A Bayesian-inspired approach to constructing synthetic controls with `SyntheticControls.jl`

TP Prescott *

May 2022

The following notes aim to explain some of the mathematical and statistical reasoning underlying the code in `SyntheticControls.jl`.

1 Context

Synthetic controls are a means by which causal conclusions can be drawn. Suppose we have a set of units, one of which has an intervention applied to it at a particular time. We refer to this unit as the *intervention unit*; the other units, which do *not* have the intervention applied, are called the *candidate control units*. We assume that data on a certain quantity is collected before and after the time of intervention in all units, which we will refer to as the units' *output*. It is further assumed that the intervention is intended to affect the output for the intervention unit. To assess whether the intervention does so, we need to compare the observed post-intervention output against what the output of the intervention unit would have been without the intervention applied. Clearly, this *counterfactual* cannot be directly observed.

Synthetic controls provide a method for constructing a counterfactual. The approach taken is to construct a *synthetic control* as a weighted average of the candidate control units. This weighted average is typically taken to be a *convex combination* of the other units in the set, that have *not* had the intervention applied. The output of the synthetic control plays the role of the counterfactual.

Beyond the statistical underpinning of this approach, the key *practical* question is how to choose the weighted average defining the synthetic control. We assume that there

*The Alan Turing Institute. London NW1. England.

is pre-intervention data on the output for each of the units, including the intervened-on unit.[1] We also allow that there are other covariates for each of the units. Taken together, the covariates and the pre-intervention output provide a data space, and each unit corresponds to a single observed point in this (potentially high-dimensional) space. If the intervention unit lives in the convex hull of the candidate control units, as embedded in that data space, then we can construct a synthetic control that precisely matches the pre-intervention and covariate data of the intervention unit. Otherwise, the best synthetic control that can be found corresponds to the point on the surface of the convex hull that is closest, in data space, to the intervention unit, where the distance metric is usually some weighted Euclidean distance (discussed below).

Suppose that there are more units than dimensions. Then, in the case where the intervention unit lives in the convex hull of the candidate controls, this implies that there are potentially many different synthetic controls that all match the intervention unit perfectly. More generally, there is no reason to suppose that a single "point estimate" of the synthetic control provides the best comparison. A more robust causal analysis requires some uncertainty quantification.

This observation motivates a Bayesian-inspired approach to constructing a synthetic control. The space of all possible synthetic controls is the space of all convex combinations of the candidate control units. Embedded in data space, this is the convex hull of the candidate control units. The distance, in data space, of any synthetic control from the intervention unit induces a likelihood, in the sense that synthetic controls that are closer in data space to the intervention unit are preferred. The variance applied to this likelihood allows us to 'inflate' the set of synthetic controls away from a single 'maximum-likelihood' point in data space (i.e. the nearest point in the convex hull) and instead consider a set of synthetic controls that are all good matches to the intervention unit. A Bayesian approach also enables a user to impose a non-uniform prior, in the sense of specifying that certain candidate control units should be preferred, perhaps due to easier data collection or prior knowledge of similarity.

2 Notational Setup

[1] An aside. We usually have chosen the specific unit to intervene into because its output is somehow an outlier compared to the average of the other units. For example, we would test localised measures to improve a certain health outcome in locations where that outcome is particularly poor. Do this mean that including pre-intervention output data to construct a synthetic control is in some sense unreliable, or even invalid, thinking causally?

Let's not use pre-intervention output data to construct synthetic controls. Instead, we will just construct a synthetic control ensemble based on the values of the covariates.

- Units $j = 1, \ldots, J, J+1$, where the intervention unit is indexed as $j = J+1$;
- A $K \times J$ matrix, X_0, of covariates in the candidate control units, $j = 1, \ldots, J$;
- A $K \times 1$ vector, X_1, of covariates in the intervention unit, $j = J+1$;
- A row vector of J post-intervention output observations, Y_0, corresponding to the candidate control units, $j = 1, \ldots, J$;
- A single post-intervention output observation, Y_1, corresponding to the intervention unit, $j = J+1$;

Consider the standard unit simplex embedded in $\mathbf{R}^J$, denoted

$$\Delta^{J-1} = \left\{ w \in \mathbf{R}^J \mid \sum_{j=1}^{J} w_j = 1,\ w_j \geq 0 \right\}.$$

For each $w \in \Delta^{J-1}$, we identify a synthetic control. That is, the counterfactual to the post-intervention output observation, Y_1, is given by the post-intervention *synthetic* output observation,

$$Y_1^\star = Y_0 w = \sum_{j=1}^{J} w_j Y_{0j}$$

defined by $w \in \Delta^{J-1}$, for $Y_0 = (Y_{0j})_{j=1}^J$. In the following, we drop the superscript $J-1$ from the simplex notation.

Now consider an arbitrary probability density, $\pi(w)$, with positive support on the simplex, $w \in \Delta$. The density π induces a distribution on the space of synthetic control counterfactual outputs, such that

$$\mathbf{P}(Y_1^\star \in A) = \int_\Delta \mathbf{1}(Y_0 w \in A) \pi(w)\, \mathrm{d}w.$$

Thus, given the distribution π, we can then assess the likelihood of the observed post-intervention observation, Y_1, against the *uncertain counterfactual distribution* of $Y_1^\star$ induced by π. Given π, we can thus essentially construct an approximate p-value for Y_1, corresponding to the zero-effect null hypothesis.

2.1 Bayesian Counterfactual Distribution

The paragraph above begs the question of what the distribution, π, on the simplex should look like.

We can begin with a prior. A natural prior to use in this context is a *Dirichlet* prior, with density function π_α parametrised by the vector $\alpha = (\alpha_1, \ldots, \alpha_J)$, such that

$$\log \pi_\alpha(w) = \sum_{j=1}^{J} (\alpha_j - 1) \log w_j - \log B(\alpha)$$

on $w \in \Delta$, where we note that the beta function $B(\alpha)$ is constant in w. Note that $\alpha_j = 1$ for all j gives a uniform prior on Δ.

At this stage, we can use the information provided by the K covariate dimensions, encoded for the candidate control units in the $K \times J$ data matrix X_0 and for the intervention unit in the data column vector X_1. Our aim is that the synthetic control should be close, in the covariate data space, to the intervention unit. We define a Gaussian-style weighted sum-of-squares log-likelihood function,

$$\log L(w) = -\frac{1}{2}(X_1 - X_0 w)^T \Sigma^{-1} (X_1 - X_0 w) = -\frac{1}{2}\|X_1 - X_0 w\|_\Sigma^2,$$

for a to-be-determined positive semi-definite hyperparameter, Σ, corresponding to a Gaussian covariance in covariate data space. A sensible choice of Σ might be based on a whitening transformation or something similar. In the following, we assume that covariate data space has been transformed such that we can set $\Sigma = \sigma I$ to a scaling of the identity matrix, whereby

$$\log L(w) = -\frac{1}{2\sigma}\|X_1 - X_0 w\|^2 \tag{1}$$

is a scaling of the squared Euclidean distance. The parameter, σ, is then a free parameter.

The classical point estimate of a synthetic control corresponds to finding w to minimise $\log L(w)$. In contrast, a Bayesian-inspired approach seeks to construct the *posterior* distribution

$$p(w \mid X_0, X_1) \propto L(w)\pi_\alpha(w)$$

from the prior and likelihood. Up to an additive constant, the log probability density

of the posterior is given by

$$\log p(w \mid X_0, X_1) = -\frac{1}{2\sigma}\|X_1 - X_0 w\|^2 + \sum_{j=1}^{J}(\alpha_j - 1)\log w_j.$$

Given this expression, the free parameter σ essentially acts to balance the prior against the likelihood. As $\sigma \to 0$, the prior has less influence, and the posterior density converges onto the maximum likelihood synthetic control(s).

2.2 MCMC

We can produce a posterior sample using, for example, Metropolis–Hastings MCMC. Given w, we can produce a proposal w' from the Dirichlet proposal distribution centered on w, with log density

$$\log \pi_{\beta w}(w') = \sum_{j=1}^{J}(\beta w_j - 1)\log w'_j - \log B(\beta w)$$

where larger values of the bandwidth parameter $\beta > 1$ concentrate proposals closer to w. Given this (asymmetric) proposal density and the posterior density above, the Metropolis–Hastings acceptance probability $P(w' \mid w)$ is calculated as the exponential of

$$\begin{aligned}
\log P(w' \mid w) &= \log p(w' \mid X_0, X_1) - \log p(w \mid X_0, X_1) + \log \pi_{\beta w'}(w) - \log \pi_{\beta w}(w') \\
&= \frac{1}{2\sigma}\left(\|X_1 - X_0 w\|^2 - \|X_1 - X_0 w'\|^2\right) \\
&+ \sum_j (\beta w'_j - \alpha_j)\log w_j - \sum_j (\beta w_j - \alpha_j)\log w'_j \\
&+ \log B(\beta w) - \log B(\beta w').
\end{aligned}$$

The MCMC algorithm produces a Monte Carlo sample from the posterior, which we can consider to be an ensemble of synthetic controls from the posterior. In particular, we have a set of synthetic counterfactuals, $Y^\star_{1,n}$ for $n = 1, \ldots, N$. Thus we can compare Y_1 to the set of $Y^\star_{1,n}$ to make causal statements about the intervention.

2.3 SMC

The arguments above rely on a set value of σ. One approach to selecting σ is to follow an SMC approach, whereby σ is scaled from generation to generation. However, I wouldn't know how to choose an optimum value for σ; I think that it is, in the end, arbitrary.

3 Geometry: penalising distant units

The posterior on w is determined purely by the likelihood in Equation (1). Suppose that, when embedded in covariant data space, the intervention unit is a perfect interpolation between two candidate control units that are each very distant from the intervention unit. There may be two other units that are both far closer to the intervention unit, but the interpolation is not perfect. Intuitively, it is not clear that interpolating the distant candidate control units will give a reliable estimate of the counterfactual post-intervention output in the intervention unit.

The classical synthetic control approach uses the distant units in the single, optimal, synthetic control. In the Bayesian-inspired approach descibed here, the likelihood formulation in Equation (1) allows a synthetic control formed only by the nearby units into the posterior distribution of synthetic controls. However, all things being equal in the prior distribution (i.e. for a flat prior), then this likelihood still prefers to use the distant units, as that gives a closer match in covariate data space.

This observation motivates a potential adaptation of the likelihood function to express a preference towards the nearby units. We write

$$\log \Lambda(w) = -\frac{1}{2\sigma} \sum_j w_j \|X_1 - X_{0,j}\|^2$$

and note that the $w \in \Delta$ that minimises $\log \Lambda(w)$ is the unit vector corresponding to the candidate control unit nearest to the intervention unit in covariate data space. Here, large distances between individual candidate control units and the intervention unit are penalised.

We propose that the posterior then be adapted to take into account this additional cost, such that

$$\log p_\gamma(w \mid X_0, X_1) = \gamma \log L(w) + (1 - \gamma) \log \Lambda(w) + \log \pi_\alpha(w) \tag{2}$$

for the hyperparameter $\gamma \in [0, 1]$. Here, $\gamma = 1$ corresponds to the situation in the previous section. That is, $\gamma = 1$ imposes no penalty on distant units, such that the maximum likelihood synthetic control is equal to the classical synthetic control. At the other extreme, $\gamma = 0$ corresponds to a maximum likelihood synthetic control equal to the candidate control unit nearest to the intervention unit in covariate data space. For γ between these two extremes, the distance of individual candidate control units is balanced against the fit of the interpolation. This approach produces synthetic controls as convex combinations of candidate control units that are similar to the intervention unit, as measured in covariate space.

Like σ, the hyperparameter γ is freely chosen to give good results. Again, I wouldn't know how to choose an optimum value for γ; I think that it is, in the end, arbitrary. In any case, the alternative posterior in Equation (2) can be substituted into the Metropolis–Hastings acceptance probability, and the MCMC sampling algorithm proceed much as before.