Heterogeneous Treatment Effects in Panel Data

Retsef Levi
Sloan School of Management
Massachusetts Institute of Technology
retsef@mit.edu

Elisabeth Paulson
Harvard Business School
epaulson@hbs.edu

Georgia Perakis
Sloan School of Management
Massachusetts Institute of Technology
georgiap@mit.edu

Emily Zhang
Operations Research Center
Massachusetts Institute of Technology
eyzhang@mit.edu

Abstract

We address a core problem in causal inference: estimating heterogeneous treatment effects (HTEs) using panel data with general treatment patterns. Many existing methods either do not utilize the potential underlying structure in panel data or have limitations in the allowable treatment patterns. In this work, we propose and evaluate a new method that first partitions observations into disjoint clusters with similar treatment effects using a regression tree, and then leverages the underlying structure of the panel data to estimate the average treatment effect (ATE) for each cluster. Computation experiments with semi-synthetic data show that our method achieves superior accuracy for ATE and HTE estimation compared to alternative approaches. This performance was achieved using a regression tree with no more than 40 leaves, making the method both accurate and interpretable, and a strong candidate for practical applications.

1 Introduction

Panel data arise when outcomes are observed across n units over T periods. Each unit may have been subject to an intervention during certain time periods that influences the outcome. As a concrete example, a unit could be a geographic region affected by a new economic policy, or an individual consumer or store influenced by a marketing promotion. Our goal is to understand the impact of this intervention on the outcome. Since the effect of the intervention might vary across individual units and time periods as a function of the unit-level and time-varying covariates, we aim to estimate the heterogeneous treatment effects (HTEs). This is a key problem in econometrics and causal inference, enabling policymakers or business owners to make more informed decisions about which units to target for future interventions.

Existing methods face limitations. Synthetic control and matrix completion approaches only address average effects. In contrast, machine learning methods (e.g., causal forests (Wager and Athey, 2018), meta-learners (Künzel et al., 2019)) estimate heterogeneous treatment effects (HTEs) but assume i.i.d. data, neglecting the panel structure, which reduces their accuracy. We introduce PaCE, a method that accommodates general treatment patterns while incorporating panel structure, achieving both accurate and interpretable HTE estimates.

Notation. For any matrix A, $\|A\|_{\mathrm{F}}$ denotes the Frobenius norm, and $\|A\|_{\star}$ denotes the nuclear norm. We use $\circ$ to denote element-wise matrix multiplication.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Model and algorithm

PaCE is designed for panel data, where an outcome of interest is observed across n distinct units for T time periods. Let $O \in \mathbb{R}^{n\times T}$ be the matrix of these observed outcomes, where each row represents a unit and each column a time period. Our objective is to discern how these outcomes were influenced by various treatments under consideration.

We assume that, in a hypothetical scenario where the treatments were not applied, the expected outcomes exhibit some underlying structure, such as seasonality or patterns of variation across units. Our method leverages this panel structure, allowing it to outperform methods that ignore such structure and treat all observations as i.i.d. Specifically, we model the matrix of untreated outcome as an unknown low-rank matrix $M^\star \in \mathbb{R}^{n\times T}$, which can capture such structured variations.

To incorporate the influence of the external factors, we introduce the covariate tensor $\mathbf{X} := [X_1, \dots, X_p] \in \mathbb{R}^{p\times n\times T}$, where an element X_{izt} signifies the i-th covariate of unit z at time t. For convenience, we will write $X^{zt} := [X_{1zt}, \dots, X_{pzt}]$ as the vector of relevant covariates for unit z at time t. We consider q distinct treatments, each represented by a binary matrix $W_1, \dots, W_q$ that encodes which observations are subjected to that treatment. For each treatment $i \in [q]$ applied to unit z at time period t, the treatment effect is modeled as a non-parametric Lipschitz function of the covariates $\mathcal{T}_i^\star(X^{zt})$. We consider the treatment effects of distinct treatments to be additive.

Combining all of these elements, the observed outcome matrix O can be expressed as the sum of $M^\star$, the combined effects of the treatments, and a noise matrix E. Mathematically,

$$O = M^\star + \sum_{i=1}^{q} \mathcal{T}_i^\star(\mathbf{X}) \circ W_i + E,$$

where $\mathcal{T}_i^\star(\mathbf{X})$ denotes the matrix where the element in z-th row and t-th column is $\mathcal{T}_i^\star(X^{zt})$.

2.1 Algorithm

Our goal is to estimate the treatment effect function $\mathcal{T}_i^\star$ for each $i \in [q]$ using a regression tree, having observed O, $\mathbf{X}$, and $W_1, \dots, W_q$.

The key challenge is that the true treatment effects are unknown, so the splitting criterion must be approximated. The way that we propose to do this is specialized for panel data. We leverage global structural information from the entire panel of data to construct the trees. By utilizing this underlying structure, PaCE—using just one tree for each of the q treatments—is able to outperform alternative methods on most of our experiment instances.

The regression tree approximations are obtained by solving a min–min optimization problem: the outer minimization searches over candidate tree structures, while the inner minimization fits a counterfactual matrix and leaf values to the trees. Specifically, we minimize the discrepancy between the observed data and the sum of an estimated counterfactual matrix and the estimated treatment effects. Formally, we aim to solve the following:

$$\min_{T_1,\dots,T_q \in \mathcal{T}_{\ell_{\max}}} \quad \min_{\substack{M \in \mathbb{R}^{n\times T} \\ \tau \in \mathbb{R}^{q\times \ell}}} \quad \frac{1}{2} \left\| O - M - \sum_{i=1}^{q} \sum_{j=1}^{\ell} \tau_{i,j}\, C_j^i(T_i) \right\|_F^2 \; + \; \lambda \, \|M\|_\star,$$

where $\mathcal{T}_{\ell_{\max}}$ is the set of regression trees with at most $\ell_{\max}$ leaves, and each $C_j^i(T_i) \in \mathbb{R}^{n\times T}$ is a binary cluster matrix indicating which entries both receive intervention i and fall into the j-th leaf of tree T_i.

Because solving this problem exactly is computationally intractable, PaCE instead constructs trees greedily. That is, in each iteration, we greedily choose a *valid split* on a covariate which allows the objective to be minimized. A valid split is one that complies with predefined constraints, which will be discussed in our theoretical guarantees.

3 Theoretical guarantees

We show that PaCE recovers the treatment effect functions $\{\mathcal{T}_i^\star(\mathbf{X})\}_{i=1}^q$ under mild conditions. The final treatment effect estimate given by PaCE has two sources of error. First, there is the *approximation error*, which stems from the approximation of the non-parametric functions $\mathcal{T}_i^\star$ by piece-wise constant functions. The second source of error, the *estimation error*, arises from the estimation of the ATE of each cluster. Note that the deviation between our estimate and the true treatment effect is, at most, the sum of these two errors. We will analyze and bound these two different sources of error separately to demonstrate the convergence of our method.

3.1 Approximation Error

We adopt the *α-regularity* condition from Definition 4b of Wager and Athey (2018). This condition requires that each split retains at least a fraction $\alpha \in (0, \frac{1}{2})$ of the available training examples. We further require that the depth of each leaf be on the order $\log \ell$ and that the trees are *fair-split trees*. That is, during the tree construction procedure, for any given node, if a covariate j has not been used in the splits of its last πp parent nodes, for some $\pi > 1$, the next split must utilize covariate j. A *valid split* is defined as one that retains at least one treated observation on each side of the split and satisfies both the α-regularity and fair-split tree conditions. Additionally, we assume that the proportion of treated observations with covariates within a given hyper-rectangle should be approximately proportional to the volume of the hyper-rectangle. This is a 'coverage condition' that allows us to accurately estimate the heterogeneous treatment effect on the whole domain using the available observations.

Assumption 3.1. Suppose that all covariates belong to $[0,1]^p$. Let $x^{(1)} \leq x^{(2)} \in [0,1]^p$ be the lower and upper corners of any hyper-rectangle. We assume that the proportion of observations that have covariates inside this hyper-rectangle is proportional to the volume of the hyper-rectangle $V := \prod_{p' \in [p]} \left(x_{p'}^{(2)} - x_{p'}^{(1)} \right)$, with a margin of error $M := \sqrt{\frac{\ln(nT)(p+1)}{\min(n,T)}}$. More precisely,

$$\underline{c}V - c_m M \leq \frac{\#\left\{(z,t) : x^{(1)} \leq X^{zt} \leq x^{(2)}\right\}}{nT} \leq \bar{c}V + c_m M,$$

for some fixed constants $\bar{c} \geq \underline{c} > 0$ and some $c_m > 0$.

In the following result, we demonstrate that, as the number of observations grows, a regression tree, subject to some regularity constraints, contains leaves that become increasingly homogeneous in terms of treatment effect.

Theorem 3.2. *Let $\mathbb{T}$ be an α-regular, fair-split tree, split into ℓ leaves, each of which has a depth of at least $c \log \ell$ for some constant c. Suppose that $\ell \leq \left(\frac{\alpha}{2c_m M} \right)^{\frac{1}{c \log 1/\alpha}}$ and that Assumption 3.1 is satisfied. Then, for each treatment $i \in [q]$ and every leaf of $\mathbb{T}$, the maximum difference in treatment effects between any two observations within the cluster C of observations in that leaf is upper bounded as follows:*

$$\max_{X_1, X_2 \in C} |\mathcal{T}_i^\star(X_1) - \mathcal{T}_i^\star(X_2)| \leq \frac{2L_i\sqrt{p}}{\ell^s},$$

where $s := \frac{c}{(\pi+1)p} \frac{\alpha \underline{c}}{4\bar{c}}$ and L_i is the Lipschitz constant for the treatment effect function $\mathcal{T}_i^\star$.

This demonstrates that it is possible to approximate the true treatment effect functions with a regression tree function with arbitrary precision. As the number of leaves ℓ in the tree $\mathbb{T}$ increases, the approximation error of the tree decreases polynomially.

3.2 Estimation error

We also bound the estimation error, showing that it converges as $\max(n, T)$ increases. We state the result informally here due to space constraints and defer the full theorem and proof to the extended version.

We require constraints on the treatment application patterns. Intuitively, we require that linear combinations of these patterns with $M^\star$ do not yield a low-rank matrix, ensuring that treatment effects are not confounded with the low-rank counterfactual. In addition, we assume the treatment matrices are not collinear; otherwise, the individual effects of each treatment could not be separately identified.

Theorem 3.3. *The estimation error for a given cluster is inversely proportional to the number of treated observations in that cluster, provided that there are $O(\log n)$ clusters.*

The convergence rate of PaCE is primarily determined by the slower asymptotic convergence rate of the approximation error. However, our result in Theorem 3.3 is of independent interest, as it extends previous results on low-rank matrix recovery with a deterministic pattern of treated entries to allow for multiple treatment matrices.

4 Empirical evaluation

In real-world observational data, the true treatment effects are unobserved, making it impossible to evaluate the accuracy of various methods. Therefore, the use of semi-synthetic data, where the treatment effects are artificially generated and thus known, is common in causal inference (e.g., Arkhangelsky et al. (2021)).

To assess the accuracy of PaCE, we compare its performance on semi-synthetic data with alternative methods. Using publicly available panel data as the baseline $M^\star + E$, we introduce an artificial treatment and add a synthetic treatment effect to treated entries to generate the outcome matrix O. Since the true HTEs are known, we can verify the accuracy of various methods. Our results show that PaCE often surpasses alternative methods in accuracy, while using a tree with no more than 40 leaves to cluster observations. Thus, PaCE not only offers superior accuracy, but also provides a simple, interpretable solution.

Data To demonstrate the effectiveness of our methodology, we use two publicly available U.S. economic datasets. The first source comprises monthly Supplemental Nutrition Assistance Program (SNAP) user counts by zip code in Massachusetts, spanning January 2017 to March 2023, obtained from the Department of Transitional Assistance (DTA) public records (Massachusetts Government, 2024). The second data source comprises nine annual demographic and economic data fields for Massachusetts zip codes from 2005 to 2022, provided by the U.S. Census Bureau. We use the SNAP user count as the baseline, $M^\star + E$, with the zip code-level census data for Massachusetts as covariates. The panel size is 517 zip codes over 70 months.

Treatment generation We conduct empirical evaluations across a range of scenarios. To generate the treatment pattern, we vary two parameters: 1) the proportion α of units receiving the treatment $\alpha \in \{0.05, 0.25, 0.5, 0.75, 1.0\}$ and 2) the functional form of the treatment—either adaptive or non-adaptive. In the non-adaptive case, a random α proportion of units receive the treatment for a random consecutive period of time. In the adaptive case, for each time period, we apply the treatment to the $\alpha/2$ proportion of units that show the largest absolute percentage change in outcome in the previous time period. The purpose of the adaptive treatment is to mimic public policies that target either low-performing or high-performing units.

The treatment effect is generated by randomly sampling two covariates and either adding or multiplying them, then normalizing the magnitude of the effects so that the ATE is 20% of the average outcome. The treatment effect is added to the outcome to simulate an intervention increasing SNAP usage.

Each set of parameters is tested on 200 instances, with the treatment pattern and treatment effect being randomly regenerated for each instance. In Table 1, we present the average nMAE obtained by each method, with standard deviations shown in parentheses. Results are shown for the five top-performing methods. PaCE attains the best accuracy, achieving the lowest average nMAE. This result is particularly notable given that PaCE relies on a single tree estimator, whereas the competing methods employ more sophisticated black-box machine learning models.

Table 1: Average normalized mean absolute error(nMAE) (lower is better).

Method	nMAE	(sd)
PaCE	**0.21**	(0.18)
Causal Forest (Athey et al., 2019)	0.24	(0.16)
LinearDML (Chernozhukov et al., 2017a,b)	0.27	(0.17)
DML (Chernozhukov et al., 2017a,b)	0.43	(0.30)
XLearner (Künzel et al., 2019)	0.56	(0.25)

References

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(3):1148–1178.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017a). Double/debiased machine learning for treatment and causal parameters. Technical report.

Chernozhukov, V., Goldman, M., Semenova, V., and Taddy, M. (2017b). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv*, pages arXiv–1712.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Massachusetts Government (2024). Department of transitional assistance caseload by zip code reports. `https://www.mass.gov/lists/department-of-transitional-assistance-caseload-by-zip-code-reports`. Accessed: 2024-05-09.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.