Regularizing the Forward Pass

Patrick Power, Shomik Ghosh and Markus Schwedeler

DRAFT

May 1, 2023

Abstract

We introduce an estimation framework that generalizes ordinary least squares, allows for nonparametric cluster effects,[1] and is inherently compositional, even under regularization.

Keywords: Causal Inference, Deep Learning

[1]To clarify this point, in many applied microeconomic contexts, individuals grouped together in some fashion. In our context, this grouping is done via a shared zip code. We indicate this grouping via an indicator variable. The nonparametric component is that we allow these the effect of this group memebership on the outcome of interest to vary across the other controls/features. That is, we don't place a functional form assumption on the way the cluster indicator variable interacts with the other controls.

1 Introduction

1.1 Context

Many of the most influential economic studies make use of data consisting of "clusters" of individuals – Moving to Opportunity, Oregon Health Insurance Experiment, the STAR Experiment. Such settings are attractive because they often balance identification with general equilibrium results.

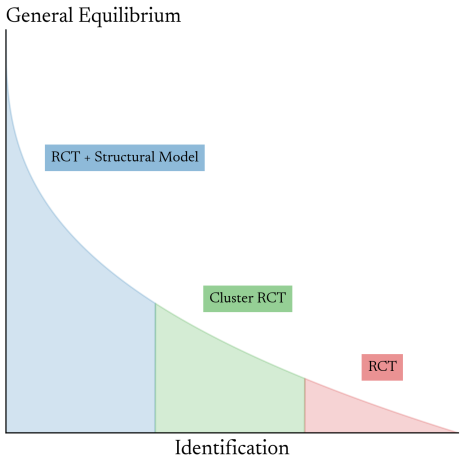


Figure 1

In a large number of of these contexts, whether it's to estimate treatment heterogeneity, adjust for the propensity score, or control for pre-treatment outcomes, it is common to estimate the conditional expectation function with a cluster specific feature. The statistical challenge is how to do this in a way such that it generalizes across the unobserved/missing clusters. To this aim, we propose a framework that generalizes ordinary least squares, allows for nonparametric cluster effects, and is inherently compositional, even under regularization.

1.2 Challenge

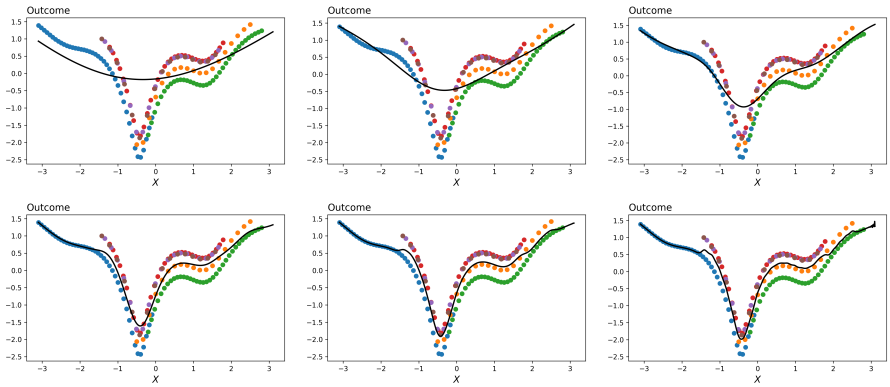
We begin by describing the context in which estimating the conditional expectation function, $\mathbb{E}[Y|X,D]$[2] can be quite challenging. In section 6, we introduce the necessary mathematical structure to formalize the problem.

[2]We abuse terminology here by writing "the" conditional expectation function.

$$\omega \longmapsto \mathbb{E}[Y|X,D](\omega)$$
$$\int Y d\mathbb{P}_A = \int \mathbb{E}[Y|X,D] d\mathbb{P}_A, \quad \forall A \in \sigma(X \times D)$$

Estimating the condition expectation function can be challenging when (1) we observe only a subset of the clusters in both the treated and control groups, (2) the distribution of covariates differ across clusters, and (3) the distribution of outcomes conditional on covariates differ across clusters. We illustrate a stylized 1-dimensional simulation of the *tragic triad of clustered data* in figure 2.[3] A standard nonparametric model (black line) with a

Figure 2: The Tragic Triad of Clustered Data



Reproduced Here: In this figure, we assume away within-cluster variation. Each dot corresponds to an observation. The different colors highlight the various clusters.

"smoothing" hyperparameter struggles in this context: in order to fit the 'v'-shaped component of the data where there is general consensus across the clusters, the bandwidth of the estimator must be small. In doing so, it over fits the fails. Intuitively, what's needed in this context is an estimator that is "locally" aware of the cluster structure of the data.

The tragic triad most frequently occurs when an identification strategy relies on a control variable that only varies at the cluster level. Taken from our work on The Right to Counsel at Scale, which uses the average outcome within zip codes prior to the policy as a control, figure 4 illustrates how the neighborhood of an observation differs dramatically by whether the feature vector associated with the observation contains an element that varies at the cluster level. It's interesting to note that these issues are perhaps only magnified as we increase the dimensionality of the data. Extending the work of Balestriero et al. [2021], we illustrate in figure 5a that clustered sampling doesn't change the fundamental issue of learning in high dimensions (extrapolation) so much as it motivates us to reconsider how we go about it, as highlighted in figure 5b.

[3]We borrow the expression "the tragic triad" from Chelsea Finn's paper on Gradient Surgery

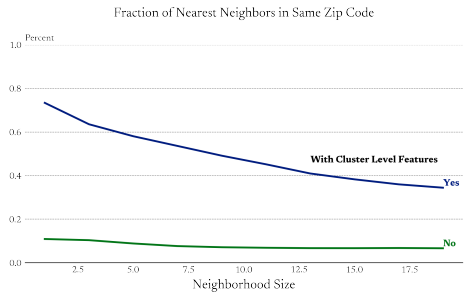
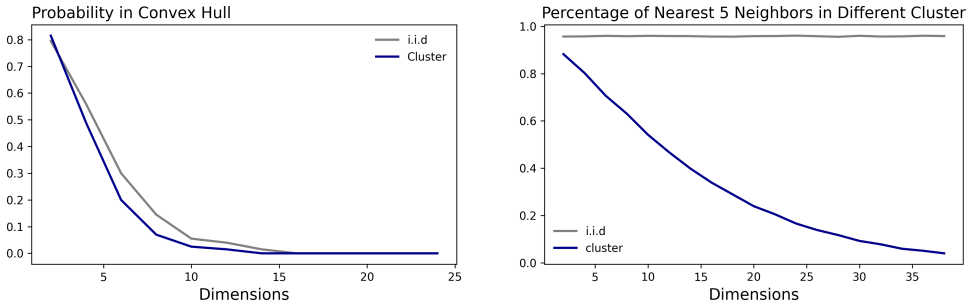


Figure 4: Using data from our paper The Right to Counsel at Scale, we plot the fraction of the 'k' nearest observations that are in the same sample where we allow 'k' to vary across the x-axis.



(a) (b)

Figure 5: (a) Probability that a observation in the test set is in the convex hull formed by the training set (b) Fraction of the five nearest neighbors in a different cluster. Data consists of 25 cluster and 25 observations per cluster. Clusters differ only in the mean which is drawn from an isotropic gaussian distribution: Reproduced Here and Here

1.3 Preview of Results

Thus far, we've (1) highlighted that typical smoothing methods can perform poorly in the context of *Tragic Triad of Clustered Data* and (2) that many empirical strategies involve controlling for a cluster level feature which therefore likely brings them into this context.[4] We wrap up this introduction by illustrating how our framework, building off of Finn et al. [2017] and Kelly et al. [2020], enables one to account for the clustered nature of one's data in a way that is conceptually similar to standard deep learning practices.

Deep learning models consists of two key components: forming predictions by composing functions and updating predictions via reverse-mode automatic differentiation. In our approach to learning from clustered data, these two ingredients remain the same. The difference is that instead of working in the category of Sets (where objects are types and morphisms are functions), we work in the Kelesie Category where types are "expanded" to include a penalty term, functions are "embellished" passing through both the predictions

[4]Of course, this is not necessary nor sufficient

and penalty values, and composition is "augmented" so that a function with a single input can be composed with multi-valued output function.

Figure 6: Regularization via Composition

	Prediction	Training
Application	$f(x)$	$\tilde{f}(x) := \big(f(x), m(f)(x)\big)$
	$f \overset{F_m}{\longmapsto} \tilde{f}$	
Composition	$f \circ g(x)$ $=f(g(x))$	$\tilde{f} \mathbin{\tilde{\circ}} \tilde{g}(x)$ $=\big(f(g(x)), m(f)(g(x)) + m(g)(x)\big)$

Regularizing the forward pass can me understood as functor mapping the Category of Sets into the Kelesie Category

We highlight in figure 8 that subject to the typical caveats of hyperparameter tuning, our model fits the 'v'-shaped nature of the data where there is consensus across the clusters without overfitting in the tails." That is, our model is implicitly locally aware of the cluster structure of the data.

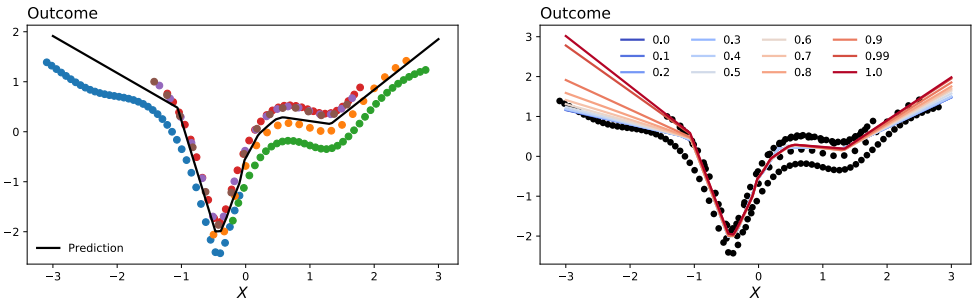


Figure 8: Reproduced Here

2 Formal Problem

2.1 Set-Up

Haskell-like Notation

We would like to avoid the following notation: $\gamma(\hat{F}_n)$. Following many in the broader machine learning/statistics/functional programming literatures, we use this haskell-like notation because (1) it allows for correctness unlike the term above and (2) because expressing functions in their curried form can often make the math cleaner.[5]

$$\begin{aligned} \hat{F}_n &:: \Omega_n \to \mathcal{X} \to [0,1] \\ \gamma &:: (X \to [0,1]) \to \mathcal{R} \\ \gamma \circ \hat{F}_n &:: \Omega_n \to \mathcal{R} \end{aligned}$$

Identification

The fundamental aim in causal inference is to generalize across the entire population of interest. The typical approach in economics is to define a probability space of interest and make some form of selection on observable assumption. Doing so turns the problem of generalization into a repeatedly sampling problem. We begin this process by describing the probability spaces of interest.

> **Population**: The sample space consists of duplicates of all individuals in the population. One under treatment and the other under control. The probability measure on this space, $\mathbb{P}_0$, is not assumed to be the uniform measure.
>
> **Sample**: The sample space consists of 'n' observations from the population
>
> **Estimator**: The sample space is simply the real line or $[0,1]$ depending on the context.

Given this probability space set-up, we're going to assume that conditional on X, treatment is good as randomly assigned.[6] That is, under the probability measure $\mathbb{P}_0$, treatment is not independent of the potential outcome variables. The typical reason for this is that the propensity score varies at the cluster level. Specifically, the selection into treatment at the

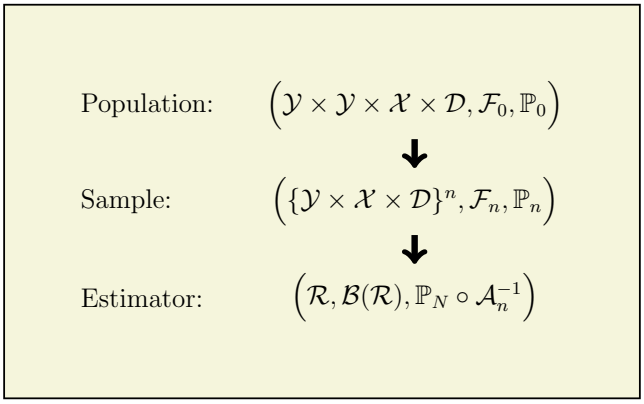
[5] If we were working with uncurried functions, we wouldn't be able to use composition.

$$\begin{aligned} f &:: \mathcal{R} \to \mathcal{R} \to \mathcal{R} \\ X &:: \Omega \to \mathcal{R} \\ f \circ X &:: \Omega \to \mathcal{R} \to \mathcal{R} \end{aligned}$$

[6] Given a probability space $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$, two random variables, X, Y, are independent if

$$\forall A, B \in \sigma(X), \sigma(Y), \quad \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Probability Spaces



cluster level is generally non-random. For instance in our accompaning work, the Right to Counsel was generally targeted at zip codes that had a higher number of evictions. The hope is that by conditioning on X, the variables are independent with respect to the generated measures.

Conditioning on a Random Variable

$$\tilde{C} :: \mathcal{M} \to (\Omega \to \mathcal{R}) \to \mathcal{F}_{\sigma(X)} \to \mathcal{M}$$

$$\{Y_i(1), Y_i(0)\} \not\perp D_i \qquad \{Y_i(1), Y_i(0)\} \perp D_i$$

With the probability spaces of interest defined, the selection on observable assumption in place, the final component in this set-up is a more complete description of the probability measure in the sample probability space. The key aspect is that we're not working with the product measure.

$$\mathbb{P}_{n,\mathcal{A}_n} := \mathbb{P}_n \circ \mathcal{A}_n^{-1}$$

In many applied contexts,
this is not the product measure

Rather, the sample probability measure can be instead realized by conditioning on the event that there are k_1 clusters in the treated group and k_2 clusters in the control group.

Two random variables are said to be conditional independent of Z if

$$\forall C \in \sigma(Z), \quad X \perp Y \ w.r.t. \ (\Omega, \mathcal{F}, \mathbb{P}_C)$$