Regularizing the Forward Pass

Patrick Power, Shomik Ghosh and Markus Schwedeler

DRAFT

April 29, 2023

Abstract

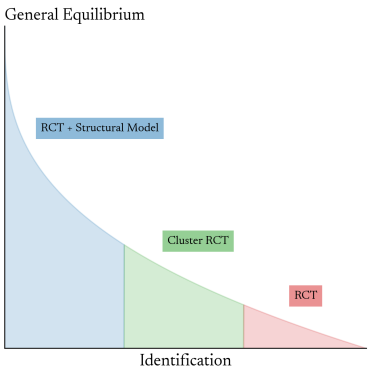
We introduce an estimation framework that generalizes ordinary least squares, allows for nonparametric cluster effects,[1] and is inherently compositional, even under regularization.

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 Motivation

In many of the most influential works in Economics, the data set consists of "clusters" of individuals. Consider for instance Moving to Opportunity, the Oregon Health Insurance Experiment, the STAR Experiment. In addition to being generally easier to implement and better adhere to the potential outcome framework, such a data setup can be attractive because it can provide insight in to the effects of the policy or treatment at scale.



For the applied microeconomist, the general intuition is that one ought to adjust one's standard errors to account of the clustered nature of the data. The following by @jondr44 captures the intuition succinctly.

> "Suppose I have sampled 1000 people i.i.d from 3 states, say CT, MA, and RI. I estimate average wages. Should I cluster my standard errors? Suppose first that I care about those three states in particular. Maybe I'm advising the governors of Southern New England. I have an i.i.d. sample from the population I care about, so clearly I don't need to cluster at all. Now, suppose I care about the average in the entire US. But I only had money to send surveyors to 3 states. And I happened to draw CT, MA, and RI. Now, I need to cluster because I effectively only have 3 observations out of 50 instead of 1000 out of the population of the US. So, with exactly the same data, the answer to whether I need to cluster depends on the question I'm trying to answer. Or, in other words, the *estimand*. – April 10, 2023

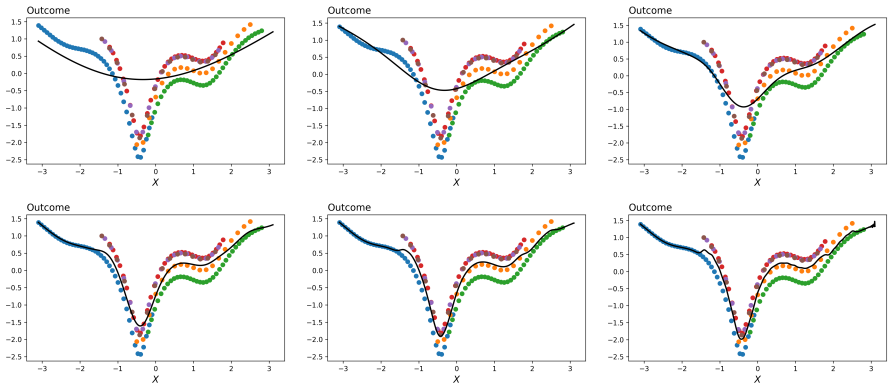
Often in applied work, though, there's an interest in estimating the Conditional Expectation Function in order to estimate treatment heterogeneity, adjust for the propensity score, or reduce sampling variation. And in this case, the statistical question becomes not merely how to adjust one's standard errors, but more broadly, how to estimate the CEF in such a way that it generalizes across the unobserved clusters.

2 Challenge

We begin be providing the intuition behind why estimating the conditional expectation function differs with clustered data. In section 3, we introduce the necessary mathematical structure to formalize the problem.

The aim is to estimate the conditional expectation function: $\mathbb{E}[Y|X, D]$.[2] This can be challenging when (1) in both the treated and control groups, we observe only a subset of the clusters, (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 1.[3] We see that a

Figure 1: 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.

standard nonparametric model (black line) with a "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, though, 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. That is, an estimator which incorporates the cluster feature during training time, even though the cluster indictor doesn't enter the CEF of interest.

It's interesting to note that these issues, the , are perhaps only magnified as we increase the dimensionality of the data. Extending the work of Balestriero et al. [2021], we illustrate in figure 3a 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 3b.

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

$$\omega \longmapsto \mathbb{E}[Y|X, D](\omega)$$

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

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