Regularizing the Forward Pass

Patrick Power, Shomik Ghosh and Markus Schwedeler

DRAFT

April 30, 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 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.

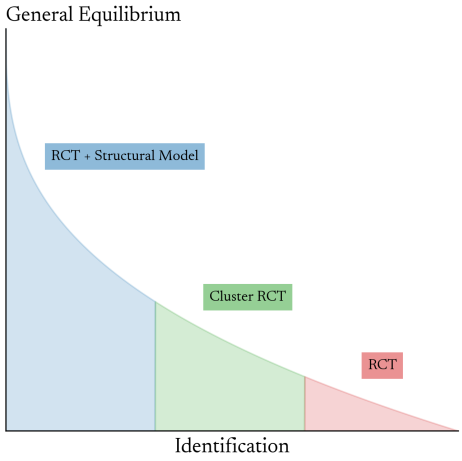


Figure 1

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.

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 3, we introduce the necessary mathematical structure to formalize the problem.

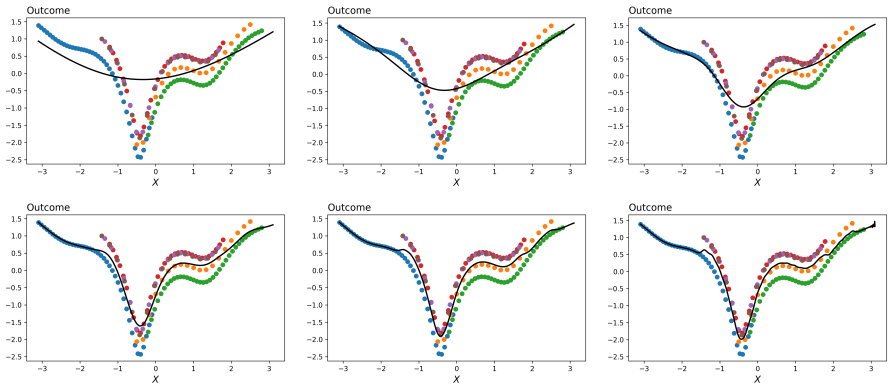
Estimating the condition expectation function can be challenging when (1) we observe only a subset of the clusters in both the treated and control groups, (2) the distribution

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

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

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.

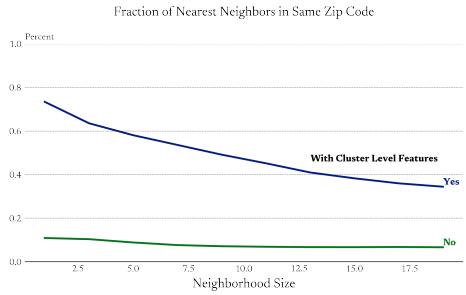
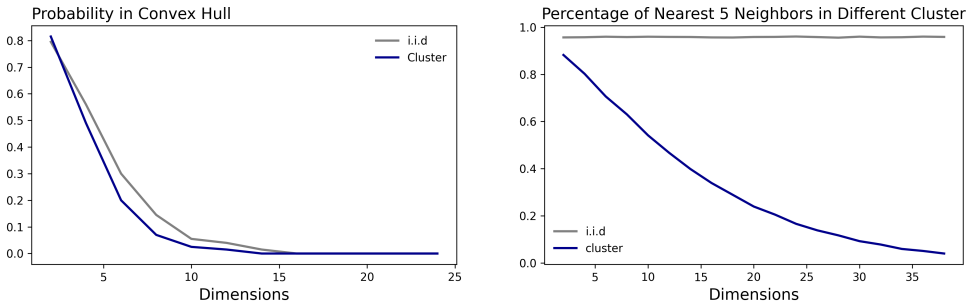


Figure 4

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

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.



(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