}(0)$.

Approaches proposed in literature: (i) vertical regression (vr) (ii) horizontal regression (hr) (iii) matrix completion (mc)

Vertical Regression (VR)

Find weights by solving:

$$
{}_{\omega_0,\omega_1}\left\{ \sum_{t=1}^{T-1} \left(Y_{NT} - \omega_0 - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2 \right\}
$$
$$
+\lambda\left(\alpha||\omega||_1 + \frac{1-\alpha}{2}||\omega||_2^2 \right)
$$

The imputed value is then: $Y_{NT}^{vt} = \omega_0 + \sum_{i=1}^{N-1} \omega_i Y_{iT}$. Note that coefficients depends just on i. Penalty: standard elastic net with λ and α selected through cross validation using different time periods.

Horizontal Regression (HZ)

Find weights by solving:

$$_{\beta_0,\beta_1}\left\{\sum_{n=1}^{N-1}\left(Y_{NT}-\beta_0-\sum_{t=1}^{T-1}\beta_t Y_{it}\right)^2\right\} + \lambda\left(\alpha||\omega||_1+\frac{1-\alpha}{2}||\beta||_2^2\right)$$

The imputed value is then: $Y_{NT}^{hz}=\beta_0+\sum_{t=1}^{T-1}\beta_t Y_{Nt}$. Note that coefficients depends just on t. The same as vertical regression after the matrix Y is transposed.

Matrix Completion (MC)

Find the optimal singular value decomposition of L:

$$\min_{L,\alpha,\beta}\left\{\sum_{(i,t)\neq(N,T)}\left(Y_{it}-\alpha_i-\beta_t+L_{it}\right)^2\right\}+\lambda||L||_*$$

where $||L||_*$ is the nuclear norm and L_{it} takes into account the covariates and the error term. By solving the optimization problem, $L_{it}=\sum_{r=1}^{R}A_{ir}B_{ir}$ where $\hat{A}=S\Sigma^{1/2}$, $\hat{B}=R\Sigma^{1/2}$, R is the rank of L determined through the penalization (i.e. elements of L put to 0), Σ diagonal matrix of singular values, S a unitary matrix.

$Y_{NT}^{mc}=L_{NT}+\alpha_N+\beta_T$

Vertical and Horizontal Cross-validations (VC and HC)

VC

First go through units $i =, \ldots, N$. Put aside Y_{iT} for any i and observations for N^{th} unit. Store $\hat{Y}_{it}$ for each method.

HC

Go through the S pre-treatment periods $s = 1 \ldots, S$ (avoid predicting past with future). Estimate $\hat{Y}_{NT-s}$ for each method.

VC estimates the ensemble weights as:

$$\min_{\theta \geq 0} \left\{ Y_{iT} - \theta_{vt} \hat{Y}_{it}^{vt} - \theta_{hz} \hat{Y}_{it}^{hz} - \theta_{mc} \hat{Y}_{it}^{mc} \right\}^2$$

HC estimates ensemble weights as:

$$\min_{\theta \geq 0} \left\{ Y_{iT} - \theta_{vt} \hat{Y}_{it}^{vt} - \theta_{hz} \hat{Y}_{it}^{hz} - \theta_{mc} \hat{Y}_{it}^{mc} \right\}^2$$

Application set up

- Outcomes' definitions: GDP, log GDP, GDP growth rate. Actual values of outcomes available
- $N = 51$ States
- Four different T considered: $T \in \{10, 25, 100, 270\}$ to ensure variability of results: methods with coefficients not depending on time (i.e. HZ) may perform poorly for large T

Application procedure

- Take any i; apply intervention from t to T
- Estimate $Y_{it}(0)$ using all observations $Y_{js}, \quad s \leq t$ with one of the methods
- Compare the estimated Y_{it} with the actual value and square the difference
- Average over the $T - t$ periods of intervention and over the N units. This gives the overall error (first 3 columns of Tab. I)
- Repeat ensembling the estimated Y_{it} through VC or HC to obtain overall error for ensembled methods (Ens-VC and Ens-HC)

Discussion

Ensemble in both his versions gives an overall error which is lower or at most equal to the error given by other methods.

If one of the individual methods does very poorly, the ensemble method takes that into account and puts little weight on that method.

Variation by the choice of cross-validation procedure: with many time periods one can use the time series observations for a particular unit to choose the weights (HC), but with few time periods one needs the cross-section variation and VC performs better.

Discussion

Ensemble in both his versions gives an overall error which is lower or at most equal to the error given by other methods.

If one of the individual methods does very poorly, the ensemble method takes that into account and puts little weight on that method.

Variation by the choice of cross-validation procedure: with many time periods one can use the time series observations for a particular unit to choose the weights (HC), but with few time periods one needs the cross-section variation and VC performs better.

Once imputed missing values, one can perform estimation using: Fixed Effects or Arellano-Bond...

To rely or not to rely? (ii)

Purpose of the Fixed effect approach is to estimate:

$$Y_{it} = D'_{it}\alpha + X'_{it}\gamma + \varepsilon_{it}$$

where $X_{it} := (W'_{it}, Q_i, Q_t)$ and Q_i, Q_t are dummies for time and individual effects.

BIAS: Estimation of N parameters with NT observations. It can be shown that the order of the bias is larger than the one of the stochastic error $\longrightarrow$ cannot rely on significance tests.

To rely or not to rely?

Arellano-Bond eliminates the unit effects a_i by taking differences across time. Specifically it estimates:

$$\Delta Y_{it} = \Delta D'_{it}\alpha + \Delta X_{it}\gamma + \Delta\varepsilon_{it}$$

where $X_{it} := (W'_{it}, Q_t)$ and $\Delta\varepsilon_{it} \perp\!\!\!\perp (D'_{is}, W_{is})_{s=1}^{t-1}$

BIAS: Due to the last independence assumption, A.B. can be rewritten as an overidentified GMM with scores:

$$g(Z_i, \alpha, \gamma) = \{(\Delta Y_{it} - \Delta D'_{it}\alpha - \Delta X'_{it}\gamma)M_{it}\}_{t=2}^{T}$$

where $M_{it} = [(D'_{is}, W_{is})_{s=1}^{t-1} Q_t]$. Since individual effects are eliminated and scores depend on time, with large T too many moment conditions are used to estimate parameters, incurring in a bias.

Formal notation

1. F.E. bias: $p = dim(\gamma)$: dimension of nuisance parameter. $p \to \infty, \quad \text{when} \quad n \to \infty$. $d_\alpha = dim(\alpha)$ is held fixed
2. A.B. bias: $m = dim(g(Z_i, \alpha, \gamma)) \to \infty \quad \text{when} \quad n \to \infty$

Regularity conditions

1. If $(p \wedge m)$ is small w.r.t. n, then

$$(p \wedge m)^2/n \to 0 \quad as \quad n \to \infty$$

2.