Synthetic control

Soheil Eshghi

May 2019

Abstract

We aim to create a convex-optimization pipeline to create better synthetic controls when we have an abundance of data for the unintervened population.

1 Introduction

To evaluate the effectiveness of an intervention, it is necessary to create plausible counterfactuals. One way to create such counterfactuals is to aggregate information about covariates and outcomes linearly from unintervened upon individuals (controls) to create a plausible unintervened synthetic individual with the same covariates as the intervened (a synthetic control), and with a composite outcome created linearly from the same individual [Abadie 2010]. The goal of this project is to improve the counterfactual created by incorporating some of the logical extensions of the synthetic control method.

These methods typically fit regression models for control subjects over the whole of the parameter space, or cluster control individuals to create locally approximate outcomes. Furthermore, they are typically developed for cases where there is a dearth of data-points for the creation of the counterfactual, and they do not address the case of creating a counterfactual in large data-sets when there may be many methods to match covariates and create synthetic controls.

In particular:

1. None of these methods explicitly consider the variance of the synthetic outcome created. In particular, the methods created focus on removing bias and minimizing variance in the co-variate space without incorporating the uncertainty in the outcome space.

2. More subtly, once a model for local relationships in the covariate space is established, it can also be used to judge the credibility of control observations used to create the synthetic control and thus exclude possibly noisy observations.

3. Finally, we aim to integrate the "local approximation" stage that chooses the input to the synthetic control with that which creates the synthetic

control, which removes some of the guesswork related to the appropriateness of ad hoc notions of locality.

We aim to create a convex optimization framework to minimize the variance of the synthetic control while limiting the contribution of noise from control measurements. As the synthetic control is used to judge the effectiveness of an intervention, creating a synthetic control with lower variance will provide evidence for effective interventions with smaller effect sizes that may have been discarded with weaker tests.

2 Model

In the simplest case, we will consider the creation of one counterfactual with data available for one time-point. Nothing precludes the generalization of this to the full model of [Abadie 2010].

Assume we have covariate-observation pairs (X_i, Y_i) for n points, and that X^*, the covariates of the intervened unit, lies within the convex hull of $(X_i)_{i=1}^n$. One way to match co-variates is to pick a vector w such that $w'1 = 1$ and $\|X^* - Xw\|$, where X is the matrix of available covariates (arranged in column form), is minimized.

One can incorporate a notion of locality by penalizing the use of covariates far from the covariates of interest, through incorporating the following term in either a constraint or as part of the objective: $\sum_{i=1}^n w_i \|X^* - X_i\|$.

The use of the transformed linear model (linking covariates to observations of outcomes) to create the counterfactual show our belief in the model specification. Under these conditions, the prediction of the model for the outcome at covariates X_i based on all other observed covariates X_{-i} (if it is within the convex hull) should closely match its observed outcome. Any discrepancy should make us less confident in the observation of the outcome (if we hold the model specification to be correct), so we would want Y_i to play less of a role in the creation of the counterfactual in that scenario. Thus, we will put an upper limit on w_i, the weight assigned to the covariate-observation pair by a function of the discrepancy between Y_i and the created synthetic prediction for it $\hat{Y}_i$ from X_{-i}.

The question of how to minimize variance in the Yw can be addressed in multiple ways. One is to incorporate the weighted sample variance (weighted by w) into a constraint or as part of the objective. However, this is not the right way to address this issue, as it abstracts out the dependence on the covariates. Thus, it is likely that we will need to create a covariate-weighted variance measure for the synthetic controls.