Double Machine Learning for Continuous Treatment Effects in Panel Data: An Application to Extreme Heat in Agriculture

Sylvia Klosin[*]and Max Vilgalys[†]

February 11, 2022

Abstract

This paper introduces and proves asymptotic normality for a new semi-parametric estimator of continuous treatment effects in panel data. Specifically, we estimate a weighted average derivative of the regression function, or the average causal effect of a marginal increase in treatment. Our estimator uses the panel structure of data to account for unobservable time invariant heterogeneity and machine learning methods to estimate functions of high-dimensional inputs flexibly. We use tools from double debiased machine learning (DML) and automatic double machine learning (ADML) literature to construct our estimator. This estimator is helpful for many questions in climate policy, where it's crucial to measure the economic impacts of weather while accounting for unobservable spatial factors and without imposing strong parametric restrictions. We demonstrate the performance of our estimator in a simulation exercise. We then apply our estimator to study the elasticity of crop yield with respect to extreme heat in the United States.

Keywords: Weighted average derivative, double machine learning, panel data

JEL Classification: C14, C21, C55, Q51, Q54

[*] email: klosins@mit.edu

[†] email: vilgalys@mit.edu

1 Introduction

Import questions in environmental and energy economics often involve unobservable spatial components and nonlinear relationships between high-dimensional environmental factors. For example, researchers wishing to measure the impact of extreme heat on crop yield must account for persistent spatial factors like soil quality and complex interactions between temperature and precipitation. Typically, researchers assume a functional form for the environmental factors and partial out the unobservable spatial components using standard econometric approaches. A common approach is assuming a linear functional form and using fixed effects, first differences, or within transformations (Wooldridge, 2010). Machine learning (ML) methods allow researchers to avoid making restrictive functional form assumptions while preserving statistical power. However, standard ML methods fail to account for unobservable spatial components and may induce biases. We propose a debiased machine learning approach that addresses these econometric challenges.

Specifically, we propose a semi-parametric estimator of causal treatment effects in panel data settings and prove asymptotic normality. We demonstrate the estimator's performance with a simulation exercise and an application measuring the impact of extreme heat on crop yield. We focus on using the Least Absolute Shrinkage and Selection Operator (Lasso), although our procedure holds for general ML algorithms as long as they satisfy a convergence rate condition.

Our exposition and proof focuses on the weighted average derivative, although our debiasing procedure applies to related causal parameters such as continuous treatment effects and average derivatives. In future work, we plan to extend our proof to these cases.

We believe this estimator is useful for many policy-relevant questions. Environmental applications often involve measuring causal parameters in the presence of high-dimensional control variables, such as the damage to crop yield from increasing levels of extreme heat (Schlenker and Roberts, 2009; Burke and Emerick, 2016) or the mortality effects of exposure to criteria air pollutants (Di et al., 2017; Deryugina et al., 2019). In spatial data, the continuous treatment effect is widely used; Liu and Su (2020) use distance as a treatment to study the spatial gender wage gap, and Diamond and McQuade (2019) use distance as a treatment to study the impact of low-income housing on neighborhoods.

We contribute to the literature on high dimensional panel data. Primarily, we generalize the high dimensional panel treatment estimators by Belloni et al. (2016). They use a partially linear model (PLM) and a binary treatment - we relax the PLM assumption and generalize to a model that allows heterogeneous treatment effects for the continuous treatment setting. Furthermore, while Belloni et al. (2016) focused on within transformed data, we primarily focus on first differences.

Key assumptions that we share with Belloni et al. (2016) include that our high dimensional covariates are sparse but that the fixed effects are not sparse. Like them, we also assume errors are additive and strict exogenity conditional on covariates, which are two assumptions almost always made in applied work.

Alternate nonparametric panel approaches exist that relax the two previous assumptions (e.g. Evdokimov 2010; Chernozhukov et al. 2013). Such approaches often lead to partial identification challenges, which we avoid in our approach, and are not well-suited to high dimensional covariates. We see our estimator as a great alternative - offering the flexibility of modern ML while still allowing for the treatment effect to be identified.

This project also contributes to a literature on measuring economic damages from climate change by estimating causal parameters while relaxing functional form restrictions. Hsiang (2016) summarizes the measurement challenge and some leading approaches. These methods generally rely on linear approximations of functions, often guided by domain experts to select candidate models. Machine learners are increasingly used to measure economic impacts of environmental variation when domain expertise is not available. Crane-Droesch (2018) shows that in agriculture, machine learners that account for spatial heterogeneity significantly improve prediction, and Deryugina et al. (2019) use machine learners to flexibly model the mortality response from increased air pollution. However, naive machine learners can introduce biases when estimating causal parameters. Our procedure allows approximately debiased estimation of causal parameters while relaxing common linear assumptions.

The rest of our paper is organized as follows. In Section 2, we introduce our parameter of interest and introduce the assumptions needed for identification of the target parameter. Section 3 lays out the learning problem and defines our estimation procedure. Section 4 shows the results from our simulation exercise and empirical application. Section 5 concludes.

2 Setup

2.1 Notation

We work in a panel data setting with n individuals and T time periods. As is often the case in economic data, we assume that n is large but T is small. We assume we have independent and identically distributed data $(W_1, \cdots, W_n)$ where the $W_i = \{(X_i, D_i, Y_i)\}_{t=1}^T$ are copies of a random variable W with support $\{\mathcal{W} = \mathcal{X} \times \mathcal{D} \times \mathcal{Y}\}_{t=1}^T$, with a cumulative distribution function (cdf) $F_{YDX}(Y, D, X)$. We use capital letters to denote random variables and lowercase letters to denote their possible values. For each unit in a large population $X_{i,t} \in \mathbb{R}^p$ denotes a vector of covariates, with p potentially large, and $D_{i,t} \in \mathbb{R}$ as the treatment.

For a given variable X, we use the notation $\Delta X_{i,t} := X_{i,t} - X_{i,t-1}$ for the first difference transformation.

2.2 Parameter of Interest

To be specific we are estimating the general additive fixed effects panel model.

$$Y_{i,t} = a_i + \gamma(X_{i,t}, D_{i,t}) + \epsilon_{i,t} \qquad E[\epsilon_{i,t} | a_i, X_{i,1}, \cdots, X_{i,T}, D_{i,1}, \cdots, D_{i,T}] = 0 \tag{1}$$

Here a_i represents individual fixed effects. The γ is a flexible high dimensional function of treatment, covariates, interactions, and higher order terms. In writing our outcome model in this way, we are imposing a few key functional assumptions. For one we impose that the individual fixed effects a_i are additively separable. Second, the γ function is not indexed by time t - the function is assumed to be constant over time. We are assuming there are no treatment dynamics - the lagged values of our treatments or covariates are not included in γ.[1] We assume that γ can be estimated well with Lasso, which imposes a form of sparsity on covariates. However we do not impose that the fixed effects a_i are sparse. We choose these modeling assumptions because they match those commonly used in applied work, while relaxing functional form assumptions on γ.

Our estimation target is a weighted average derivative defined by the following function:

$$\tau_0 = \mathbb{E}[m(W_{i,t}, \gamma)] = \mathbb{E}\left[\int \omega(u) \frac{\partial \gamma_0(u, X)}{\partial d} du\right] \tag{2}$$

The causal interpretation of eq. (2) is the average causal effect of a marginal increase in treatment, for some researcher-specified probability distribution $\omega(u)$.

Let's unpack equation (2). First, note that because treatment is continuous and our functional form is very flexible the derivative $\frac{\partial \gamma_0(D_{i,t}, X_{i,t})}{\partial D_{i,t}}$ is allowed to be different at each value of $D_{i,t}$. We aggregate weighting some marginal effects more than others using a weight function ω. This ω is a function the applied researcher can choose. For example, one could use the pdf of a normal distribution centered at the average value of the treatment to put more weight on marginal effects at average values. Or to simulate how a change in distribution of the treatment effect may impact the population, a researcher could use a counterfactual distribution.

It is also interesting to note an an alternate form of eq. (2) by adding two more equalities:

$$\tau_0 = \mathbb{E}[m(W_{i,t}, \gamma)] = \mathbb{E}\left[\int \frac{\partial \omega(u)}{\partial u} \frac{-1}{\omega(u)} \gamma_0(u, X) du\right] = \mathbb{E}\left[\frac{\partial \omega(U)}{\partial u} \frac{-1}{\omega(U)} \gamma_0(U, X)\right] \tag{3}$$

The second equality follows by integration of parts as explained in Chernozhukov et al. (2018b). The random variable U can be thought of as a simulation draw from the specified probability distribution, and is independent of X. This equivalence is very useful as it enables us to estimate an average derivative without having to numerical differentiate the γ function. It can be challenging to establish a convergence rate for numerical differentiation because the derivative operator is unbounded, and there are few statistical guarantees on the convergence of a derivative in a high-dimensional setting. By loading the derivative operator onto ω, we avoid this challenge.

In future work, we plan to extend our theory to include the following related causal parameters:

[1] This assumption is in contrast to the dynamic panel case. Chernozhukov et al. (2017) propose an estimation strategy for this case.

Average Derivative

$$\tau_0 = \mathbb{E}[\frac{\partial \gamma_0(D,X)}{\partial D}]$$

This parameter is the average causal effect of a marginal increase in yield at the observed treatment levels in the data. This parameter has been studied by (among others) Imbens and Newey (2009); Rothenhäusler and Yu (2019).

Average Structural Function

$$\beta_0(d) = \mathbb{E}[\gamma_0(d,X)]$$

This parameter is the average outcome variable if all units were assigned the treatment d. It is also known as the dose-response function, and has been studied by (among others) Blundell and Powell (2001); Colangelo and Lee (2020)

Example 1. *Simulation*
Now we define a τ_0 for an simple example data generating process (DGP) in order to help the reader better understand our object of interest. Let's say we have scalar control variable $X_{i,t}$ and our treatment variable is $D_{i,t}$. Let's say our outcome $Y_{i,t} = a_i + 2X_{i,t} + 3D_{i,t}X_{i,t}^2$. Our fixed effect $a_i \sim N(1,1)$. Our covariate and treatment are distributed jointly normal

$$\begin{pmatrix} X_{i,t} \\ D_{i,t} \end{pmatrix} \sim N\left(\begin{pmatrix} a_i \\ a_i \end{pmatrix}, \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right).$$

Hence our fixed effects, treatment, and covariate are all correlated with one another. In this case $\frac{\partial \gamma(D_{i,t},X_{i,t})}{\partial D_{i,t}} = 3X_{i,t}^2$. Therefore $\tau_0 = \mathbb{E}[3X_{i,t}^2] = \mathbb{E}[3(1+a_i^2)] = 3+6 = 9$. Here we did not use a special weight and took the expectation over all possible values of the treatment.

In this example, the weighted average is also able to recover the average derivative for any probability ω. This is because the derivative is not a function of $D_{i,t}$; generally, the value of τ_0 depends on the choice of weight function. Consider $\omega(u) = \text{Uniform}([0,1])$; then, $\mathbb{E}[\int \omega(u) \frac{\partial \gamma_0(u,X)}{\partial d} d_u] = \mathbb{E}[\gamma_0(1,X_{i,t}) - \gamma_0(0,X_{i,t})] = \mathbb{E}[3X_{i,t}^2]$ as above.

Example 2. *Crop Yields*
To see how this parameter could apply to a policy question of interest, consider the relationship between crop yields and exposure to extreme heat. Here, the treatment variable $D_{i,t}$ is annual aggregate exposure to temperatures above 29°C. This variable is of interest because it roughly captures the amount of heat stress a plant experiences. As Schlenker and Roberts (2009) demonstrate, crop yields are generally increasing in damaging heat exposure, while increasing in heat exposure below 29°C. The covariates $X_{i,t}$ include other weather features, such as heat exposure below this temperature threshold and precipitation. Heat exposure is measured in Growing Degree Days (GDD), the amount of time a crop is exposed to temperatures between an upper and lower bound during the March-August growing season.

The weighted average derivative captures the average damaging effect of $D_{i,t}$ over a specified distribution. This is especially of interest when considering the effects of climate change; we could compare the average damages under weather from global climate projections under varying carbon emissions pathways. Burke et al. (2015) conduct a similar exercise, but with a more restrictive functional form.

In our applied example, the effect of marginally increasing damaging heat ($D_{i,t}$) on crop yield ($Y_{i,t}$) differs depending on covariates ($X_{i,t}$). For example, heat stress may be partially offset by increased precipitation; we would expect

$$\frac{\partial\gamma_0(D_{i,t}, X_{i,t} = \text{high precipitation year})}{\partial D_{i,t}} \leqslant \frac{\partial\gamma_0(D_{i,t}, X_{i,t} = \text{low precipitation year})}{\partial D_{i,t}} \leqslant 0.$$

It is also important to account for unobservable spatial heterogeneity. Crane-Droesch (2018) demonstrates that, even after nonparametrically controlling for a wide array of observable soil characteristics, accounting for unobservable factors considerably improves accuracy when predicting corn yields. In our empirical application, we focus on corn yields from the continental United States. In Figure 1, we illustrate the magnitude of county-level unobservable components in the United States, by plotting the average residual from a nonparametric regression of corn yield on weather variation. These persistent components can be quite large; the maximum magnitude exceeds one standard deviation of log corn yields.

2.2.1 First Difference

We want to emphasize that our object of interest is the derivative of γ with respect to the treatment - not the derivative of y. We want to control for the time invariant unobserved a_i. As we assume T is small, we do not have sufficient data to consistently estimate a_i. A classic technique for controlling for a_i is the first difference transformation. For space we don't go into details of the transformation here, but details can be found in Wooldridge (2010). Because of the transformation we estimate $\Delta\gamma(X_{i,t}, D_{i,t}) := \gamma(X_{i,t}, D_{i,t}) - \gamma(X_{i,t-1}, D_{i,t-1})$ rather than $\gamma(X_{i,t}, D_{i,t})$ directly. Given that the derivative is with respect to current-period $D_{i,t}$ we are still able to recover our original object of interest.[2] The τ_0 in equation (2) is the panel data is the same as the τ_0 in equation (4).

$$\tau_0 = \mathbb{E}[\int \omega(u)\frac{\partial\Delta\gamma_0(u, X_{i,t})}{\partial d_{i,t}} d_u] = \mathbb{E}\left[\frac{\partial\omega(U)}{\partial d_{i,t}}\frac{-1}{\omega(U)}\Delta\gamma_0(U, X_{i,t})\right] \tag{4}$$

Here, $\Delta\gamma_0(U, X_{i,t}) := \gamma_0(U, X_{i,t}) - \gamma_0(D_{i,t-1}, X_{i,t-1})$. When our estimator of $\Delta\gamma_0$ has a form such that γ_0 is consistently estimated from first differences, it is possible to write a simplified form for τ_0. The estimators we propose satisfy this condition because we assume the function is time stationary and the operator is linear in basis functions of $\{D, X\}$. Then, we can write:

[2] If we write $h(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) := \gamma_0(D_{i,t}, X_{i,t}) - \gamma_0(D_{i,t-1}, X_{i,t-1})$, the target derivative is the partial derivative of h with respect to its first argument.

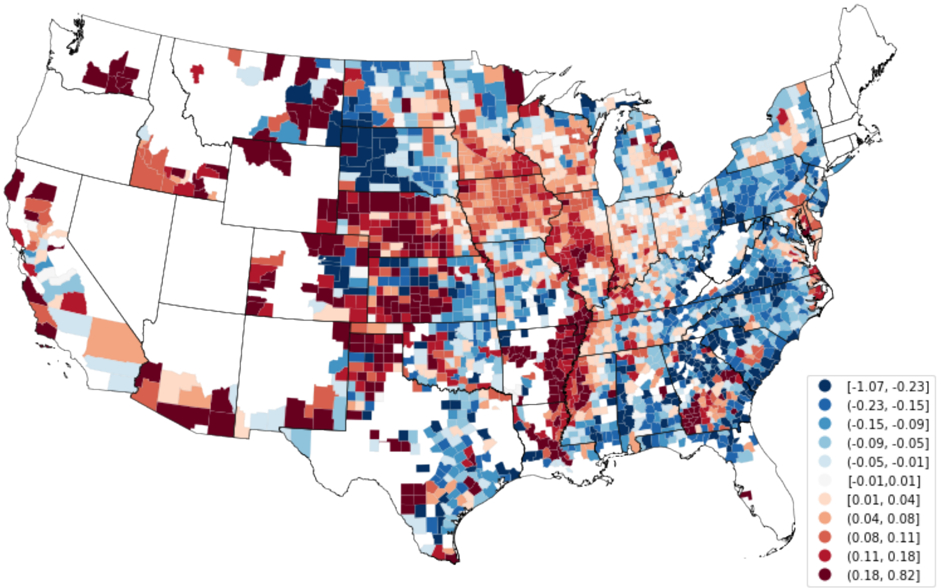


Figure 1: Spatial heterogeneity in average corn yields, after flexibly controlling for weather. Units are standardized as the share of standard deviation of log corn yield. Specifically, the county average of residuals from the following regression: $\hat{\varepsilon}_{i,t} := \log Y_{i,t} - \mathbb{E}[\log Y_{i,t}|X_{i,t}]$, where the vector $X_{i,t}$ includes annual precipitation and GDD within 1°C temperature bins from 0°C to 40°C. We estimate the regression using a random forest regressor with 100 estimators, on county-level annual yield and temperature exposure from 1990-2019. Weather data are generously shared by Schlenker and Roberts, and crop yields are from the US Department of Agriculture.

$$\begin{aligned}\tau_0 &= \mathbb{E}\left[\int \omega(u) \frac{\partial \Delta \gamma_0(u, X_{i,t})}{\partial d_{i,t}} d_u\right] \\ &= \mathbb{E}\left[\int \omega(u) \left(\frac{\partial \gamma_0(u, X_{i,t})}{\partial d_{i,t}} - \frac{\partial \gamma_0(D_{i,t-1}, X_{i,t-1})}{\partial d_{i,t}}\right) d_u\right] \\ &= \mathbb{E}\left[\frac{\partial \omega(U)}{\partial d_{i,t}} \frac{-1}{\omega(U)} \gamma_0(U, X_{i,t})\right]\end{aligned} \tag{5}$$

Note that the third equality follows because we assume that treatment is independent of its lagged values; this implies $\frac{\partial \gamma_0(D_{i,t-1}, X_{i,t-1})}{\partial d_{i,t}} = 0$. This is a convenient form because the estimator is a function only of $X_{i,t}$.

The parameter of interest τ_0 is implicitly defined by the moment function g.

$$g(W_{i,t}, \tau_0, \gamma_0) = \frac{\partial \omega(U)}{\partial u} \frac{-1}{\omega(U)} \gamma_0(U, X_{i,t}) - \tau_0 \tag{6}$$

$$\mathbb{E}[g(W_{i,t}, \tau, \gamma_0)] = 0 \text{ iff } \tau = \tau_0 \tag{7}$$

Here $\Delta\gamma_0(D, X)$ is a nuisance parameter that must be estimated in order to estimate the parameter of interest.

2.3 Identification

Our parameter of interest is identified with the following assumptions:

Assumption 1. *(Identification)*

1. *(Strict Exogeneity)* $E[\epsilon_{i,t}|a_i, X_{i,1}, \cdots, X_{i,T}, D_{i,1}, \cdots, D_{i,T}]$
2. *(Overlap) For any* $d_{i,t} \in \mathcal{D}$*, we have* $f_{D_{i,t}|X_{i,t},X_{i,t-1},D_{i,t-1}}(d_{i,t}|X_{i,t}, X_{i,t-1}, D_{i,t-1})$ *is bounded away from zero*
3. *(Consistency)* $D_{i,t} = d_{i,t}$ *implies* $Y = Y(d_{i,t})$

We use the classic strict exogenity assumption from the panel literature (Wooldridge, 2010). This assumption is potentially restrictive, as it does not allow treatment variables to be correlated with lagged values of the outcome variable.

Overlap is also known as the "positivity" assumption. We are assuming that the propensity score is uniformly bounded away from 0 for all values in the support of the pre-treatment variables. As discussed in Imbens (2000), when one has a continuous treatment this may be harder to satisfy than the more commonly studied binary treatment case. This condition also becomes harder to satisfy when the dimensionality of the covariates is large. A key aspect of out papers that current treatment $D_{i,t}$ is not just a deterministic function of $D_{i,t-1}$ due to the time varying in every period nature of continuous treatments - which is needed for the overlap condition.

Consistency is a causal assumption that is not always explicitly stated, but is basically always assumed in some form. Consistency means that the observed outcome for individuals with treatment level d equals her outcome if she had received treatment d (Hernán and Robins, 2010).

3 Learning Problem

3.1 Debiased Moment

If a researcher wanted to estimate the τ_0 parameter with ML, a first pass at the problem could be a naive "plug-in" approach. In such an approach, we would use ML to fit the model for nuisance parameter $\Delta\gamma(D_{i,t}, X_{i,t})$, and then predictions of the model would be used to create our parameter τ_0 according to the moment equation given in (6).

Naive use of a machine learner can bias estimation of our causal parameters, due to regularization and overfitting bias. We overcome this by introducing an ADML estimator, following

Chernozhukov et al. (2018a,b). The ADML estimator is created by creating a new debiased moment function which enables us to avoid the bias of the naive approach. We denote this new debiased moment function by ψ.

$$\begin{aligned}\psi(W_{i,t}, \tau_0, \gamma_0, \alpha_0) &= g(W_{i,t}, \tau_0, \gamma_0) + \phi(W_{i,t}, \tau_0, \gamma_0, \alpha_0) \\ &= \frac{\partial \omega(u)}{\partial u} \frac{-1}{\omega(u)} \Delta\gamma_0(D_{i,t}, X_{i,t}) - \tau_0 + \alpha_0(W_{i,t})(\Delta Y_{i,t} - \Delta\gamma_0(X_{i,t}, D_{i,t}))\end{aligned} \tag{8}$$

We see that now in addition to the original g function we know also have a function ψ that is a function of a new parameter α. The introduction of a new parameter introduces a new learning problem, but the overall estimate will be *doubly robust* in the sense that small estimation error in $\Delta\gamma$ and/or α do not introduce bias to the overall estimate. See Chernozhukov et al. (2018b) for more details.

This form of the estimator comes from the Riesz representation theorem. One consequence of this theorem is, for data W and functional $m(W, \gamma)$, there exists a function α such that for any γ:

$$\mathbb{E}[m(W_{i,t}, \gamma)] = \mathbb{E}[\alpha(W_{i,t})\gamma(W_{i,t})] \tag{9}$$

This implies that the function α is identified from data orthogonally from estimating γ, and can be used to construct a doubly robust estimator of our target parameter.

For this setting, let $W_{i,t} := \{D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}\}$.

The Riesz Representation theorem gives us:

$$\mathbb{E}[m(W_{i,t}, \Delta\gamma_0)] = \mathbb{E}[\alpha_0(W_{i,t})\Delta\gamma(D_{i,t}, X_{i,t})] - \mathbb{E}[\alpha_0(W_{i,t})\Delta Y_{it}] \tag{10}$$

Where the second equality follows from iterated expectations. We discuss the estimation of α in section 3.3.

3.2 Estimation

We use the empirical analog of the debiased moment function (8) as our estimator. Note that we work with $(T-1)$ time periods rather than T because we removed one time period by first-differenceing the data.

$$\hat{\tau} = \frac{1}{n(T-1)} \sum_{\ell=1}^{L} \sum_{i\in\ell} \sum_{t=2}^{T} \left[\left(\int \frac{\partial \omega(u)}{\partial u} \frac{-1}{\omega(u)} \Delta\hat{\gamma}(u, X_{i,t}) du \right) + \hat{\alpha}_\ell(W_{i,t})(\Delta Y_{i,t} - \Delta\hat{\gamma}_\ell(X_{i,t}, D_{i,t})) \right] \tag{11}$$

Before we start estimating the functions, we need to take several transformations of the data. First, we take basis functions of (X, D). For Lasso, common practice is to include polynomial expansions of each term as well as interactions between terms. For other machine learning approaches, the user does not typically define basis functions; instead, the machine learner selects appropriate functions from a large set of candidate functions. Given an iterative procedure that selects these functions, we can think of these selected functions as analogous to basis functions. For a discussion of how this relates to a neural network, see B. It is important that basis functions are applied before differencing of the data. For example, to include interacted terms in a first-differenced model, we wish to include the terms $\Delta(X_{i,t}D_{i,t}) := X_{i,t}D_{i,t} - X_{i,t-1}D_{i,t-1}$ and not $\Delta X_{i,t}\Delta D_{i,t} = (X_{i,t} - X_{i,t-1})(D_{i,t} - D_{i,t-1})$

Stage 1

i Start with data splitting. First pick the number of splits L, where $L \in \{2, \cdots, n\}$[3]. Then partition the unit's indices into the L different groups. We use ℓ to denote these groups $\ell = 1, \cdots, L$. Denote observations in group ℓ by W_ℓ. It is important that for each person in the data, all of their observations go into the same fold. This is because the data in different folds has to be independent, and there is a dependence within the observations of a person.

ii For each fold ℓ estimate the nuisance parameters $\Delta\hat{\alpha}_\ell$ and $\Delta\hat{\gamma}_\ell$

Stage 2

iii Using the nuisance parameters predicted on the left out folds construct the new debiased moment function ψ to create our estimate of τ by summing across all observations in (11)

iv Calculate the variance using the new moment function

$$\hat{V} = \frac{1}{n(T-1)} \sum_{\ell=1}^{L} \sum_{i\in\ell} \sum_{t=2}^{T} \hat{\psi}_{\ell it}^{2} \tag{12}$$

where $\hat{\psi}_{\ell it} := \psi(W_{i,t}, \hat{\tau}, \hat{\gamma}_\ell, \hat{\alpha}_\ell) = g(W_{i,t}, \hat{\tau}, \hat{\gamma}_\ell) + \phi(W_{i,t}, \hat{\tau}, \hat{\gamma}_\ell, \hat{\alpha}_\ell)$

Now we go into details of the construction.

3.3 Stage 1: Nuisance Parameters

In this section, we introduce estimators for $\hat{\alpha}(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})$ and $\Delta\hat{\gamma}(D_{i,t}, X_{i,t})$. We suppress the ℓ notation and index from $t = 1, \ldots, T$; understand that these estimators are constructed using the cross-folds procedure and first-differencing procedure above.

For exposition, we focus on Lasso[4]. We consider a rich, flexible specification by including polynomial basis functions. Let $b(D_{i,t}, X_{i,t})$ denote the $p \times 1$ dictionary of functions.[5]. Then let $\Delta b(D_{i,t}, X_{i,t}) := b(D_{i,t}, X_{i,t}) - b(D_{i,t-1}, X_{i,t-1})$.

[3] Common default numbers of splits include $L = 5$ and $L = 10$

[4] please see appendix for extension to NN

[5] For example, when we estimate γ we set $b(D_{i,t}, X_{i,t})$ to be a third order polynomial set of the covariate variables and interactions

constructing $\hat{\alpha}(W)$ We assume that α can be represented well linearly with the series of basis functions we have selected [6].

Our goal is to find a vector of coefficients $\hat{\rho} \in \mathbb{R}^p$ for our dictionary such that we can write an estimator $\hat{\alpha}$ of α_0.

$$\hat{\alpha}(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1}) = (b(D_{i,t}, X_{i,t}) - b(D_{i,t-1}, X_{i,t-1}))'\hat{\rho} \tag{13}$$

α is identified from eq. (10) by the following restriction. Note that this equality holds for any function γ; we set $\gamma = b(D_{i,t}, X_{i,t})$.

$$\mathbb{E}\left[\frac{\partial\omega(U)}{\partial u}\frac{-1}{\omega(U)}b(U, X_{i,t})\right] = E[\alpha(W_{i,t})(b(D_{i,t}, X_{i,t}) - b(D_{i,t-1}, X_{i,t-1}))] \tag{14}$$

Note that from eq. (5), we are able to write an expression involving $\gamma_0(U, X_{i,t})$ instead of $\Delta\gamma_0(U, X_{i,t})$. Let $m(W_{i,t}, b) := \int \frac{\partial\omega(u)}{\partial u}\frac{-1}{\omega(u)}b(u, X_{i,t})d\omega(u)$. We can find the appropriate $\hat{\rho}$ by solving the following regularized minimum distance problem:

$$\hat{\rho} = \underset{\rho}{\operatorname{argmin}}\left\{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{p}\left(m(W_{i,t}, b_j) - \Delta b(D_{i,t}, X_{i,t})'\rho\Delta b_j(D_{i,t}, X_{i,t})\right)^2 + 2r_L|\rho|_1\right\}; \quad |\rho|_1 = \sum_{j=1}^{p}|\rho_j|, \tag{15}$$

Where b_j denotes a single basis function. When $\Delta b(D_{i,t}, X_{i,t})$ is set so that each basis function has mean zero and variance 1, this expression simplifies:

$$\hat{\rho} = \underset{\rho}{\operatorname{argmin}}\{-2\hat{M}'\rho + \rho'\hat{Q}\rho + 2r_L|\rho|_1\} \quad |\rho|_1 = \sum_{j=1}^{p}|\rho_j|, \tag{16}$$

where $\hat{M}$ and $\hat{Q}$ are defined as

$$\hat{M} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} m(W_{i,t}, b) \tag{17}$$

and

$$\hat{Q} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} b(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})b(D_{i,t}, X_{i,t}, D_{i,t-1}, X_{i,t-1})' \tag{18}$$

constructing $\Delta\hat{\gamma}(D, X)$ Our goal is to find a vector of coefficients $\hat{\beta}$ for our dictionary such that

$$\Delta\hat{\gamma}(D, X) = \Delta b(D, X)\hat{\beta}.$$

[6] this is formally given in Appendix A in assumption 3

We can find the appropriate $\hat{\beta}$ by solving the following Lasso problem,

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\Delta Y_{i,t} - \Delta b(D_{i,t}, X_{i,t})\beta)^2 + r_L |\beta|_1 \right\} \quad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|, \tag{19}$$

3.4 Stage 2: Debiased Score

From stage 1 we have functions $\Delta\hat{\alpha}(D, X)$, $\Delta\hat{\gamma}(D, X)$, and $\frac{\partial \Delta \hat{\gamma}_\ell}{\partial D_{i,t}}$ for each of the folds ℓ. Recall that $\hat{\alpha}_\ell(D, X)$ and $\hat{\gamma}_\ell(D, X)$ were fit using W_ℓ^C, and now in stage two we use these functions to find fitted values for observations in W_ℓ. Sum up all observations in all folds, and sum up all folds and divide by n as in equation (11)

Similarly, sum up over all observations at in equation (12) to calculate the estimate of the variance. Code for this procedure is available upon request.

3.5 Asymptotic Normality

We now state our main theorem in which we prove the asymptotic normality of the estimator we just described. The details of the rate and regularity conditions as well as proof are left to Appendix A

Theorem 3.1. *(asymptotic normality) Given Assumptions* (2) (3) (4) (5)*. Let* $n \to \infty$

$$\sqrt{n}(\hat{\tau} - \tau_0) \xrightarrow{d} N(0, V), \quad \hat{V} \xrightarrow{p} V \tag{20}$$

Where

$$V_= (G'\Upsilon G)^{-1} G'\Upsilon\Psi\Upsilon G (G'\Upsilon G)^{-1} \tag{21}$$

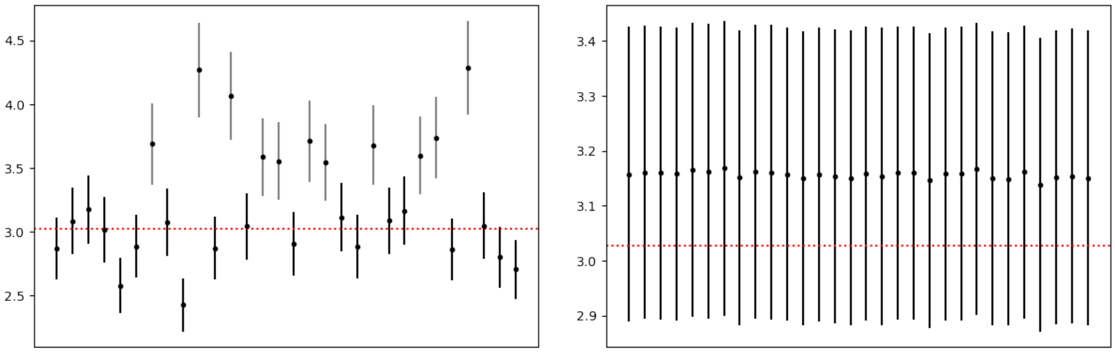
Where $G = \frac{\partial g(\tau)}{\partial \tau}$, Ψ is the average of ψ terms above, and Υ is a GMM variance matrix

4 Applications

4.1 Simulation

We demonstrate the performance of our estimator through a simulation exercise, similar to Example 1.

We have scalar control variable $X_{i,t}$ and our treatment variable is $D_{i,t}$. We specify a DGP, $Y_{i,t} = a_i - 0.1D_{i,t} + X_{i,t} + D_{i,t}X_{i,t}^2 + \varepsilon_{i,t}$. Our fixed effect $a_i \sim N(1, 1)$ and error term $\varepsilon_{i,t} \sim N(0, 1)$. Our covariate and treatment are distributed jointly normal



(a)

(b)

Figure 2: 95% confidence intervals of our debiased estimator using draws from a Gaussian distribution (fig. 2a) and the analytical derivative fig. 2b. The dotted red line denotes the true average derivative. The analytical derivative estimator has 100% coverage, while the Gaussian weight estimator has 63.3% coverage.

$$\begin{pmatrix} X_{i,t} \\ D_{i,t} \end{pmatrix} \sim N\left(\begin{pmatrix} a_i \\ a_i \end{pmatrix}, \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right).$$

We simulate a dataset for $N = 1000$ and $T = 2$ for this dataset, and estimate the average derivative using our procedure described above. We set $L = 5$ for the number of cross-folds and use basis functions up to third order polynomials of all interacted inputs. We use Pedregosa et al. (2011) to estimate $\hat{\gamma}$, and estimate $\hat{\alpha}$ using Diamond and Boyd (2016) to solve the minimization problem in Equation (15).

We estimate the average derivative using two approaches: (1) using simulation draws from a standard normal distribution, i.e. $m(W_{i,t}, \gamma) = \int \frac{\partial \omega(u)}{\partial u} \frac{-1}{\omega(u)} \gamma(u, X_{i,t}) du$, for $\gamma(u) = \phi(u)$, where ϕ denotes the probability density function of the standard normal distribution, and (2) using the analytical derivative, i.e. $m(W_{i,t}, \gamma) = \frac{\partial \gamma(D_{i,t}, X_{i,t})}{\partial d_{i,t}}$. Note that for the normal distribution with mean μ and variance σ^2, the score has the following convenient form: $\frac{\partial \omega(x)}{\partial x} \frac{-1}{\omega(x)} = \frac{x - \mu}{\sigma^2}$.

The 95% confidence intervals of simulation trials for each analysis are shown in Figure 2. The confidence intervals contain the true parameter in 63.3% using Gaussian weight estimator and 100% of trials using the analytical derivative. In general, the derivative operator is unbounded and we do not expect this procedure to deliver consistent rates for truly non-parametric specifications; we expect that the good performance of the analytical derivative is due to the relatively simple specification of the data generating process.

4.2 US Agriculture

We estimate the elasticity of US corn yield with respect to extreme heat. This parameter is of interest because it summarizes the extent that warming temperatures decrease crop yields,

	OLS	Lasso	
	(1)	(2)	(3)
GDD above	-0.0058	-0.0069	-0.0069
29°C	(0.0001)	(0.0000)	(0.0000)
MSE	0.0412	0.0223	0.0217
Simulations		30	30
Debiased			X

Table 1: Preliminary results from estimating elasticity of corn yield with respect to extreme heat in US agriculture. Lasso estimation of average derivative uses the analytical derivative. Standard errors are given in parentheses. All procedures indicate that yields decline from exposure to extreme heat, although the more flexible models find a greater degree of damage.

one of the most direct economic impacts of climate change. We follow Schlenker and Roberts (2009) and define extreme heat as GDD exposure above 29°C; this is D_{it} for our example. Our X_{it} includes GDD exposure below 29°C and annual precipitation. Our data are at the US county level, and span 1990-2019. Weather data are generously shared by Schlenker and Roberts, and crop yields are from the US Department of Agriculture's National Agricultural Statistics Service[7].

We estimate γ and α in this setting using third-order polynomials of all interacted terms. We present preliminary results from the analytical derivative for this estimation. In Table 1 we report central estimates and standard errors from our procedure, as well as ordinary least squares (OLS) estimates. The Lasso model has considerably lower mean squared error (MSE), as would be expected from a more flexible model. Our procedure finds higher effects than OLS, although all approaches indicate (as expected) that yields decline from exposure to extreme heat. This indicates that standard models may underestimate the extent of damages from extreme heat, although we stress that these results are preliminary.

5 Conclusion

In this paper, we introduced a new estimator of continuous treatment effects in panel data. Our estimator controls for time-invariant unobserved heterogeneity, which is empirically relevant for panel and spatial data sets. We also allow for novel flexible non-parametric estimation of the treatment and its interaction with observed covariates using tools from the DML and ADML literatures. We prove asymptotic normality for our proposed estimator and provide empirical simulation results that display that our estimator performs well in practice. We also discuss how our estimator is relevant to significant economic and policy topics like climate change. For example, when measuring the impact of important economic quantities like the impacts of temperature change on crop yield, it is essential to model them flexibility, which our estimator allows.

[7] Crop yield data are available here

References

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Blundell, R. and Powell, J. L. (2001). Endogeneity in nonparametric and semiparametric regression models. *NA*.

Burke, M. and Emerick, K. (2016). Adaptation to climate change: Evidence from US agriculture. *American Economic Journal: Economic Policy*, 8(3):106–140.

Burke, M., Hsiang, S. M., and Miguel, E. (2015). Global non-linear effect of temperature on economic production. *Nature*, 527(7577):235–239.

Chatterjee, S. and Jafarov, J. (2015). Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

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

Chernozhukov, V., Newey, W., and Singh, R. (2018a). De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.

Chernozhukov, V., Newey, W. K., and Singh, R. (2018b). Automatic debiased machine learning of causal and structural effects. *arXiv preprint arXiv:1809.05224*.

Colangelo, K. and Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett*, 13:114003.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., and Schwartz, J. D. (2017). Air pollution and mortality in the medicare population. *New England Journal of Medicine*, 376(26):2513–2522.

Diamond, R. and McQuade, T. (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy*, 127(3):1063–1117.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*, 1.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics*, 8:43–75.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Liu, S. and Su, Y. (2020). The geography of jobs and the gender wage gap. *Available at SSRN*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rothenhäusler, D. and Yu, B. (2019). Incremental causal effects. *arXiv preprint arXiv:1907.13258*.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 106(37):15594–8.

Tibshirani, R. and Wasserman, L. (2016). A closer look at sparse regression.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

A Asymptotic Normality

Assumption 2. *(Bounded dictionary) There exists a C such that with probability one*

$$\max_{1\leqslant j\leqslant p} |b_j(W)| \leqslant C \tag{22}$$

Assumption 3. *(Sparse regime) Assume that the following hold.*

1. *There exists $C, \xi > 0$ such that for all $\bar{s}$ with $\bar{s} \leqslant C(\sqrt{\frac{\ln(p)}{n}})^{\frac{-2}{(1+2\xi)}}$ there is a $\bar{\rho} \in \mathbb{R}^p$ with $|\bar{\rho}|_1$ and $\bar{s}$ nonzero elements s.t.*

$$\|\alpha_0 - b'\bar{\rho}\| \leqslant C(\bar{s})^{-\xi} \tag{23}$$

2. *$Q = \mathbb{E}[b(W)b(W)']$ is nonsingular and has the largest eigenvalue uniformly bounded in n*

3. *for $\rho = \bar{\rho}$ and $\rho = \arg\min_\rho\{\|\alpha_0 - b'\bar{\rho}\|^2 + 2r_L \sum_{j=1}^p |\rho_j|\}$ there is $k > 3$ such such that*

$$\inf_{\{\delta:\delta\neq 0, \sum_{j\in\mathcal{J}_\rho^c} |\delta_j| \leqslant k \sum_{j\in\mathcal{J}_\rho} |\delta_j|\}} \frac{\delta' Q \delta}{\sum_{j\in\mathcal{J}_\rho} \delta_j^2} > 0 \tag{24}$$

   where $\mathcal{J}_\rho = support(p)$

Part 3 of assumption 3 is a population version of the restricted eigenvalue condition of Bickel et al. (2009) as adapted in Chernozhukov et al. (2018b). A clear introduction of the restricted eigenvalue condition is given in Tibshirani and Wasserman (2016).

Assumption 4. *(Regularization)*

$$r_n = a_n(\sqrt{\frac{\ln(p)}{n}}) \textit{ for some } a_n \to \infty \tag{25}$$

To satisfy this assumption we set $a_n = \ln(\ln(n))$, following Chatterjee and Jafarov (2015). In practice, we pick a data-driven value following Chernozhukov et al. (2018b).

Lemma A.1. *(Sparse regime) If assumption 3 holds, we get a rate of*

$$\|\hat{\alpha}_\ell(t, X) - \alpha_0(t, X)\| = O_p(\epsilon_n^{\frac{-1\xi}{(1+2\xi)}} r_L) \tag{26}$$

Assumption 5. *(Regression rate)*

For each $\ell = 1, ..., L$

$$\|\Delta\hat{\gamma}(D, X) - \Delta\gamma_0(D, X)\| = O_p(n^{-d_\gamma}) \tag{27}$$

$d_\gamma \in (\frac{1}{4}, \frac{1}{2})$

Corollary A.1.1. *Under assumptions 3, 5 and 4*

$$\|\hat{\alpha}(W) - \alpha_0(W)\|\|\Delta\hat{\gamma}(W) - \Delta\gamma_0(W)\| = o_p((n)^{-\frac{1}{2}}) \tag{28}$$

Corollary A.1.1 shows the trade off in the error permitted in estimating our two nuisance parameters - the regression function $\Delta\gamma_0$ and the α_0.

Theorem A.2. *(asymptotic linearity) Given Assumptions* (2) (3) (4) (5)

$$\sqrt{n}(\hat{\psi}(\tau_0)) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}}\psi(W, \tau_0, \gamma_0, \alpha_0) + o_p(1) \tag{29}$$

Proof follows by Lemma 15 from *Chernozhukovet al.* (2016). Three conditions need to hold to follow this proof structure. First are the mild mean square consistency conditions - which are satisfied given the rate conditions given in A.1 for the debiasing term and in Assumption 5 for our conditional expectation function $\Delta\gamma$. Second is an assumption that the controls the interaction of the nuisance parameters, that is controlled by Corollary A.1.1. Lastly, there is an assumption that controls that average of the double robustness term ψ.

Theorem A.3. *(asymptotic normality) Let the same assumptions hold. Let $n \to \infty$*

$$\sqrt{n}(\hat{\tau} - \tau_0) \xrightarrow{d} N(0, V), \quad \hat{V} \xrightarrow{p} V \tag{30}$$

Where

$$V_{=}(G'\Upsilon G)^{-1}G'\Upsilon\Psi\Upsilon G(G'\Upsilon G)^{-1} \tag{31}$$

Where $G = \frac{\partial g(\tau)}{\partial\tau}$, Ψ is the average of ψ terms above, and Υ is a GMM variance matrix

B Extension to NN

We also consider an alternate estimator for γ: the semiparametric neural network (SPNN) from Crane-Droesch (2018). This approach takes two stages. First, an iterative procedure selects an appropriate set of basis functions by minimizing squared loss from a neural network. The output of this neural network is the sum of a per-unit fixed effect estimate and a fully connected transformation of the flexible inputs D, X. Note that estimates of a_i from this procedure are inconsistent for small panels, due to the incidental parameters problem. In the second stage, we apply the within function using the basis functions identified in the first stage, resulting in a consistent estimate of γ.

It is straightforward to compute derivatives of neural networks because the training process involves computing gradients of outputs with respect to inputs. As Baydin et al. (2018) discuss, automatic differentiation is an important tool to rapidly train large or deep neural networks. By using this same machinery, we can directly estimate gradients from the neural network. This algorithmic constraint also implies that neural networks can only represent functions with bounded derivatives. While we are not aware of a rate condition for the derivative of a neural network, this suggests that neural networks can be used to estimate average partial derivatives. In future work, we plan to include an implementation using this property of a neural network.