Assessing the performance of nuclear norm-based matrix completion methods on CO_2 emissions data

Rodolfo Metulini[1], Francesco Biancalani[2], Giorgio Gnecco[2], Massimo Riccaboni[2]

1. Department of Economics - University of Bergamo
2. Laboratory for the Analysis of CompleX Economics Systems (AXES), IMT School for Advanced Studies Lucca

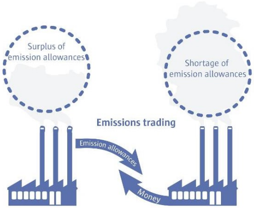
SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation - Università Politecnica delle Marche - Ancona

June 22th, 2023

The framework

Carbon Dioxide (CO_2) emissions represent a rising concern in relation to pollution and climate change (Yoro & Daramola, 2020)

Economic systems produce large amounts of CO_2 by the use of fossil energy. Governments are addressing the production to new systems aimed to minimize emissions.



The European Union (EU) implemented a market of emission rights called the **Emissions Trading System** (ETS) that was launched in 2005, aimed at reducing greenhouse gas emissions.

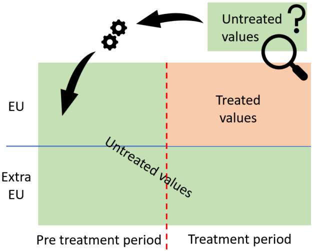
- The idea is to set an annual limit on CO_2 emissions for companies belonging to specific industries.
- Inside this cap, firms are allowed to sell and buy emission rights.

A **counterfactual analysis** for policy evaluation would permits to quantify the reduction of CO_2 emissions due to ETS

The Aim

Due to the ETS policy, untreated CO_2 emissions for the EU countries are unknown in the treatement period.

Matrix Completion (MC) (Hastie et al., 2015) is a supervised statistical learning method to reconstruct a partially incomplete matrix.



We use MC to generate estimates of such untreated CO_2 emissions based on values of the EU countries in the pre-treatment period and on values of extra-EU countries in the treatment period.

To obtain a **robust** counterfactual, we have to study the performance of MC method in reconstructing the original matrix (in absence of treatment).

We develop a simulation study to test the **performance** of Nuclear Norm-based MC methods for panel data.

Nuclear Norm-based MC

Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, MC works by finding a suitable low-rank approximation of $\mathbf{M}$, by assuming the model $\mathbf{M} = \mathbf{C}\mathbf{G}^T + \mathbf{E}$, where $\mathbf{C} \in \mathbb{R}^{m \times r}$, $\mathbf{G} \in \mathbb{R}^{n \times r}$, whereas $\mathbf{E} \in \mathbb{R}^{m \times n}$ is a matrix of errors.

Mazumder (2010) optimization problem - MC Baseline (MCB):

$$\underset{\hat{\mathbf{M}} \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \left(\frac{1}{|\Omega^{\mathrm{tr}}|} \sum_{(i,j) \in \Omega^{\mathrm{tr}}} \left(M_{i,j} - \hat{M}_{i,j} \right)^2 + \lambda \|\hat{\mathbf{M}}\|_* \right)$$

Athey et al. (2021) methodological advancements (MC Fixed Effects - (MCFE) and MC Time Fixed Effects - (MCTFE)) explicitly includes individual and time fixed effects in the optimization problem:

$$\underset{\hat{\mathbf{L}} \in \mathbb{R}^{m \times n}, \hat{\mathbf{\Gamma}} \in \mathbb{R}^{m \times 1}, \hat{\mathbf{\Delta}} \in \mathbb{R}^{n \times 1}}{\text{minimize}} \quad \left(\frac{1}{|\Omega^{\mathrm{tr}}|} \sum_{(i,j) \in \Omega^{\mathrm{tr}}} \left(M_{i,j} - \hat{M}_{i,j} \right)^2 + \lambda \|\hat{\mathbf{L}}\|_* \right)$$

$$\text{subject to} \qquad \hat{\mathbf{M}} = \hat{\mathbf{L}} + \hat{\mathbf{\Gamma}} \mathbf{1}_n^\top + \mathbf{1}_m \hat{\mathbf{\Delta}}^\top$$

$\hat{\mathbf{\Gamma}} \mathbf{1}_n^\top$ and $\mathbf{1}_m \hat{\mathbf{\Delta}}^\top$ model row (individual) and column (time) fixed effects

Differently from MCB the nuclear norm $\|\hat{\mathbf{L}}\|_*$ is used instead of $\|\hat{\mathbf{M}}\|_*$.

Design of experiment

Freely available database on total CO_2 emissions (in thousand of tons) by country and sector (Corsatea et al, 2019 – `https://joint-research-centre.ec.europa.eu/document/download/b572c87b-a2fb-4ab6-af38-ff0451273e9e_en?filenameco2em56.zip`), covering years 2000 – 2016 and 42 countries (29 European and 13 non-European).

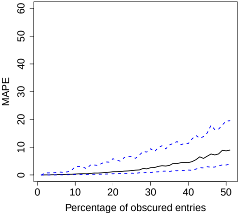
Years: from 2000 to 2005, in order to avoid possible treatment effects coming from the ETS. **Countries**: 26 (14 EU, 12 extra-EU, we dropped small and extra-EU countries having special agreements with the EU)

We compare the performance of `MCB`, `MCTFE` and `MCFE`, with respect to the **original matrix** (`raw`) and to a suitably **pre-processed matrix** (l_1 row-normalization by country), using the Mean Absolute Percentage Error (**MAPE**).

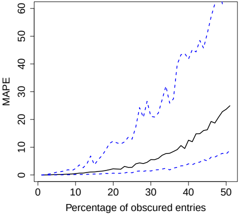
For any specific percentage of **unknown entries** (**from 0 to 50%**, at intervals of 1), **200 replications** have been generated, where the unknown entries (test set) are chosen at random according to the desired percentage.

Computations performed with `mcnnm_cv` function in `MCPanel` R package.

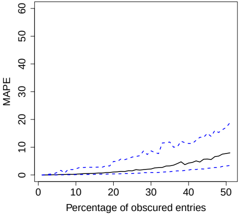
Results - MAPE



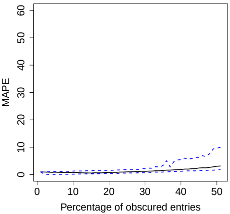
(a) MCB



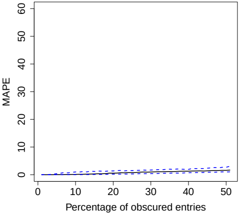
(b) MCTFE



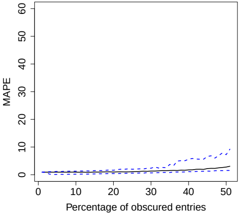
(c) MCFE



(d) MCB_l_1

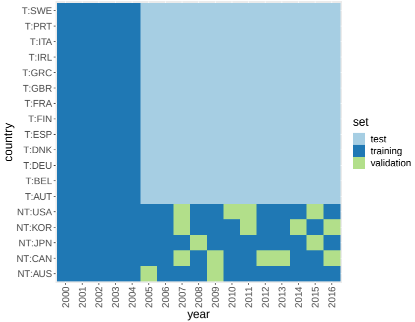


(e) MCTFE_l_1



(f) MCFE_l_1

Counterfactual Strategy



Training set: Values of the pre-treatment period + 75% of randomly selected extra-EU countries values in the treatment period. **Validation set**: remaining 25% of extra-EU countries values in 2005–2016. **Test set**: values of EU countries in the treatment period (2005–2016) (around 50% of missing entries).

`MCFE` on by country l_1 row-normalized values is applied to estimate the counterfactual CO_2 emissions on the test set.

To draw best and worst case scenario, we represent, for each treated country, **10^{th}, 50^{th} and 90^{th} percentiles** from **80 replications** with randomly selected different training and validation sets.

Counterfactual Results

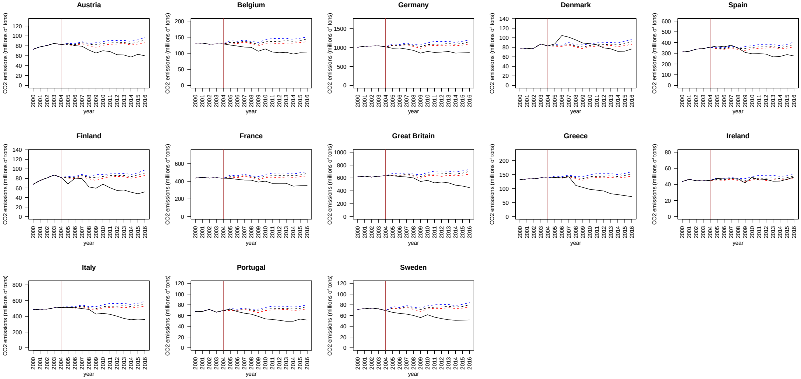


Figure: Total CO_2 emissions of treated countries. Actual values (black lines) compared to counterfactual values calculated by MCFE (test set). Medians (black dashed lines), 10^{th} percentiles (red dashed lines), and 90^{th} percentiles (blue dashed lines) considering the 80 MCFE random simulations. Vertical red lines divide the period into pre-treatment and treatment.

Discussion

In previous works of us we developed MC strategies to:

1. Impute missing entries in World Input/Output tables

   → Metulini, R., Gnecco, G., Biancalani, F., & Riccaboni, M.: Hierarchical clustering and matrix completion for the reconstruction of world input-output tables. AStA Advances in Statistical Analysis, 1-46 (2022)

2. Predict CO_2 emissions at sector-year-country level

   → Biancalani, F., Gnecco, G., Metulini, R., Riccaboni, M. (2023). Matrix Completion for the Prediction of Yearly Country and Industry-Level CO2 Emissions. In "Machine Learning, Optimization, and Data Science". LOD 2022. Lecture Notes in Computer Science, vol 13810. Springer, Cham.

In this work:

- We assessed the performance of different versions of nuclear norm-based MC in imputing missing CO_2 emissions. → MCFE and MCTFE performs well (even for large amounts of missing entries) when applied to row-normalized matrices.
- With a robust counterfactual analysis, we are able to quantify the amount of CO_2 emissions saved due to the ETS is in place.

References

1. Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K.: Matrix completion methods for causal panel data models. *Journal of the American Statistical Association* **116(536)**, 1716-1730 (2021)
2. Corsatea T.D., Lindner S., Arto, I., Roman, M.V., Rueda-Cantuche J.M., Velazquez Afonso A., Amores A.F., Neuwahl F.: World Input-Output Database Environmental Accounts. Update 2000-2016, EUR 29727 EN, *Publications Office of the European Union*, Luxembourg, (2019)
3. Hastie T, Tibshirani R, Wainwright M, Statistical Learning with Sparsity: The Lasso and its Generalizations. *CRC Press*, New York (2015).
4. Mazumder R, Hastie T, Tibshirani R,: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research* **11**, 2287-2322 (2010)
5. Yoro, K. O., & Daramola, M. O., CO2 emission sources, greenhouse gases, and the global warming effect. In: *Advances in carbon capture*, pp. 3-28. Woodhead Publishing (2020).