FINAL PROJECT: SUMMARIZING PAPERS ON PANEL DATA MODELS

TO MINH ANH

2024-04-28

Contents

I. Shen et al. (2022), Same root different leaves: Time Series and Cross-sectional Methods in Panel Data

The paper examines panel data analysis, which observes the behavior of multiple units over time to evaluate the causal effect of a treatment. It discusses two widely used methodologies: horizontal regression (i.e., unconfoundedness), which employs time series patterns, and vertical regression (e.g., synthetic controls), which utilizes cross-sectional data. These methods are typically viewed as fundamentally different; however, the authors challenge this conventional belief by showing that under certain standard settings, these two approaches can yield identical point estimates. Notably, when the point estimates are the same, each approach quantifies uncertainty with respect to a distinct estimand, where the confidence interval developed for one estimand can have incorrect coverage for another. This underscores how assumptions about the source of randomness critically influence inference accuracy.

1. Introduction

The paper begins with the compelling example of estimating the economic impact of terrorism in the Basque Country using panel data analysis, referencing the foundational work of Abadie and Gardeazabal (2003). Their seminal study quantitatively isolated the economic effects of terrorism using synthetic controls, a method that creates a synthetic Basque Country from a weighted combination of control regions unaffected by instability to estimate the region's economic evolution in the absence of terrorism. The authors aim to further this methodological conversation by exploring whether horizontal (HZ) and vertical (VT) regressions, commonly used in economic evaluations, produce identical estimates and how their differing assumptions about randomness affect inference. They address these questions by classifying several widely studied regression formulations into symmetric classes that yield identical point estimates and asymmetric classes that do not. Within the symmetric class, they study the properties of the estimators with randomness stemming from different sources and examine the issues from both model-based and design-based perspectives. The paper contributes to the understanding of these regression models by clarifying when and why they might offer similar point estimates and the impact of their differences in handling randomness on economic inference.

2. The Panel Data Framework

The paper sets the stage by introducing a panel data framework with N units over T time periods and establishes the necessary notation for subsequent discussions. It then provides a brief comparison of two bodies of literature, which are explored in more detail in the second paper.

Specifically, the unconfoundedness literature operates on the concept that "history is a guide to the future." Accordingly, unconfoundedness methods model outcomes in the treated period as a weighted composition of outcomes from pretreatment periods. This modeling is accomplished by regressing the control units' outcomes in the treated period on their lagged outcomes and applying the learned regression coefficients to the treated unit's lagged outcomes to predict the missing outcome.

In the Synthetic Controls literature, which is based on the principle that "similar units behave similarly," the treated unit's outcomes are expressed as a weighted composition of the control units' outcomes. This is done by regressing the treated unit's lagged outcomes on the control units' lagged outcomes and applying the learned regression coefficients to the control units' outcomes in the treated period to predict the missing outcome.

The paper then acknowledges the viewpoint from Athey et al. (2017), as discussed in the second paper, that despite the perceived asymmetry between HZ and VT suggesting fundamental differences, both can be applied to the same setting with appropriate regularization.

3. Point Estimation

In this section, the authors aim to answer the question: "When are HZ and VT point estimates identical?" They start by outlining various widely studied regression formulations from both the HZ and VT literature, including Penalized Regression—Ordinary Least Squares (OLS), Principal Component Regression (PCR), Ridge and Lasso Regression, Elastic Net Regression; as well as Constrained Regression—Simplex regression. To answer the question, they categorize these regression formulations into symmetric classes (where HZ and VT point estimates agree) and asymmetric classes (where HZ and VT disagree). Their central findings are summarized in two theorems:

Theorem 1: HZ is equal to VT for i) OLS, ii) PCR, and iii) ridge regression under some specific conditions.

Theorem 2: HZ is not equal to VT for i) lasso, ii) elastic net, and iii) simplex regression.

Firstly, Theorem 1 not only aligns with the results from previous articles on the synthetic control method, which demonstrated that HZ OLS and HZ ridge share the same linear forms with the corresponding VT estimates, but it also establishes numerical equivalence, suggesting a deeper dual relationship between the two perspectives. Secondly, the authors emphasize that Theorem 1 holds for any data configuration, suggesting that HZ and VT for OLS are valid when $T > N$ and $N > T$, contrary to prior belief. Although, as seen later in the second paper, the unconfoundedness approach does not perform well when $T >> N$, and the synthetic control approach struggles when $N >> T$, compared to the newly proposed estimator. Thirdly, they conjecture that the geometry of the ℓ_2-ball in the OLS, PCR, and ridge regression models is likely responsible for the HZ and VT estimation symmetry. Conversely, they infer from Theorem 2 that the common thread between the objective functions in the asymmetric class is a penalty or constraint promoting sparse models. Hence, the geometries of the ℓ_1-ball and the simplex, which encourage sparsity, are likely sources of HZ and VT estimation asymmetry.

4. Inference

The authors proceed to address the question: "When HZ and VT point estimates are identical, how does the source of randomness impact inference?" The article considers both a (i) model-based approach, assuming randomness arises from the distribution of potential outcomes, and a (ii) design-based approach, attributing the source of randomness to the treatment assignment mechanism.

Within **the model-based framework**, they consider a classical regression model. They start by specifying the assumptions for the horizontal, vertical and mixed models. Essentially, the HZ models assumes a time series correlation pattern where we can predict future potential outcomes Y_{iT} from past outcomes $\mathbf{Y}_0$, whereas the VT model assumes a cross-sectional dependency where we can predict the potential outcomes Y_{iT} based on other different units at the same point in time. Moreover, the mixed model considers both time series and cross-sectional correlations while maintaining constraints on the distribution of errors and their independence. The error terms ϵ_{iT} for those models have a 0 mean and are independent across time for each unit (HZ), across units (VT) or across time and units (mixed model). Upon establishing these assumptions, the authors delve into statistical inference, which is captured in **Theorem 3**. Each part of this theorem establishes that, given the model's assumptions and suitable moment conditions, the normalized in-sample prediction of the potential outcome will converge in distribution to the standard normal distribution. This is important for conducting hypothesis tests and creating confidence intervals.

In determining the confidence interval, the authors discuss estimating covariance matrices and the behavior of these estimates under conditions of homoskedastic and heteroskedastic errors within HZ, VT, and mixed models. They present three lemmas that distinguish between the two types of error variances and illustrate the performance of estimators under each scenario, examining the unbiasedness and conservativeness of these estimators. Homoskedasticity, with its assumption of constant variance, simplifies the estimation, while heteroskedasticity, which assumes varying variances, necessitates more sophisticated methods like jackknife or heteroskedasticity-robust (HR) estimators.

Within **the design-based framework**, the authors consider the assumptions that the treatment period is randomly selected from all possible time periods and that treatment is randomly assigned to units within the population. Here, the authors propose a design-based estimator that considers all possible treatment assignments and predicts the mean counterfactual outcome under control for unit i at time t, denoted as $Y_{it}(0)$. This estimator is the average outcome for all units and times where the unit did not receive treatment. Although the model-based and design-based estimators arrive at the same point estimate for the realized assignment, they diverge in their formulations and the source of randomness they attribute.

Regarding **inference properties**, the article presents a table comparing the estimands associated with the model-based and design-based estimators under three sources of randomness: (i) time, (ii) unit, and (iii) both time and unit. There are two key takeaways: (i) both model-based and design-based estimators yield similar estimands for each source of randomness, and (ii) different sources of randomness result in distinct estimands.

The article also offers several insightful remarks, one of which underscores the practical importance of identifying the source of randomness in the data. For instance, in case studies like the Basque Country analysis, it might be more reasonable to ascribe randomness to temporal factors rather than spatial, given the possibility that terrorism might have occurred in a different year rather than a different region. When the source of randomness is difficult to ascertain, however, the authors suggest a structured framework to rigorously compare the predictive qualities of various estimators under review.

5. Illustration and Conclusion

This section outlines the application of statistical models to panel data across three case studies: the impact of terrorism in the Basque Country (Abadie and Gardeazabal (2003)), the effects of California's Proposition 99 on tobacco consumption (Abadie et al. (2010)), and the economic repercussions of German reunification on West Germany (Abadie et al. (2015)). Through 500 replications of a data generating process, the analysis assesses the stability and coverage of 95% confidence intervals across various models (Figure 1). Their formal results and simulations reveal two critical insights: the accuracy of inferences is directly influenced by the choice of estimand and the variance estimation developed for one estimand may not have the correct coverage for another. The section concludes by challenging standard views on regression analyses for panel data and urges careful consideration of the source of randomness to guide the selection of estimands and inferential methods.

II. Athey et al. (2017), Matrix Completion Methods for Causal Panel Data Models

1. Introduction

In the domain of panel data analysis, the paper introduces a novel approach for imputing missing potential control outcomes to estimate the average treatment effect on treated units. The authors propose a matrix completion estimator that effectively approximates the original incomplete matrix with reduced complexity. They offer new insights into how the literatures on matrix completion, interactive fixed effects models, and program evaluation using unconfoundedness (HZ) and synthetic control (VT) methods are interconnected. Their contributions are in three ways: (i) presenting formal results from the generalized matrix completion literature where the missing data patterns allow for time-based correlation, (ii) demonstrating that their approach, along with synthetic control and unconfoundedness, can be considered matrix completion methods with a shared objective function but differing in identification—either through regularization or hard restrictions, and (iii) evidencing through simulations that their proposed matrix completion with nuclear norm minimization estimator (MC-NNM) generally outperforms traditional HZ and VT estimators.

2. Set up

The paper lays the groundwork for addressing causal questions in panel data, observing N units across T periods. It outlines the necessary assumptions, defines the matrix of realized outcomes $\mathbf{Y}$, treatment assignment $\mathbf{W}$, and observed covariate matrices $\mathbf{X}$ (unit-specific covariate columns) and $\mathbf{Z}$ (time-specific covariates). The section also discusses notations connected to the matrix completion literature, which are used to impute missing entries in the $\mathbf{Y}_0$ matrix for treated units. This is crucial for estimating the average effect for the treated or the average treatment effects.

3. Patterns of Missing Data, Thin and Fat Matrices, and Horizontal and Vertical Regression

The authors discuss several configurations of the matrices $\mathbf{Y}$ and $\mathbf{W}$ that are the focal points in different segments of the general literature. This discussion aims to contextualize the problem and motivate previously developed methods from the literature on causal inference under unconfoundedness, synthetic control, and interactive fixed effect literature, ultimately drawing formal connections between these areas and matrix completion literature. Note that the literature has focused primarily on the case where $\mathbf{W}$ is completely random, which may not be applicable in social science applications. Therefore, they move to discuss non-random patterns of missing data, such as block structure, where a subset of units consistently adopts a treatment at a specific time, resulting in a block pattern of missing matrix values. This pattern is consistent with methods used for single-treated period and unit analyses in both unconfoundedness and synthetic control literature. Another pattern is staggered adoption, where units adopt the treatment at different times, and once adopted, the treatment is irreversible, often seen when new technology is adopted. The section also considers how the shape of matrix Y—whether thin ($N >> T$), fat ($N << T$), or approximately square ($N \approx T$)—influences the analysis approach and the need for regularization.

The paper continues by detailing two specific combinations of missing data patterns and matrix shapes that have been thoroughly explored. The Horizontal Regression and Unconfoundedness Literature primarily focuses on the single-treated-period block structure within a thin matrix. This approach involves a considerable number of treated and control units and imputes missing potential outcomes for the last period using control units with comparable lagged outcomes. On the other hand, the Vertical Regression and Synthetic Control Literature employs a method suited for a single-treated-unit block structure with a relatively fat or square matrix. It entails regressing outcomes for the treated unit before treatment against outcomes for control units during the same periods.

Finally, the paper discusses fixed effects and factor models, methods that capture stable patterns over time and across units, as exemplified by two-way fixed effects and interactive factor models. Previous research often involved scenarios where the number of units N greatly exceeded the number of time periods T, with N increasing while T remained constant. In contrast, current literature permits both N and T to increase, which allows for the consistent estimation of loadings $\mathbf{U}$ and factors $\mathbf{V}$, with the number of factors R typically fixed. The estimation of the rank R is critical in these models. Although there have been an attempt at an interactive fixed effects model with blocked assignment, which is computationally simpler, it is unsuitable for complex patterns of missing data. Consequently, the paper refers to machine learning and statistical literature that focuses on matrix completion. The goal here is not to directly estimate $\mathbf{U}$ and $\mathbf{V}$ but to impute missing elements in matrix $\mathbf{Y}$. This is achieved using low-rank matrix models based on regularization techniques like nuclear norm regularization, providing an alternative approach to handling missing data in complex scenarios.

4. The matrix completion with Nuclear Norm Minimization Estimator

This section presents the paper's first major contribution, discussing a matrix completion method to estimate a complete outcomes data matrix $\mathbf{Y}$ using nuclear norm minimization. It models $\mathbf{Y}$ as the sum of a low-rank matrix $\mathbf{L}^*$, which absorbs fixed effects, and an independent error term . The expectation of $\mathbf{L}^*$ is zero, indicative of measurement error. The error is assumed to be independent of $\mathbf{L}^*$, with its elements being sub-Gaussian and independent.

The paper then focuses on estimating the low-rank matrix $\mathbf{L}^*$ and notes that directly minimizing the sum of squared differences between observed and estimated matrices is ineffective due to the missing values issue. To resolve this, they add a nuclear norm penalty term to the objective function for regularization. However, they do not wish to regularize the fixed effects, aiming to improve imputation quality, so they explicitly propose variables to estimate these effects. Consequently, they introduce a formal estimator for $\mathbf{L}^*$ (subsequently named MC-NNM), which minimizes the sum of squared residuals and the nuclear norm penalty term. Unlike traditional methods for $\mathbf{L}^*$ discussed in previous papers, the authors incorporate fixed effect estimates, which are notably accurate due to the high proportion of observed values in matrix completion literature. This method involves iteratively updating the estimates of $\mathbf{L}^*$ and the fixed effects with a shrinkage operator until convergence, employing a fast convex optimization algorithm. Cross-validation is utilized to select the optimal regularization parameter. Additionally, the section briefly considers confidence intervals for the asymptotic distribution of $\mathbf{L}^*$ and outlines methods for their construction, highlighting the importance of re-sampling methods for statistical evaluations.

5. The relationship with Horizontal and Vertical Regressions

The authors demonstrate that matrix completion, horizontal regression, vertical regression, synthetic control regression, the elastic net version, and difference-in-differences estimators are all closely related. Specifically, they can all be expressed as minimizing the same objective function $Q(Y, R, A, B, \lambda, \delta)$ under different restrictions or with different approaches to regularization of the unknown parameters R, A, B, λ, δ. **Theorem 1** outlines these differences in hard restrictions and regularization approaches as follows: nuclear norm matrix completion involves direct minimization of the nuclear norm subject to matrix rank constraints; horizontal regression is applicable when the number of units exceeds the number of time periods, focusing on the last period's outcome and regressing it on previous outcomes to impute missing data; vertical regression is used when the number of time periods exceeds the number of units, regressing outcomes across different units; the elastic net is a version of vertical regression that applies regularization to the regression coefficients; and difference-in-differences regression is modeled with no rank constraint, effectively setting the rank to zero.

The paper emphasizes the need to add structure to the optimization problem and discusses how each method's approach to regularization and the restrictions imposed on the parameters affect the estimation of missing outcomes. It also underscores the importance of cross-validation for comparing approaches that are difficult to test against each other directly.

6. Main results

The section begins with additional notations and definitions necessary for the theoretical discussion. It mentions an observation process that defines the set of observed entries and assumes the randomness in this process is independent of the low-rank matrix $\mathbf{L}^*$. However, it acknowledges that the randomness can still be a function of $\mathbf{L}^*$. Subsequently, the paper presents a significant theoretical result — the estimation of an upper bound for the root-mean-squared-error (RMSE) of the estimator $\hat{\mathbf{L}}$. Under certain assumptions, including that the rank of $\mathbf{L}^*$ is fixed and the penalty parameter λ is properly chosen, the RMSE is bounded by a function of the number of units N, time periods T, and the parameters controlling the complexity of $\mathbf{L}^*$. This bound decreases as N and T increase, indicating better estimation accuracy with larger data sets.

The interpretation of the theorem clarifies that when $\mathbf{L}^*$ is low-rank and the lower bound for the average number of control units (indexed by a parameter p_c) grows quickly enough, the RMSE converges to zero, suggesting that the estimator becomes perfect as the amount of data increases. The paper also contrasts the result with existing matrix-completion literature, pointing out that their estimator, which allows for time series dependency structures in the missing data, requires more observations to achieve a consistent estimation than what is suggested in previous research.

7. Illustrations, Generalizations and Conclusion

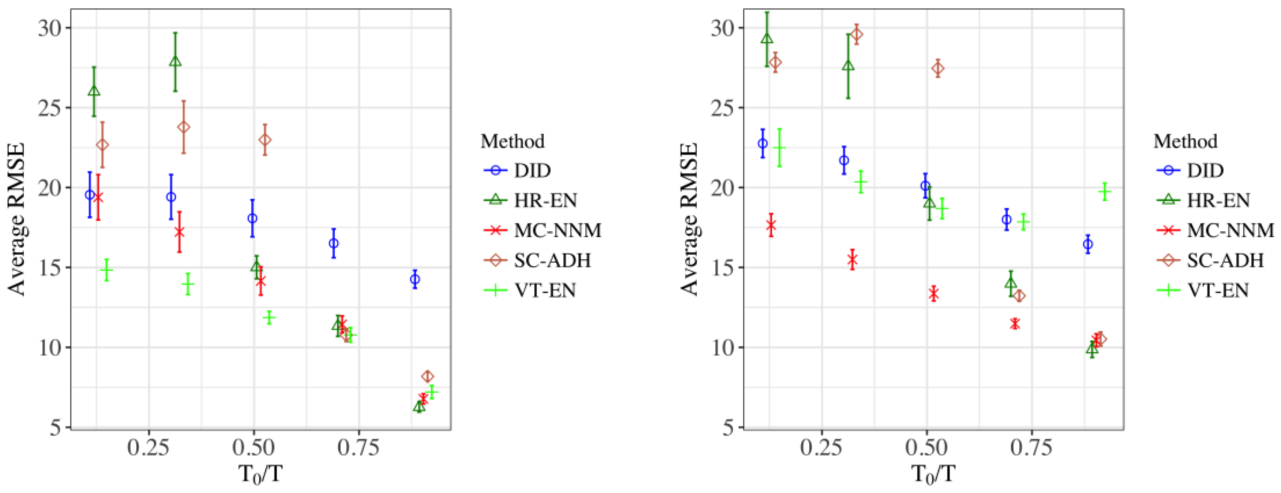
The paper assesses the MC-NNM estimator against traditional methods using two datasets. The first uses California Smoking Data (Abadie et al. (2010)) to predict outcomes under simultaneous and staggered treatment adoption scenarios, showing MC-NNM's superior predictive ability (Figure 2). The second analysis examines daily stock returns, with MC-NNM outperforming HR-EN and VT-EN, especially when matrix dimensions are nearly square, highlighting its adaptability and effectiveness in various data settings (Figure 3). Following a concise discussion on the broader applicability of the proposed estimator, the article concludes by summarizing its main findings and examples, as outlined in the introduction.

Appendix - Additional Figures

Coverage for nominal 95% confidence intervals across 500 replications. The coverage length is normalized by the magnitude of the corresponding point estimate.

Case study	$\widehat{v}_0^{\mathrm{hz}}$			$\widehat{v}_0^{\mathrm{vt}}$			$\widehat{v}_0^{\mathrm{mix}}$		
	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}
Basque (CP)	0.92	0.74	0.63	0.99	0.93	0.88	1.00	0.97	0.94
Basque (AL)	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04
California (CP)	0.95	1.00	0.92	0.64	0.93	0.60	0.98	1.00	0.95
California (AL)	0.07	0.07	0.07	0.03	0.03	0.03	0.08	0.08	0.08
W. Germany (CP)	0.94	1.00	0.93	0.49	0.94	0.49	0.96	1.00	0.95
W. Germany (AL)	0.03	0.03	0.03	0.01	0.01	0.01	0.03	0.03	0.03

Figure 1: Simulation Results



(a) Simultaneous adoption, $N_t = 8$

(b) Staggered adoption, $N_t = 35$

Figure 2: California Smoking Data

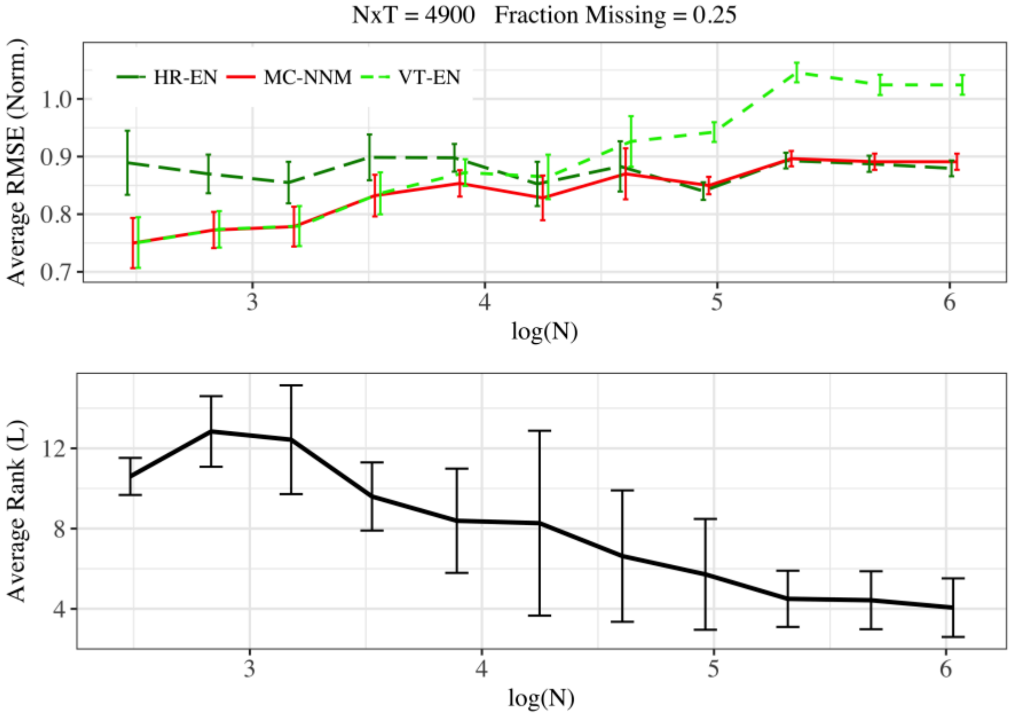


Figure 3: Stock Market Data

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

Abadie, A., Diamond, A., and Hainmueller, J. (2015), "Comparative Politics and the Synthetic Control Method: COMPARATIVE POLITICS AND THE SYNTHETIC CONTROL METHOD," *American Journal of Political Science*, 59, 495–510. https://doi.org/10.1111/ajps.12116.

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. https://doi.org/10.1257/000282803321455188.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017), "Matrix Completion Methods for Causal Panel Data Models." https://doi.org/10.48550/ARXIV.1710.10251.

Shen, D., Ding, P., Sekhon, J., and Yu, B. (2022), "Same Root Different Leaves: Time Series and Cross-Sectional Methods in Panel Data." https://doi.org/10.48550/ARXIV.2207.14481.

STA 640 FINAL PROJECT: SUMMARIZING PAPERS ON PANEL DATA MODELS

Minh Anh To

Department of Statistical Science
Duke University

Watch the project video here.

Introduction

I. Shen et al. 2022, Same root different leaves : Time Series and Cross-sectional Methods in Panel Data
II. Athey et al. 2017, Matrix Completion Methods for Causal Panel Data Models

Time Series and Cross-sectional Methods in Panel Data

- Examines Panel Data Analysis Methodologies :
  - Horizontal Regression (i.e., unconfoundedness) : Adopts time series data patterns.
  - Vertical Regression (e.g., synthetic control) : Uses cross-sectional data patterns.
- Main Contributions :
  - Challenges the conventional belief that these two methods are fundamentally different by proving that both approaches yield identical point estimates under several standard settings.
  - Shows that in the case where the point estimates are the same, each approach quantifies uncertainty with respect to a distinct estimand, where the confidence interval developed for one estimand can have incorrect coverage for another.
  - Underscores how assumptions about the source of randomness critically influence inference accuracy.

Matrix Completion Methods for Causal Panel Data Models

- In the domain of panel data analysis, the paper introduces a novel approach for imputing missing potential control outcomes to estimate the average treatment effect on treated units.
- Main contributions :
  - Presents formal results from the generalized matrix completion literature where the missing data patterns allow for time-based correlation. Specifically, proposes a matrix completion estimator that effectively approximates the original incomplete matrix with reduced complexity (MC-NNM).
  - Demonstrates that their approach, along with synthetic control and unconfoundedness, can be considered matrix completion methods with a shared objective function but differing in identification—either through regularization or hard restrictions.
  - Evidences through simulations that their proposed matrix completion with nuclear norm minimization estimator (MC-NNM) generally outperforms traditional HZ and VT estimators.

References I

Athey, Susan et al. (2017). "Matrix Completion Methods for Causal Panel Data Models". In : doi : 10.48550/ARXIV.1710.10251. url : https://arxiv.org/abs/1710.10251 (visité le 27/04/2024).

Shen, Dennis et al. (2022). "Same Root Different Leaves : Time Series and Cross-Sectional Methods in Panel Data". In : doi : 10.48550/ARXIV.2207.14481. url : https://arxiv.org/abs/2207.14481 (visité le 27/04/2024).