DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

The Bayesian Synthetic Control: A Probabilistic Framework for Counterfactual Estimation in the Social Sciences

Citation
Tuomaala, Elias. 2019. The Bayesian Synthetic Control: A Probabilistic Framework for Counterfactual Estimation in the Social Sciences. Bachelor's thesis, Harvard College.

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364627

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

The Bayesian Synthetic Control: A Probabilistic Framework for Counterfactual Estimation in the Social Sciences

A thesis presented

by

Elias Tuomaala

to

Applied Mathematics

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 29, 2019

Abstract

Social sciences know limited methods for counterfactual estimation, i.e. approximating how a variable of interest would have developed without a particular policy intervention. Recent frequentist work has developed for the purpose a tool known as the Synthetic Control Method (SCM). SCM continues to suffer of certain flaws, including its inability to produce a confidence interval. In this senior thesis I develop a new Bayesian statistical methodology for counterfactual estimation inspired by the synthetic control method. I use pre-intervention data to model the target society's trajectory on underlying developments that can be inferred from data on control societies. My proposed method is less prone to overfit than synthetic controls and better able to describe estimate uncertainty. I implement the method for two previously studied research questions: German re-unification in 1990 and a California tobacco control reform in 1988. To estimate the model computationally, I use a standard MCMC sampling approach. My approach outperforms SCM in a simple test of predictive accuracy.

Acknowledgements

I cannot thank enough the advisor of this senior thesis, Rahul Dave from the Harvard School of Engineering and Applied Sciences. His inspired teaching, active mentorship, insightful feedback, and tireless support were vital for the writing of this thesis at its every stage. I also thank Professor Matthew Blackwell at the Harvard Department of Government and Professor Kosuke Imai at the Harvard Department of Statistics for their prized advice and feedback.

I also want to thank the various friends and colleagues who have shared their thoughts, questions, and feedback to my project at its various stages. Many of these conversations proved vital in steering the project to its successful completion. These individuals include but are by no means limited to Chris Truong at Cambridge Econometrics, Mikko Silliman at Harvard University, Paul Stainier at UCLA, Jonne Sälevä, Fadhil Abubaker, Otto Ahoniemi, Siddhant Agrawal, Ben Barrett, Sam Goldman, Alberto Cialandri, Krister Koskelo, Pinja Raitanen, Thomas Culp, and Ritika Philip.

Finally, I am endlessly grateful to my parents Tuija and Seppo Tuomaala, my roommates Pulkit Agarwal and Jacob Link, the Harvard Lowell House community, and many other friends and loved ones for their loving support without which this senior thesis could never have come to fruition.

Contents

1 Introduction

Social scientists often find themselves in the unfortunate circumstance where they would like to investigate a counterfactual scenario. What would have happened to this variable of interest without that one-off policy intervention? We cannot observe alternative histories, so the conclusive answer must ultimately remain elusive. However, statistical estimation can in principle give us both good approximations of the counterfactual trajectory and reliable measures of the uncertainty associated with that estimate. Alas, statistical methods appropriate for this purpose remain scarce.

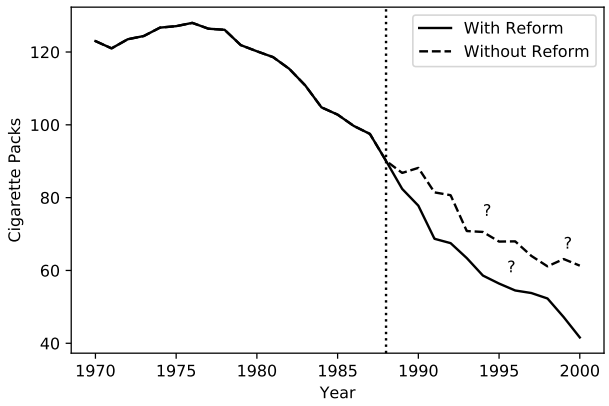
In this senior thesis, I propose a novel statistical framework for this type of counterfactual estimation. Inspired by prior frequentist work, I base my approach on the insight that societal developments are often driven by inter-society trends. Consequently, the counterfactual trajectory can be predicted from the observed trajectories of societies unaffected by the intervention. Unlike previous researchers, I lay out a fully probabilistic causal model to justify my methodology.

I then demonstrate that my model can be computationally estimated in the Bayesian framework using Markov Chain Monte Carlo (MCMC) algorithms. In contrast to prior work, my model yields full description of estimate uncertainty. I implement my model to examine two previously studied research questions: the impacts of German re-unification in 1990 and the California tobacco control reform of 1988. I demonstrate that my model outperforms the original frequentist methodology on a simple test of predictive accuracy.

1.1 Motivation

A defining feature of the social sciences is the absence of controlled experiments. Instead, researchers often need to estimate the impact of some one-off policy intervention. Say, we might ask how much the 1988 California tobacco control program reduced the state's smoking rate. Logically, that impact must equal the difference between the current rate and the smoking rate we would have observed without the intervention. Thus, we often have a pressing need to ask what could be termed the counterfactual question of the social sciences: how would the variable of interest have developed in the target society absent a given policy intervention?

Figure 1: Motivating Question: California Tobacco Control Reform



Conventional statistics might suggest a particular simple way of answering this question. We could use pre-intervention data to train a model to predict the variable of interest from some other features of that society. Post-intervention data on those covariates, then, would let us predict what should have happened to the target feature. This corresponds to explict modeling of causal a relation between features. Alas, the important covariates are often unobserved. Fur-

ther, even the observed covariates may be distorted by a treatment effect of the intervention. Together these issues of latent variables and covariate endogeneity make this inter-feature modeling approach inappropriate for most policy intervention contexts.

A promising alternative is to predict the variable of interest in the target society using not other features but other societies. For instance, we could predict the California smoking rate from data on smoking rates of other US states. This approach lacks some of the flaws of the feature model. First, endogeneity is usually mitigated. The smoking rate in Illinois, for instance, is unlikely to change much due to tobacco regulation reform in California. Second, the threat of latent variables is mitigated. Many unobserved variables will have observable effects on other societies, and the society model can in principle capture these effects.

An ostensible concern is that a society-based model does not reflect causal relations. Surely smoking rate in one state is not caused by smoking rates in others. Indeed, this is the very reason we need not worry about endogeneity. Yet, the society model in fact can be grounded in good causal understanding. Note that the unobserved variables that determine our variable of interest will often follow inter-society trends.

Smoking rates across states, for instance, may be affected by federal legislation, Big Tobacco PR scandals, and poorly understood cultural phenomena. If we suppose each state to have a unique, constant set of exposures to these trends, we can estimate jointly both the trends and the exposures. This gives the inter-society model its causal interpretation. The model estimates the underlying causal trends and their impact on the target society using data on other societies.

Prior work in quantitative social science has developed an increasingly pop-

ular frequentist method based on this inter-society principle. It is known as the synthetic control method (SCM). SCM models the target society as a weighted average of other societies. This frequentist method, however, has substantial flaws. First, it is suspect to substantial overfit concerns. Second, it fails to produce confidence intervals or other similar measures of esimation uncertainty. Third, the approach is not well grounded in a probabilistic causal model. Fourth, SCM discards much of the information contained in available data.

The problem can be better solved using the Bayesian paradigm of statistics. Bayesian models are known to be quite resistant to the threat of overfitting. More importanty, a Bayesian model will always yield a full, quantitative description of estimation uncertainty. Finally, full Bayesian treatment will naturally allow formal modeling and computational estimation of the kinds of unobserved trends discussed above.

1.2 Contribution Summary

In this senior thesis I propose a new Bayesian framework for counterfactual analysis. I present an approach that can take use of the inter-society insight that underlies synthetic controls. At the same time my approach is less, if at all, suspect to the limitations of traditional SCM. I in particular demonstrate that my framework can readily quantify the uncertainty associated with its point estimates, including the calculation of 95% credible intervals. The framework is inspired by the synthetic control method, so I name it the Bayesian Synthetic Control (BSC).

The BSC framework is defined as a probabilistic latent variable model. Each of the unobserved intersociety trends is modeled explicitly as a random, latent variable, and so are the coefficients used to construct observed society data from those latent trends. The target society's post-intervention trajectory is

treated as a sum of a counterfactual trajectory and the effect of the intervention, the latter of which is estimated as an additional model variable. The model incorporates three levels of uncertainty: over the trajectory of the latent trends, over the target society's exposure to the various trends, and over the random noise that distorts our observed data.

I lay out in detail the mathematical formulation of the probabilistic model that underlies BSC. Then I implement the method in practice to re-visit two previously studied research questions: the economic impact of the German re-unification in 1990 and the effectiveness of 1988 tobacco control reform in California. I estimate these effects computationally using a Markov Chain Monte Carlo (MCMC) algorithm, an technique that sends a sampler on a 'random walk' to explore the targeted probability distribution. I show that BSC outperforms SCM in a simple test of prediction accuracy. I also demonstrate that BSC's method of quantifying its prediction uncertainty, unlike those of its frequentist predecessors, can be used to test for modeling assumption violations.

1.3 Outline

The rest of this thesis is structured as follows. Section 2 introduces the frequentist Synthetic Control Method and reviews related literature. In Section 3 I first review the workflow of Bayesian data analysis and then exhibit how to cast the SCM research question into the Bayesian paradigm of statistics in three sequential steps. In section 4, I introduce the full Bayesian Synthetic Control framework and specify in detail the underlying probabilistic latent variable model. In section 5 I apply my model to the previously studied problem of the economic impact of German re-unification in 1990, and compare its performance to SCM. Section 6 lays out a practical application to another previously studied question, that of California's 1988 tobacco control reform, and introduces a

method of endogenously choosing the supposed number of latent variables. Section 7 discusses the model's values and limitations as well as plausible extensions to it. Section 8 offers concluding remarks.

2 Synthetic Control Method and Related Work

2.1 Origins and Motivation

SCM was first developed in 2003 [Abadie and Gardeazabal, 2003]. The original motivating problem was to estimate how the Basque Country GDP would have developed without the heightening of separatist terrorist activity in mid-1970's. To do so, the authors modeled the Basque Country as a weighted average of other regions in Spain. The method they used to choose those weights, along with some later adjustments to the procedure, is what has since come to be known as the synthetic control method.

2.2 Technical Specification

The basic SCM procedure is similar to training an OLS regression to predict the target society's value from the values of comparison societies. The model is trained on pre-intervention data to derive coefficients for the comparison societies. Those coefficients and post-intervention data on the comparison societies are then used to draw a synthetic post-intervention trajectory for the target society.

Unlike OLS, SCM restricts itself to constructing weighted averages. Mathematically this translates into constraining the regression coefficients to form a convex combination. Other societies' weights are chosen to minimize the pre-intervention error between the observed trajectory and the synthetic one. For additional robustness, this error is calculated as a weighted average of squared erros in the variable of interst and some covariates. The error calculation weights are typically chosen through cross-validation.

To describe the method formally along the lines of a later paper [Abadie et al., 2010], consider a set of $J + 1$ societies over T years such that the society

1 faces the treatment effect in all years following some intermediate year T_0. Denote by y_{it} the variable of interest in society i at time t and by $\boldsymbol{Z}_i$ a vector of r society-specific covariates that are constant over time. Collect values y_{1t} for all $t \leq T_0$ and the entries of $\boldsymbol{Z}_1$ into the vector $\boldsymbol{X}_1$ of length $r+T_0$. Similarly, collect into the $(r+T_0) \times J$ matrix the corresponding values for all other societies. Let $\boldsymbol{V}$ be a matrix of variable weights for squared error calculation. Finally, let $\boldsymbol{W}$ be the vector of J weights specified to the J comparison societies.

Define the pre-intervention error thus:

$$||\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}||_{\boldsymbol{V}} = \sqrt{(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})'\boldsymbol{V}(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})}. \tag{1}$$

Then select the synthetic control weights $\boldsymbol{W}^*$ thus:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}}{\operatorname{argmin}} ||\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}||_{\boldsymbol{V}} \tag{2}$$

$$\text{such that} \tag{3}$$

$$W_j \geq 0 \ \forall j \text{ and } \sum_{j=2}^{J+1} W_j = 1. \tag{4}$$

These weights are then used to calculate the SCM estimator $\hat{y}_{1t}^N$ for the counterfactual target society value in year t thus:

$$\hat{y}_{1t}^N = \sum_{j=1}^{J+1} W_j y_{jt}. \tag{5}$$

The details of choosing $\boldsymbol{V}$ vary in literature but they typically involve using one or more parts of the pre-intervention period as validation sets, running

the SCM procedure repeatedly, and then choosing whatever $\boldsymbol{V}$ minimizes the prediction error on the validation set(s).

2.3 Motivating Causal Model

In [Abadie and Gardeazabal, 2003], the SCM methodology was introduced as an ad hoc solution to a particular research problem. In [Abadie et al., 2010], however, the model specification is described together with a motivating causal model. Namely, if we denote the intervention-free variable value of society i in year t by y_{it}^N, the authors suppose that

$$y_{it}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \varepsilon_{it} \tag{6}$$

where δ_t indicates an annual fixed effect, $\boldsymbol{\theta}_t$ is a year-specific vector of coefficients, $\boldsymbol{Z}_i$ represents society-specific observable covariates, $\boldsymbol{\lambda}_t$ is a year-specific vector of latent variables, $\boldsymbol{\mu}_i$ denotes a vector of coefficients specific to society i, and ε_{it} is an error term drawn from a distribution centered at zero.

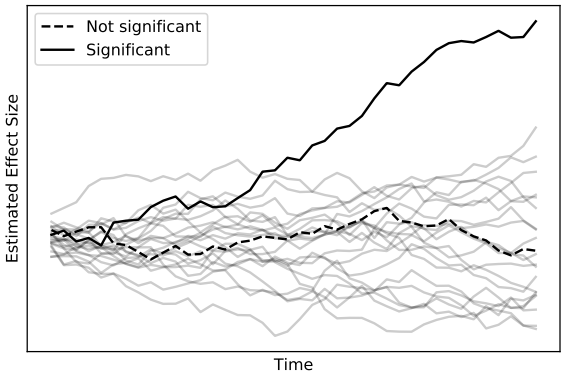
To be sure, the model is included as a motivational one. SCM does not involve estimating any of the above model parameters explicitly. However, the authors use the model to prove that under a certain set of assumptions the SCM estimator is asymptotically unbiased: the bias term converges to zero as the number of pre-treatment periods grows without bound.

2.4 Significance Testing

The same paper develops a signficance test for SCM [Abadie et al., 2010]. The test is based on relabeling and resembles Fisher's exact test. It involves constructing an SCM estimate for each of the comparison societies that did not undergo intervention. The findings are considered statistically significant if the

estimated effect is larger for the target society than for 95% of the others. This test is routinely used in the literature to check finding significance. I provide an imaginary illustration in Figure 2.

Figure 2: Canonical SCM Significance Test



2.5 Limitations

The single most important issue with the standard SCM setup is its inability to produce comprehensive description of its prediction uncertainty. The previously mentioned re-labeling technique can only test for the significance of the treatment effect's sign. (Recent researchers have also cast doubt on the test's validity as a Fisher randomization test due to lack of weighting on propensity scores [Ben-Michael et al., 2018].) Importantly, there is thus no way to set a confidence interval or to report a standard error.

SCM's design also makes it suspect to overfit considerations. Some of the variation in the comparison societies' growth trajectories is composed of random noise that doesn't reflect any trends relevant for the target society. Consequently, it is possible that the optimized weighted average correlates spuriously with the pre-intervention target society data. This threat grows as we increase

the pool of comparison societies. With many enough societies to choose from, one can always find a combination that resembles the target data very closely. To the extent that the correlation is spurious, it is unlikely to continue in the post-intervention period.

Some of SCM's design features moderate the threat of overfit. The convexity constraint of coefficients prevents the model from seeking arbitrarily overfitted linear combinations. Further, the use of the additional covariates ($\boldsymbol{Z}$) makes measuring the pre-intervention error more robust. However, it is not clear how well these design features eliminate overfit.

Additionally, these design features introduce issues of their own. The convexity constraint, for instance, prevents SCM from modeling negative correlations between societies. This amounts to discarding useful information. (Later researchers also present other reasons to find strictly nonnegative coefficients undesirable [Doudchenko and Imbens, 2017].) The use of additional covariates, on the other hand, both imposes an additional data availability constraint and burdens the analysis with an additional arbitrary specification, namely that of covariate selection.

Furthermore, the connection between SCM's implementation and the associated causal model is somewhat distant. Recall that the latter was deviced in hindsight and only serves to motivate the former. No causal relation or inter-society latent trend is estimated directly. Instead SCM relies on a (generally biased) estimator that uses other societies directly as covariates, even though we know those other trajectories to be but noisy functions of the latent explanatory factors. This limits SCM's interpretability.

Finally, we have no good way of testing whether the SCM modeling assumptions hold. This relates to the method's inability to quantify its own uncertainty. We cannot check whether predictions made for 'known counterfactuals,' or com-

parison society post-treatment data, actually fall within the confidence bounds. The closest SCM gets to a consistency check is verifying that the synthetic trajectory matches observed data in the pre-treatment period. Alas, this match can well happen even if the causal model is far from accurately describing reality.

2.6 Extensions and Related Work

The Synthetic Control Method has become the subject of active academic discourse in the past several years. Numerous applied studies have used the method to investigate topics ranging from federalism and social spending in Belgium [Arnold and Stadelman-Steffen, 2017] to the importance of the Turkey-EU customs union [Aytuğ et al., 2017] and from child mortality in the face of trade liberalization [Barlow, 2018] to the demographic consequences of HIV [Karlsson and Pichler, 2015]. The authors of the study that introduced SCM significance testing also chimed in with a follow-up article investigating the 1990 German re-unification [Abadie et al., 2015].

The SCM methodology itself has also attracted notable attention of theoretical researchers. One such study [Ferman and Pinto, 2016] investigates SCM's behavior when the method fails to find an exact match in the pre-treatment period. The authors find that under that condition the estimator is not asymptotically unbiased.

In response, another group of researchers propose an extension to SCM, the Augmented Synthetic Control Method (ASCM) [Ben-Michael et al., 2018]. The ASCM complements the SCM estimator with a bias-correction term, and shows that the result is asymptotically less biased than SCM when the pre-treatment matching is imperfect. The paper also introduces an estimator for the variance of the ASCM prediction error, and bases it on the empirical distribution of errors generated in a re-labeling exercise.

Separately, another paper proposes a closely related panel data approach (PDA) as an alternative to SCM [Hsiao et al., 2012]. The authors suppose that a number of latent inter-society factors, a society-specific intercept, and white noise determine the variable of interest. They demonstrate that the model that motivates SCM can be derived as a special case of their more general formulation. Similarly to SCM, they derive an asymptotically unbiased estimator for the counterfactual trajectory that is calculated directly using comparison society values without estimating the latent factors. Using Monte Carlo simulations the authors also find evidence of overfitting (in the form of a deteriorating bias-variance balance) as the number of comparison societies grows beyond a point. The paper does not attempt to calculate a confidence interval. Later work has since shown that PDA performs well for a wider number of simulated data generation processes than SCM [Wan et al., 2018].

Another notable contribution is the Generalized Synthetic Control, or GSC [Xu, 2017]. The GSC assumes a causal structure where observed time-invariant covariates together with latent time-variant trends drive the variable of interest. GSC differs from previous methods in that it begins by deriving an explicit point estimate for those latent trends. Its counterfactual estimator is then calculated using those latent trend point estimates. The number of factors is selected using cross-validation. The authors also introduce a novel way of setting confidence intervals. Namely, they carry out parameteric bootstrap on errors generated for comparison societies during a re-labeling exercise.

Finally, a research team at Google has made what is to my knowledge the first Bayesian contribution to the synthetic control literature [Brodersen et al., 2015]. In their paper, the authors describe a very general Bayesian state-space model for cross-sectional time-series variables. The model is in part inspired by synthetic controls and includes as one of its components a Bayesian regression

on comparison units.

At the same time, the [Brodersen et al., 2015] approach is largely designed for the study of market behavior and impact assessment for advertising. Consequently, it does not lay out a clear causal model applicable in most social scientific contexts. The abstract model is general enough to incorporate a latent linear variable structure, but the authors don't implement one. They instead focus in great part on time-series behavior such as seasonality. The authors implement a modern Markov Chain Monte Carlo (namely, NUTS) sampling method to estimate the posterior distribution of their model and thus fully quantify their prediction uncertainty.

3 Steps towards a Bayesian Synthetic Control

The recent frequentist extensions to the default SCM, namely ASCM and GSC, have taken important steps to address some of the method's limitations. However, the proposed improvements are perhaps not fully satisfying. In particular, their confidence intervals continue to depend on re-labeling methods so that the produced spread cannot distinguish between model-based uncertainty and error resulting from assumption violations. Additionally, the methods remain vulnerable to overfit.

This suggests that there remains space to improve on prior work by casting the problem in the Bayesian paradigm of statistics. Among other strengths, the Bayesian workflow is rather immune to overfitting [Bishop, 2006, p. 147] and automatically yields a full description of prediction uncertainty. The [Brodersen et al., 2015] model is a perfectly valid approach to doing so, but it is not well specified for the context of policy intervention analysis and lacks interpretability and a clear causal structure.

In this chapter I exhibit a set of intermediary steps from the original SCM method to constructing a satisfying Bayesian framework for the same purpose. First, though, I note that causal inference is most popularly done in the frequentist paradigm of statistics. I therefore begin by providing a brief review of Bayesian data analysis. For more detail, I include in the appendix a lengthier dicussion of the Bayesian and frequentist workflow differences.

3.1 Review: Bayesian Data Analysis

The conventional frequentist statistical workflow is defined by one core premise: only observable data is random. This means that observable quantites have some data generating process, they are draws from some unknown distribution,

and that running that process again would yield a different draw from that distribution. An unobserved parameter, however, is never random. It merely has some true value but no distribution or generating process from where it was drawn.

In contrast, a Bayesian analyst treats all unknown quantities as random variables, including any unseen model parameters. If we are subjectively uncertain about some quantity, it makes no difference for us whether that quantity has a generating process in the 'real world'. We can describe this uncertainty using a probability distribution. When we encounter (more) data, we update our believes to include the information contained in that data.

I illustrate the Bayesian workflow using a simplified Bayesian equivalent of the ordinary least squares regression (OLS). We begin by assuming that two variables, y_i and x_i, are correlated with some unknown slope β. Each data point y_i is normally distributed around some unknown mean μ_i with some unknown variance σ^2 so that $\mu_i = \beta x_i$.

The Bayesian and frequentist workflows agree that y_i here is properly random and could in principle be re-sampled from some distribution. That's as far as a frequentist is willing to go: the distribution's parameters β and σ are fixed even if unknown. As Bayesians, though, we consider those parameters too as random and suppose there is some underlying distribution from which they are drawn. Before actually seeing any data, we set a prior distribution for each. We could describe our prior uncertainty over β with the standard normal distribution and that for σ with the standard uniform. Figure 3 says this more visually.

For a more precise even if less exciting expression of the same model, we can also specify the Bayesian regression model algebraically:

$$y_i \sim \mathcal{N}(\mu_i, \sigma), \tag{7}$$

$$\mu_i = \beta x_i, \tag{8}$$

$$\sigma \sim \text{Uniform}(0,\ 1), \tag{9}$$

$$\beta \sim \mathcal{N}(0, 1). \tag{10}$$

Figure 3: Directed Graph of Bayesian Regression

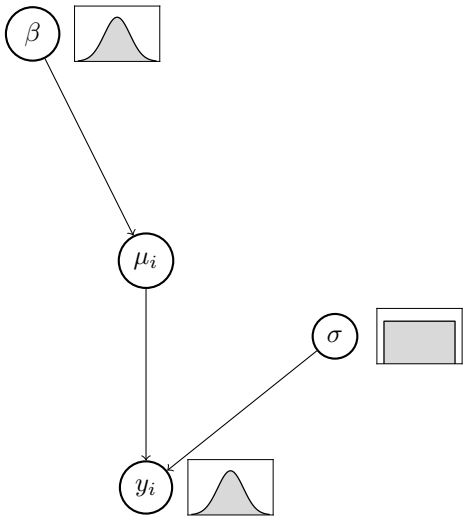


Figure 3 illustrates well our total *a priori* uncertainty over future values of y_i. We formally suppose y_i to be drawn from a gaussian. However, we don't actually know the parameters of that distribution. Therefore if we had to predict new values for y_i, we couldn't just pull them from a gaussian. Rather the distribution that captures our full uncertainty about y_i is really some complex weighted average (or integral) over many different normal distributions.

Fortunately the Bayesian paradigm allows us to disregard the details of that resulting complex distribution. The step-by-step method of doing so is referred

to as 'ancestral sampling'. First we draw new 'ancestor' samples for β and σ from their corresponding prior distributions. Then we plug those into the gaussian of y_i. Then we can easily draw a new value for y_i from a known gaussian. This procedure corresponds to climbing down the directed graph of Figure 3. Following these steps consequtively would amount to accurate estimation of the prior predictive distribution of y_i even as we may have no idea of its functional form.

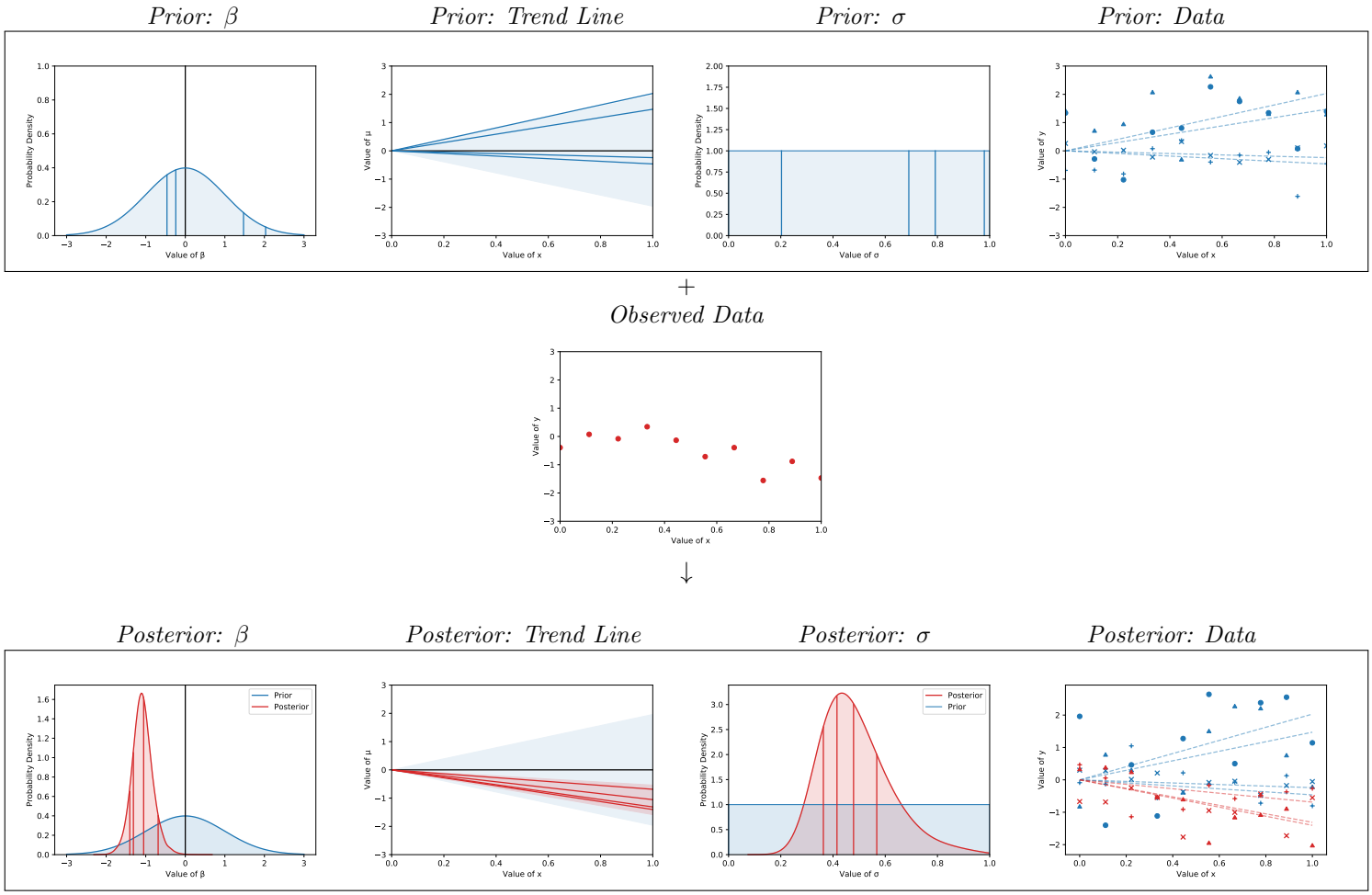
Ancestral sampling is concretely illustrated in Figure 4. The first and third blue graph on the top row visualize the prior distributions of β and σ, respectively, and highlight four random draws therefrom. The second graph on top simply takes each β draw and plots out the corresponding slope against the 95% confidence interval of such slopes. Finally, the top-right graph illustrates the output of completed ancestral sampling. For each of the four β, σ-pairs, the graph contains a 10-item dataset generated from the resulting y_i gaussian.

In a real data analysis setting, though, we observe rather than generate a dataset. For our example, we observe the data shown in the middle panel of Figure 4. The core of Bayesian data analysis consists of updating our prior beliefs, shown in the top panels, with the information contained in this observation.

This updating has a particular mathematical formulation, given by the famous Bayes' Theorem, but the task is intuitively very simple. It is really an exercise in ancestral sampling. We start going through β and σ-values in our prior distributions. For each pair we examine whether the observation could plausibly be a draw from the predictive y_i gaussian specified by that pair. We rule out all values for which this is not reasonable. We also re-weigh each remaining β, σ-pair by a combination of how probable we found it *a priori* and how plausibly exactly it could have generated the data we saw.

The outcome of this updating is a description of our new, *a posteriori* un-

Figure 4: Bayesian Workflow



certainty over the values of β and σ. These posterior distributions are superimposed in red on the blue prior distributions in the first and third bottom panel graphs of Figure 4. We can clearly see how the posterior has ruled out many values possible *a priori* but implausible given the data. The posterior trend line distribution in the second bottom panel further visualizes how the β-posterior corresponds to insisting that the true slope should resemble the downward trend of the observed dataset.

Notice that we can now return to ancestral sampling, just this time using our updated posterior distributions. The first and third bottom panel highlight four random β, σ-value pairs drawn from the posteriors. We can again plug each pair into the y_i gaussian and generate new predicted datasets. One such ten-item generated dataset is visualized for each β, σ-pair in the bottom-right panel of Figure 4. The distribution of these draws is known as the posterior predictive distribtuion.

The posterior predictive distribution is a famously useful tool in Bayesian statistics. It can obviously be used to predict the unknown. In the empirical sections of this thesis, for instance, I base my findings on counterfactual trajectories drawn from the posterior predictive distribution. However, the posterior predictive can also be used for model checking. If we use the model to predict some data we actually can observe, most of the observed data should fall within the spread of the posterior predictive distribution. A failure of this test suggests that the model is misspecified.

I hinted previously that the Bayes Theorem readily yields an expression for the posterior probability distribution. In practice, though, that expression is usually too complex for us to solve it exactly. Instead, Baeysian models are almost always estimated computationally, and that is how I derive all empirical findings in this thesis. The approach I use is referred to as Markov Chain Monte Carlo (MCMC). MCMC is a method of exploring a complex probability distribution by sending a sampler on a 'random walk' on its surface. Specifically, I use a pre-existing Python implementation of an MCMC sampler known as the No-U-Turns Sampler (NUTS) [Hoffman and Gelman, 2014], a modern extension to the famous Hamiltonian Monte Carlo (HMC) algorithm. For a more comprehensive review of MCMC sampling and of my computational implementation, please see the Appendix.

I now set aside discussion of general questions of Bayesian data analysis. Instead, I proceed to describe three intermediate steps to casting the Synthetic Control Method counterfactual prediction problem into this Bayesian paradigm of statistics.

3.2 Bayesian Dirichlet Regression

An obvious first step is to device a Bayesian regression that directly mirrors the SCM methodology. Namely, we can define a convex set of unknown weights $\boldsymbol{w}$ that are used to construct the target society from comparison societies as a weighted mean as follows:

$$y_{it}^{N} = \boldsymbol{y}_{-it}\boldsymbol{w}' + \varepsilon_{it} \tag{11}$$

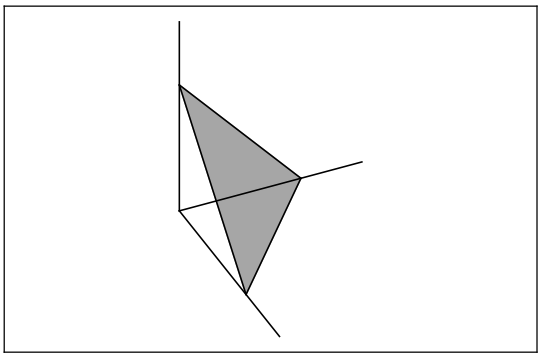
where we denote by y_{it}^{N} the non-treated outcome in society i and year t, by $\boldsymbol{y}_{-it}$ the vector of variable values in the other J societies, and by ε_{it} the year-society specific error term.

In the Bayesian setting, the prior distribution offers an obvious way of imposing the convexity constraint. Namely, we can set a prior that allocates zero probability mass to any coefficient values but those that form a convex combination.

The Beta is a well-known distribution that does so for just two coefficients. The multivariate generalization of Beta, the Dirichlet distribution, allows us to accommodate an arbitrary number of coefficients, so it is an appropriate choice for this problem. Supposing that we have no prior information about similarity between societies, we can specify a symmetric flat Dirichlet with the single concentration parameter set to one. This corresponds to setting a uniform prior on all possible weights. We can gain intuition to the shape of a uniform Dirichlet

by considering a 3-dimensional example such as the one in Figure 5. The value on each axis represents the value of one weight, and the shaded triangular plane section represents the uniform distribution on convex combinations.

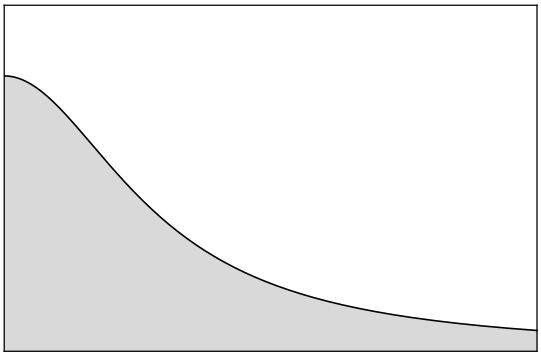
Figure 5: Three-Dimensional Dirichlet Distribution



We also need to assume a form for the error term ε_{it}. The SCM methods don't typically do so, but they always minimize some type of squared error. To achieve the same effect in a Bayesian model, we can give the term a normal prior with zero mean and an unknown standard deviation σ. Recent literature suggests that the Half-Cauchy distribution forms a good prior for standard deviation terms [Polson and Scott, 2012]. I include a visualization of the general Half-Cauchy shape in Figure 6.

If we collect all pre-intervention outcomes into the vector $\boldsymbol{y}_0^N$ and the matrix $\boldsymbol{y}_{-0}$, we can then specify the direct Bayesian SCM analogue thus:

Figure 6: Half-Cauchy Distribution



$$\boldsymbol{y}_0^N \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \sigma^2\right), \tag{12}$$

$$\boldsymbol{\mu}_0 = \boldsymbol{y}_{-0}\boldsymbol{w}', \tag{13}$$

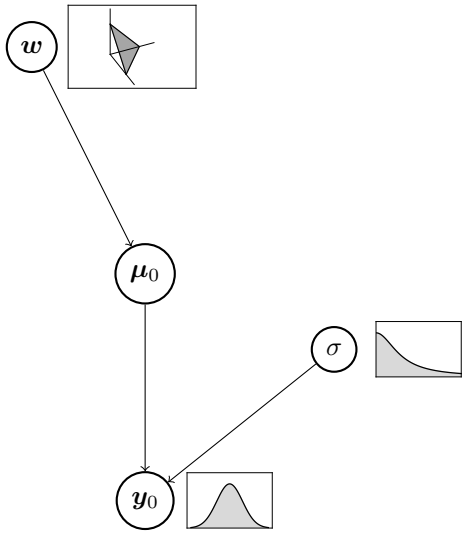
$$\sigma \sim \text{Half-Cauchy}\left(\gamma_\sigma\right), \tag{14}$$

$$\boldsymbol{w} \sim \text{Dirichlet}\left(\mathbf{1}\right) \tag{15}$$

where γ_σ is a known constant that reflects the analyst's prior belief about the scale of the error term's spread. The joint posterior distribution of $\boldsymbol{w}$ and σ can be used together with comparison society post-treatment outcomes to calculate the posterior predictive distribution of the target society's counterfactual post-treatment outcomes. The model is visualized in Figure 7.

This Bayesian SCM analogue has the ability to quantify its prediction uncertainty, and the Bayesian workflow protects it from much of the threat of overfit. However, like SCM, it is unable to capture negative correlations. Further, it is clearly not a model of any real-world causation. Whatever latent trends drive growth across societies make no appearance in the model specification. The

Figure 7: Bayesian Dirichlet Regression



next step addresses both of these issues.

3.3 Bayesian PCA Regression

The reason we think it reasonable to predict one society's outcomes from those of others is the underlying assumption that the same set of latent trends drives growth across all of them. A better approach is therefore to predict the target society directly from those trends. Borrowing a causal model from [Hsiao et al., 2012], we can set up the following formulation:

$$y_{it}^{N} = \boldsymbol{l}_t \boldsymbol{\beta}_i' + \kappa_i + \varepsilon_{it} \tag{16}$$

where $\boldsymbol{l}_t$ represents the state of some M inter-society trends in year t, $\boldsymbol{\beta}_i$ denotes a set of society-specific coefficients, and κ_i is a society-specific intercept. For notation, collect the annual $\boldsymbol{l}_t$ vectors for all years into the matrix $\boldsymbol{L}$.

Now that we have replaced society weights $\boldsymbol{w}$ with trend coefficients $\boldsymbol{\beta}_i$, there is no longer any need to impose a convexity constraint. Instead, we can set a more typical prior distribution such as the standard normal on each coefficient. Consequently the model no longer need discard information on negative correlations.

Of course, the issue with incorporating the latent trends $\boldsymbol{L}$ in the model is that they are latent, so we do not know their numeric values. An easy solution is to follow the example of GSC and replace the variable $\boldsymbol{L}$ with a point estimate $\hat{\boldsymbol{L}}$. One obvious way of doing so is to carry out a frequentist Principal Component Analysis (PCA) to derive a smaller set of M vectors. The PCA analysis can be calculated on comparison society data (so excluding for the target society) to derive the point estimate for all years.

Then we can describe the Bayesian model thus:

$$\boldsymbol{y}_0^N \sim \text{MvNormal}\left(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{I}\right), \tag{17}$$

$$\boldsymbol{\mu}_0 = \hat{\boldsymbol{L}}\boldsymbol{\beta}_0' + \boldsymbol{\kappa}_0, \tag{18}$$

$$\sigma \sim \text{Half-Cauchy}\left(\gamma_\sigma\right), \tag{19}$$

$$\boldsymbol{\beta}_0 \sim \text{MvNormal}\left(\mathbf{1}, \boldsymbol{I}\right), \tag{20}$$

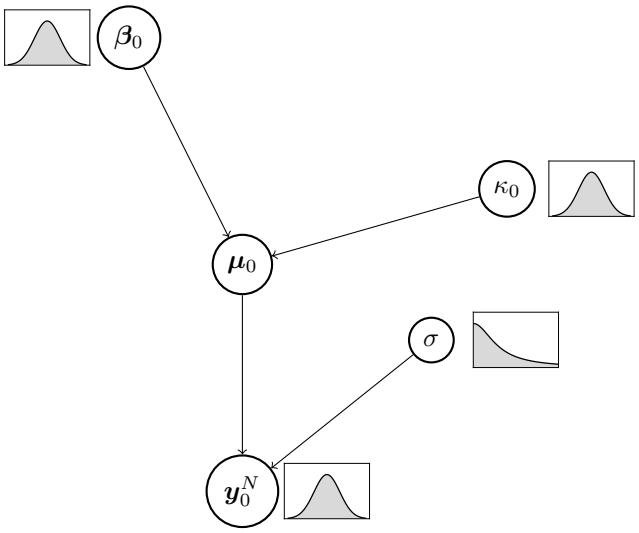
$$\boldsymbol{\kappa}_0 = \kappa_0 \mathbf{1}, \tag{21}$$

$$\kappa_0 \sim \mathcal{N}\left(\kappa^\mu, \kappa^{sd}\right) \tag{22}$$

where $\mathbf{1}$ denotes a vector of ones, $\boldsymbol{I}$ represents the identity matrix, MvNormal denotes the multivariate normal distribution, and κ^μ and κ^{sd} are known constants that describe the analyst's prior uncertainty about the target society's average outcome level in the pre-treatment period.

Again, we can fit the model in the pre-treatment period so that the joint

Figure 8: Bayesian PCA Regression



posterior distribution of $\boldsymbol{\beta}_0$ and σ can be used together with the post-treatment latent trend estimates $\hat{\boldsymbol{L}}$ to draw a posterior predictive distribution for the counterfactual estimate.

This approach finally models explicitly the relation between the target society and the laten trends. Further, it lacks the burdens of a convexity constraint. If any overfitting concerns remained previously, the dimensionality reduction here puts most of those to grave: one can look for spurious correlations from a large group of covariates but not from just a small handful.

However, this PCA regression is not really a satisfying Bayesian model. Most importantly, the core insight of the Bayesian paradigm is that we should not replace variables with their point estimates. To replace $\boldsymbol{L}$ with the PCA estimate $\hat{\boldsymbol{L}}$ is to understate drastically our uncertainty over a variable that we cannot observe. This motivates a more comprehensive latent variable model.

3.4 Bayesian Latent Factor Estimation

Let us now treat the latent trends $\boldsymbol{L}$ as unknown variables and estimate them from the data. This invites us to rewrite our model as one pertaining to the whole dataset rather than just the target society. To do so, denote by $\boldsymbol{Y}^N$, $\boldsymbol{K}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ the values of y_{it}, κ_i, $\boldsymbol{\beta}_i$, and $\boldsymbol{\varepsilon}_i$ collected into appropriately shaped matrices. Then we can describe the causal model thus:

$$\boldsymbol{Y}^N = \boldsymbol{K} + \boldsymbol{L}\boldsymbol{\beta}' + \boldsymbol{\varepsilon}. \tag{23}$$

Setting priors here is a slightly more involved process than previously. Note that several κ_i and $\boldsymbol{\beta}_i$ values are drawn, one for each society. Consequently, it would be appropriate to suppose all draws are from the same latent distribution and to estimate the parameters of that distribution explicitly. This practice, known as hierarchical modeling, provides insurance against outlier parameters.

Before I proceed, I want to draw attention to an important consequence of including the latent trends $\boldsymbol{L}$ as an explicit model variable. Denote the number of such latent trend variables by M. Any latent variable model of this kind must overcome an issue of rotational non-identitfiability.

To see what this means, note that the the M model variables are conceptually distinct from the M latent trends themselves. The former are features of the model, the latter features of reality (or at least of the data). Suppose momentarily that $M = 2$ and denote the model variables by 1 and 2 while denoting the real-world trends by A and B. Notice that the pairs can be matched with each other in either order: either $\{(A, 1), (B, 2)\}$ or $\{(A, 2), (B, 1)\}$.

These two matchings are perfectly equivalent and yield identical likelihoods. With neutral priors this leads to a multimodal posterior distribution with one

identical mode for each possible matching. The number of modes grows with the factorial $M!$ of the number of included latent trends. Unfortunately, multimodality is known to pose a serious challenge for the MCMC sampling approach I use to estimate my model computationally [Pompe et al., 2018]. Recall from section 3.1 that MCMC is based on sending a sampler on a 'random walk' around the posterior distribution. To work properly, the sampler must spend sufficient time in each neighborhood of the distribution and frequently traverse between them. This is difficult because the sampler tends to get stuck in one mode at a time for lengthy periods at a time.

I use a simple albeit imperfect way of resolving the rotational identification issue. Namely, I discard the idea of neutral prior distributions and instead use a narrow prior to pin each latent variable into the approximate shape of a unique underlying factor. This of course requires a well-justified way of picking a prior guess for each trajectory. I opt to do so using the frequentist technique of Principal Component Analysis (PCA). A PCA algorithm finds an optimal, or squared-error minimizing, linear basis for the dataset. PCA is an appealing choice because it turns out to be the maximum-likelihood estimate for the kind of a latent factor model I use [Bishop, 2006, p. 147]. I discuss the limitations of this pinning approach in section 7.

After setting these and other priors, we could ostensibly estimate this model as in the previous sections. Alas, a problem looms: we do not observe the values of $\boldsymbol{Y}^N$ for the target society post-intervention. Thus the likelihood function is undefined. The obvious solution would be to, as before, remove the post-treatment years from the model.

However, recall that we no longer have point estimates for $\boldsymbol{L}$ in the post-treatment period. Instead, it is this very model that is supposed to yield those estimates. Without them we have no use for the $\boldsymbol{\beta}$ estimates because we can

only predict the counterfactual using the two in tandem. Thus, we can no longer restrict ourselves to running the model for pre-treatment years only.

The other intuitive solution is to run the model without the target society so as to derive an estimate for the latent factors for the whole time period. However, then we would not get an estimate for the coefficients $\boldsymbol{\beta}_0$ used to derive the target society value from those trends. We could of course run a separate model to estimate those coefficients, but we would have to provide that second model with numeric values for $\boldsymbol{L}$. However, reducing $\boldsymbol{L}$ to a numeric point estimate is exactly what we are trying to avoid in this model.

Thus, we face the reality that the model in its current state can only be run if we exclude all data either for the target society or for the post-treatment years. Either exclusion renders the model useless for counterfactual estimation. The solution to this dilemma is the introduction of explicit treatment effect terms. In doing so, I lay out the full Bayesian Synthetic Control framework.

4 The Bayesian Synthetic Control Model

In this section I lay out a formal description of the Bayesian Synthetic Control model and its underlying modeling assumptions.

4.1 Structure

Consider a set of T years and J societies, and some quantitatively measured societal feature of interest. Focus first on the situation N where no policy intervention took place. Denote by y_{it}^N the value of the feature of interest in year t and society i. Suppose that the change over time in that quantity of interest is driven by a small number M of latent inter-society trends (where $M < J$), along with random noise. Suppose further that each such latent trend can be represented as a trajectory that takes a singular real value for each of the T years. Then we can associate with each year t a vector $\boldsymbol{l}_t$ of length M that captures the state of those latent inter-society trends. We can thus write that

$$y_{it}^N = \delta_t + \kappa_i + f_i(\boldsymbol{l}_t) + \varepsilon_{it} \tag{24}$$

where δ_t is an annual fixed effect, κ_i is a society-specific intercept, $f_i(\cdot)$ is some real-valued function, and ε_{it} represents random noise. Let us also suppose that $f_i(\cdot)$ takes a linear form. Then we can associate a coefficient vector $\boldsymbol{\beta}_i$ with each society i, and largely borrow the causal model of [Hsiao et al., 2012] to write equation 24 in this form:

$$y_{it}^N = \delta_t + \kappa_i + \boldsymbol{l}_t\boldsymbol{\beta}_i' + \varepsilon_{it}. \tag{25}$$

We can stack the individual parameters together into higher-dimensional vectors and matrices: let $\boldsymbol{\beta}$ denote the $J \times M$ transformation matrix of coefficients; let $\boldsymbol{L}$ denote the $T \times M$ matrix of intersociety trend values over time; let $\boldsymbol{\delta}$ denote the vector of length T of annual fixed effects; let $\boldsymbol{\Delta}$ denote the $T \times J$ matrix constructed by stacking $\boldsymbol{\delta}$ next to itself J times; let $\boldsymbol{\kappa}$ denote the vector of length J of society-specific intercepts; let $\boldsymbol{K}$ denote the $T \times J$ matrix that results when $\boldsymbol{\kappa}$ is stacked on top of itself T times; and let $\boldsymbol{\varepsilon}$ denote the $T \times J$ matrix of random noise. Finally, denote by $\boldsymbol{Y}^N$ the $T \times J$ dimensional matrix of all untreated outcomes. Then we can describe the whole system thus:

$$\boldsymbol{Y}^N = \boldsymbol{\Delta} + \boldsymbol{K} + \boldsymbol{L}\boldsymbol{\beta}' + \boldsymbol{\varepsilon}. \tag{26}$$

4.2 Reform Effect

Now suppose that some of the observations are distorted by the treatment effect of a policy intervention. Denote by α_{it} the treatment effect in society i and year t, and by y_{it}^I the outcome with the treatment effect. Then we can write that

$$y_{it}^I = y_{it}^N + \alpha_{it}. \tag{27}$$

Now let d_{it} be an indicator variable for whether the society i was treated in the year t, and let y_{it} denote the actual observed outcome. Then we can re-write equation 25 to include the effect of the intervention as follows:

$$y_{it} = \delta_t + \kappa_i + \boldsymbol{\beta}_i'\boldsymbol{l}_t + \alpha_{it}d_{it} + \varepsilon_{it}. \tag{28}$$

Let the treatment effect vary freely for each year and society. Denote by $\boldsymbol{D}$ the full $T \times J$ indicator matrix for the treated observations, and by $\boldsymbol{\alpha}$ the $T \times J$ matrix of associated treatment effects. In principle $\boldsymbol{\alpha}$ contains information also for the non-realized effects which the intervention would have had on the non-treated years and societies, but in practice its values only matter wherever the treatment is switched on in $\boldsymbol{D}$. We can then describe the full dynamics of the model thus:

$$\boldsymbol{Y} = \boldsymbol{\Delta} + \boldsymbol{K} + \boldsymbol{L}\boldsymbol{\beta}' + \boldsymbol{\alpha} \circ \boldsymbol{D} + \boldsymbol{\varepsilon} \tag{29}$$

where $\circ$ denotes the pointwise (Hadamard) product for matrices.

4.3 Estimation Goal

To estimate the parameters of the causal model in 29 we still need to specify a number of distributional assumptions and priors. Later parts of this section will do so in detail. However, it is already worth asking what exactly we aim to do with this model. Note that once we have set priors and fit the model with a dataset, we will land up with a full posterior probability distribution for each variable. What is the goal of setting up the model as it is and estimating these particular parameters?

Indeed, we would not usually be substantially interested in most of the variables listed above. Most of them, such as the intercept terms and all comparison society values, are only present to make estimating the model at large possible. There are really only two distributions we are genuinely interested in. The first one is the posterior predictive distribution of $\boldsymbol{Y}$, or rather the parts of it that relate to the target society in the treated years. The second one is the reform effect term $\boldsymbol{\alpha}$.

Using just these two distributions we can repeatedly draw a sample for $\boldsymbol{Y}$ and another one for $\boldsymbol{\alpha}$, and then subtract the latter from the former so as to get a draw of $\boldsymbol{Y} - \boldsymbol{\alpha} \circ \boldsymbol{D}$. This, of course, is by definition equal to $\boldsymbol{Y}^N$, the counterfactual trajectory. This demonstrates that estimating the equation 29 will ultimately allow us to draw samples from the posterior predictive distribution of the counterfactual trajectory of interest. These are indeed exactly the steps I use when calculating estimates in the applied sections of this thesis.

4.4 Comparison: SCM Motivating Model

The authors of [Hsiao et al., 2012] show that the model laid out in subsections 4.1 and 4.2 reflects directly the motivating model of the original SCM methodology. To see as much, recall its formal description as laid out in equation 6:

$$y_{it}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \varepsilon_{it}. \tag{30}$$

Now, let us constrain $\theta_{t1} = 1$, temporarily denote $\kappa_i \equiv Z_{i1}$, and redefine $\boldsymbol{\beta}_i$ and $\boldsymbol{l}_t$ as the following concatenations:

$$\boldsymbol{l}_t = (\theta_{t2}, \ldots, \theta_{tr}, \lambda_{t1}, \ldots, \lambda_{tF}), \tag{31}$$

$$\boldsymbol{\beta}_i = (Z_{i2}, \ldots, Z_{ir}, \mu_{i1}, \ldots, \mu_{iF}). \tag{32}$$

Then equation 30 simplifies back to equation 25:

$$y_{it}^N = \delta_t + \kappa_i + \boldsymbol{l}_t \boldsymbol{\beta}_i' + \varepsilon_{it}. \tag{33}$$

In other words, BSC and SCM share the same underlying model of causal structure. Both suppose that change over time is driven in a linear fashion by unobserved inter-society trends, each associated with society-specific exposures. The two differ only in terms of certain constrains imposed on the model. BSC fixes one of the year-specific variables to equal one, therefore introducing a geographic fixed effect. This society intercept is useful for BSC because it allows for treating the latent trends as zero-centered shapes. The same need doesn't arise in SCM because the estimator is actually calculated using the non-zero centered trajectories of other societies.

SCM, on the other hand, constrains some of the society-specific coefficients (the ones denoted by $\boldsymbol{Z}$) to take the observed values of some societal covariates. BSC imposes no such constraint, treating each societal coefficient as an unknown variable.

4.5 Distributional Assumptions

In order to set up a full probabilistic model, I need to set prior distributions for each of the model parameters. This is the crucial step that allows for Bayesian data analysis and for estimation of the posterior probability distribution. To do so in a rigorous fashion, I define some further model parameters.

4.5.1 White Noise

First, suppose that the random noise term ε_{it} follows a normal distribution with mean zero and an unknown standard deviation σ. Assume that σ is constant over societies and does not change over time. Let the value of σ have a Half-Cauchy prior distribution with the known scaling parameter γ_σ.

4.5.2 Annual Fixed Effect

Suppose that each annual fixed effect is drawn from a gaussian prior with known mean and standard deviation. Call the mean δ^{μ} and the standard deviation δ^{sd}.

4.5.3 Intercept

Suppose that each society-specific intercept κ_i is drawn from a normal distribution with unknown mean κ^{μ} and standard deviation κ^{sd}. Let the prior distribution of κ^{μ} be a gaussian with the known mean k_{μ} and standard deviation k_{sd}. For κ^{sd}, let the prior be Half-Cauchy with the known scaling parameter γ_{κ}.

4.5.4 Treatment Effect

Let each element α_{it} of the treatment effect matrix $\boldsymbol{\alpha}$ follow a gaussian prior with known mean α^{μ} and standard deviation α^{sd}. Recall from subsection 4.2 that we let the treatment effect vary freely by year and society. Thus, to prevent the estimation of one treatment effect from affecting the estimate for another, we don't allow any hierarchical structure for this prior.

It is extremely important to set α^{sd} equal to a very large number. The prior distribution should allocate almost identical probability densities for all reasonably sized intervention effects. Otherwise the model will prefer a particular effect size and adjust the $\boldsymbol{\beta}$ coefficients such that the counterfactual trajectory is estimated accordingly. This will make the model very sensitive to the treated trajectory, a strongly undesirable feature for a model aimed to estimate the counterfactual trajectory. It should pay little regard to the observations distorted by the intervention.

4.5.5 Transformation Coefficients

Let all country-specific coefficients β_{im} for the m-th latent component be drawn from a normal distribution with an unknown mean β_m^{μ} and standard deviation β_m^{sd}. For all $m = 1, \ldots, M$, let β_m^{μ} follow a gaussian prior with the known mean b_{μ} and standard deviation b_{sd}, and β_m^{sd} to have a Half-Cauchy prior with known scaling parameter γ_{β}.

Note that for the coefficients of each factor, this hierarchical structure has a regularizing effect across societies. All societies' coefficients are drawn from the same distribution so the model will estimate that distribution to have its mass wherever it is that most societies appear to fall by their likelihood. For countries that appear outliers from the group, the distribution will 'pull' their posteriors closer to the rest. Suppose for example that most societies' trajectories are constructed using a strongly positive coefficient on the m-th factor. Then the model would strongly penalize estimates that allocate a negative m-coefficient for some outlier society.

This effect acts as a soft constraint to prevent extrapolative prediction. At the same time it should be noted that there is no similar regularization across the M latent trends. The M coefficients of any particular society are estimated independently from each other.

4.5.6 Latent Factors

Let the value of the m-th latent trend in the year t have a Gaussian prior centered around the known mean p_{mt} with the trend-specific standard deviation r_m. For shorthand, denote by $\boldsymbol{P}$ the $T \times M$ matrix of these prior means, by $\boldsymbol{r}$ the vector of length M of standard deviations, and by $\boldsymbol{R}$ the $T \times M$ matrix that results when $\boldsymbol{r}$ is stacked on top of itself T times.

In principle it would be desirable to have $P_{mt_0} = 0$ and $r_{m_0} = r_{m_1}$ for all

t, m_0, m_1. This would correspond to a fully uninformative prior that treats each component identically. In practice, though, doing so yields a rotational identification issue that makes computational approximation much more difficult. Therefore in all the empirical sections of this thesis, I set informative component priors that differentiate between the components. I elaborate further on this issue in section 3.4.

4.6 Formal Model Specification

The contents of subsections 4.1 and 4.5 can be mathematically specified in a single probabilistic model. I do so in the equations 34 - 50. First, though, it bears explaining certain important features of the notation I use.

Most importantly, I distinguish between the single-variate (scalar) gaussian, denoted by $\mathcal{N}$, and the multivariate (vector) gaussian, denoted by MvNormal. The potential confusion roots to the fact that I frequently use the scalar gaussian with dimension subscripts like this: $\mathcal{N}_{X \times Z}$. In such case I include as parameter inputs a matrix of means and a matrix of variances (rather than a covariance matrix).

I refer with this notation to a collection of independent scalar normals arranged into the shape of a matrix. The (x, z)-th gaussian takes its mean and variance from the (x, z)-th elements of the mean and variance matrix, respectively, and yields its draw into the (x, z)-th cell of the output matrix. This corresponds exactly to the 'vectorized' behavior familiar to the users of software such as R and Python. At the same timem this is wholly distinct from the behavior of the multivariate gaussian which inputs a covariance matrix (rather than a matrix of variances) and yields a generally non-independent vector output.

Also consider some other points on notation. I denote vectors and matrices

by boldface characters ($\boldsymbol{\kappa}, \boldsymbol{Y}$) and scalars by regular characters (κ^{μ}, σ). Latin letters only denote matrices when in capital case and vectors (or scalars) when in small case, but this doesn't hold strictly for greek letters. All standalone vectors are treated as column vectors. I refer to a vector of X ones as $\mathbf{1}_{[\boldsymbol{X}]}$ and to a matrix of $X \times Z$ ones as $\mathbf{1}_{[\boldsymbol{X} \times \boldsymbol{Z}]}$, while $\boldsymbol{I}_{[\boldsymbol{X}]}$ refers to the $X \times X$ diagonal identity matrix.

I denote by $\circ$ the Hadamard product of two matrices of equal dimension. Hadamard multiplies the input matrices' cells elementwise. By $\otimes$ I refer to an outer product of two same-length vectors. This is the exact opposite of the dot product: if $\boldsymbol{x} \cdot \boldsymbol{z} = \boldsymbol{x}\boldsymbol{z}'$ yields a scalar, then $\boldsymbol{x} \otimes \boldsymbol{z} = \boldsymbol{x}'\boldsymbol{z}$ yields a square matrix. The outer product is used underneath exclusively for one purpose, that of stacking a vector repeatedly on top of (or next to) itself to form a matrix like thus: $\mathbf{1} \otimes \boldsymbol{x}$ (or $\boldsymbol{x} \otimes \mathbf{1}$).

Now we are finally ready to lay out in full the complete mathematical specification of the formal probabilistic model of the Bayesian Synthetic Control. The specificaion is quite lengthy, so I use hierarchical indentation to help the reader group up related equations. The expression for, or the distribution of, each named variable is independent of any other variables that are not indented underneath it. Thus, the BSC model can be formally expressed as beneath:

$$\boldsymbol{Y} \sim \mathcal{N}_{[T\times J]}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \tag{34}$$

$$\boldsymbol{\sigma} = \sigma \mathbf{1}_{[\boldsymbol{T}\times \boldsymbol{J}]}, \tag{35}$$

$$\sigma \sim \text{Half-Cauchy}(\gamma_\sigma), \tag{36}$$

$$\boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{\beta}' + \boldsymbol{\Delta} + \boldsymbol{K} + \boldsymbol{\alpha} \circ \boldsymbol{D}, \tag{37}$$

$$\boldsymbol{L} \sim \mathcal{N}_{[T\times M]}(\boldsymbol{P}, \boldsymbol{R}^2), \tag{38}$$

$$\boldsymbol{\alpha} \sim \mathcal{N}_{[T\times J]}\left(\boldsymbol{\alpha}^{\boldsymbol{\mu}}, (\boldsymbol{\alpha}^{\boldsymbol{sd}})^2\right), \tag{39}$$

$$\boldsymbol{\Delta} = \boldsymbol{\delta} \otimes \mathbf{1}_{[J]}, \tag{40}$$

$$\boldsymbol{\delta} \sim \text{MvNormal}\left(\delta^{\mu}\mathbf{1}_{[T]}, (\delta^{sd})^2 \boldsymbol{I}_{[T]}\right), \tag{41}$$

$$\boldsymbol{K} = \mathbf{1}_{[T]} \otimes \boldsymbol{\kappa}, \tag{42}$$

$$\boldsymbol{\kappa} \sim \text{MvNormal}\left(\kappa^{\mu}\mathbf{1}_{[J]}, (\kappa^{sd})^2 \boldsymbol{I}_{[J]}\right), \tag{43}$$

$$\kappa^{\mu} \sim \mathcal{N}\left(k_\mu, k_{sd}^2\right), \tag{44}$$

$$\kappa^{sd} \sim \text{Half-Cauchy}(\gamma_\kappa), \tag{45}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_{[J\times M]}\left(\boldsymbol{B}^{\boldsymbol{\mu}}, (\boldsymbol{B}^{\boldsymbol{sd}})^2\right), \tag{46}$$

$$\boldsymbol{B}^{\boldsymbol{\mu}} = \mathbf{1}_{[J]} \otimes \boldsymbol{\beta}^{\boldsymbol{\mu}}, \tag{47}$$

$$\boldsymbol{\beta}^{\boldsymbol{\mu}} = \text{MvNormal}\left(b_\mu \mathbf{1}_{[M]}, b_{sd}^2 \boldsymbol{I}_{[M]}\right), \tag{48}$$

$$\boldsymbol{B}^{\boldsymbol{sd}} = \mathbf{1}_{[J]} \otimes \boldsymbol{\beta}^{\boldsymbol{sd}}, \tag{49}$$

$$\boldsymbol{\beta}^{\boldsymbol{sd}} = \text{Half-Cauchy}_{[M]}(\gamma_b \mathbf{1}_{[M]}). \tag{50}$$

Line 34 above specifies the model's gaussian likelihood function, and lines 37 and 35 specify its mean and standard deviation inputs. The latter is not really a substantial equation, but rather a description of filling up a matrix with the single scalar σ. The prior of that variance term is expressed on line 36. Construction of the the mean term $\boldsymbol{\mu}$ is somewhat more convoluted. Indeed,

the rest of the lines 38 - 50 are devoted solely for that task.

Namely, line 38 specifies the prior of the latent trend variables, and line 39 does the same for the treatment effects. Line 41 defines the prior of the annual fixed effect vector. 40 contains notation of an outer product between $\boldsymbol{\delta}$ and a unit vector. This corresponds to the simple operation of stacking the $\boldsymbol{\delta}$-vector on top of itself repeatedly to form an appropriately shaped matrix.

Line 43 defines the prior distribution of the society-specific intercept vector. The (hyper)parameters of this distribution themselves have priors specified by 44 and 45. Again, line 42 simply stacks the intercept vector repeatedly into a matrix. Lines 46 - 50 work similarly: 46 states the prior distribution of the transition matrix of coefficients, 48 and 50 determine priors for the parameters of that distribution, and 47 and 49 describe repeated stacking of vectors.

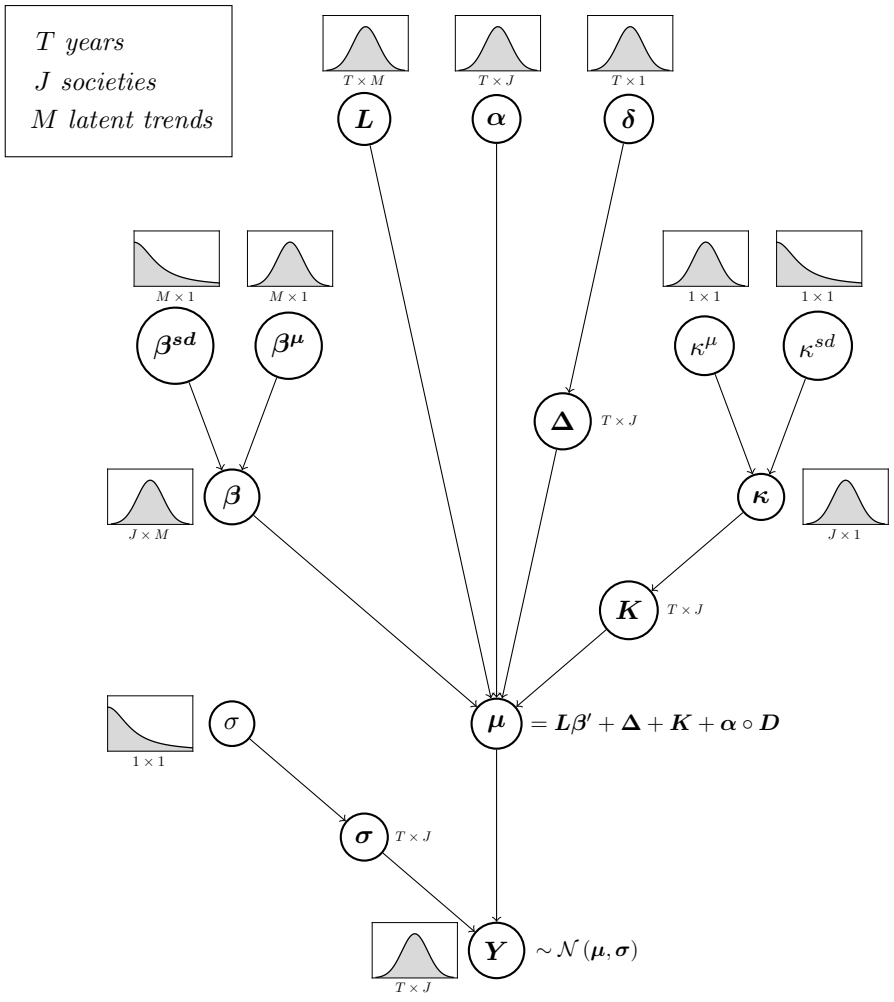
Taken together, the lines 34 - 49 can be used to precisely specify the BSC probabilistic model. However, any reader trying to develop intuition to the model's behavior will find them of little to no use. Therefore, before I move on to empirical applications, I visualize the abstract model and discuss briefly how it behaves.

4.7 Intuition

The BSC can be clearly visualized using the same directed graph notation I introduced in section 3. Figure 9 provides exactly this visualization. It pictures each model variable along with its dimensionality and the broad shape of its prior distribution.

While the full graph is fairly complex, we can ignore much of it and still capture well the core causal dynamic. First, we can immediately ignore the three intermediate variables $\boldsymbol{\sigma}$, $\boldsymbol{\Delta}$, and $\boldsymbol{K}$ placed at halfway points of the arrows. They reflect nothing beyond stacking vectors into matrices and add nothing to

Figure 9: The Bayesian Synthetic Control Model



intuition.

In fact, we need pay little attention to the whole two broad branches on the right, the $\boldsymbol{\delta}$ branch and the $\boldsymbol{\kappa}$ branch. The former represents a simple annual fixed effect, while the latter is a society-specific intercept (along with some hyperparameters). For broad intuition, it suffices for us to note that the model lets each year and society have its own average value. These intercept

variables add nothing further to the framework.

For the time being also set aside the treatment effect $\boldsymbol{\alpha}$. What remains is a representation of the four important components of the BSC model. First, there are some inter-society trends $\boldsymbol{L}$ that drive growth over time. Second, we associate with each society a set of coefficients or exposures $\boldsymbol{\beta}$ that determine how the society reacts to changes in $\boldsymbol{L}$. Third, the influence of $\boldsymbol{L}$ and $\boldsymbol{\beta}$ on the outcome variable $\boldsymbol{Y}$ can be fully summarized by an alteration in the state variable $\boldsymbol{\mu}$. Fourth, the observed data for all societies and years is distorted by white noise determined by σ.

We can impose one of two distinct interpretations on the whole these components form. The first one notes that a combination of unobserved trajectories, linear coefficients, and normally distributed noise is a well-studied one: it is that of a latent factor analysis, or of dimensionality reduction. From this point of view the core of BSC consists of doing Bayesian PCA to find a small-dimensional basis for the matrix of observed data.

The second interpretation points out that unlike in normal dimensionality reduction, in the typical BSC use-case the analyst really only cares about just one target society in the dataset. If the dataset is large, that one society cannot much impact the latent factor estimate. For that one society those latent factors are really just as external as a set of covarites would be. Given the data on other societies, the shape of $\boldsymbol{L}$ is fixed for the target society (even if only up to a probability distribution) and can be conditioned on. From this point of view, then, BSC is really just a linear regression where data on other societies is used to predict one target society

Mathematically, both interpretations hold true simultaneously. For the purposes of intuition, however, we can then conceptualize these two estimations happening in a sequence. First we estimate the latent trends using data on

other societies, and then we regress the target society on those trends. It all happens within one probabilistic model so we can do this without committing to a particular point estimate of the latent trend shapes.

Now recall from section 3.4 the problem we set off to solve in the first place. The data for the important target society years in $\boldsymbol{Y}$ is distorted by the treatment effect of a policy intervention. What we want to know is the relationship (as expressed in the $\boldsymbol{\beta}$ coefficients) between the non-treated target society outcome and the latent trends. It would be detrimental for that attempt if our estimate of that relationship were contaminated by the distorted data. This would happen if we naively regressed the observed target society outcomes on the estimated latent trends for the whole time period.

Indeed, this is the very issue that originally prevented us from running the two processes, latent variable estimation and target society regression, within a single model. To prevent contamination we needed to discard data either on the target society, preventing regression, or on the post-treatment years, preventing full latent variable estimation.

The BSC framework solves this issue by introducing the explicit treatment effect term $\boldsymbol{\alpha}$. Recall that $\boldsymbol{\alpha}$ varies freely for each post-treatment year in the target society. This has a profound impact on the model's behavior: likelihood of the observed data for the treated units becomes independent of the choice of $\boldsymbol{\beta}$. To see why, note that likelihood is determined by the distance between estimated $\boldsymbol{\mu}$ and the observed data. Now the model can use the treatment effect term $\boldsymbol{\alpha}$ to shift $\boldsymbol{\mu}$ from where $\boldsymbol{L\beta'}$ would have otherwise placed it to match exactly the observed data.

The distance between $\boldsymbol{\mu}$ and $\boldsymbol{Y}$ becomes chronically zero for all the treated outcomes. With that constant distance comes constant likelihood. As long as the prior of $\boldsymbol{\alpha}$ is sufficiently flat over all plausible values, the constant likelihood

translates directly into a constant posterior probability. In other words, the $\boldsymbol{\alpha}$ term makes all other model parameter estimates completely independent of observed data for the treated units. For all practical purposes, the model runs as if the data were missing altogether.

This is ultimately the whole purpose of the Bayesian Synthetic Control model: to carry out latent factor analysis and target society regression within one probabilistic model without contaminating parameter estimates with treated data. All other take-aways from Figure 9 are more or less secondary.

5 Application: German Re-Unification

5.1 Background

One of the important papers on the original SCM methodology investigate the impact of the German re-unification in 1990 [Abadie et al., 2015]. Namely, they focus their study on the effect on per capita incomes in the former West Germany. At the time of the re-unification, the West was markedly wealthier than the East, so the impact is speculated to have been negative. Indeed, the authors find that by 2003, West German per capita GDP would have been 12% higher without the reform. I implement the BSC framework to examine this same research question.

5.2 Data

I base my work on the dataset used by the original authors which is published online for replication purposes. The target variable is per capita GDP adjusted for purchasing power parity (PPP). The data is acquired from OECD National Accounts and, where necessary, Germany's Statistisches Bundesamt. The dataset also includes five other covariates useful for SCM, but they play no part in my BSC implementation.

The data covers 16 OECD countries. This includes all 23 member states from 1990 barring seven, which the authors discarded due to anomalous economic development. For consistency with the previous study, I use the same set of 16 countries. The time period covered is 1960-2003 of which the period 1990-2003 are considered treatment years for West Germany.

I make one alteration to the dataset. The original authors measure GDP in current US dollars rather than ones adjusted for inflation. Consequently, their variable grows at an arftifically high exponential rate. This exponential growth

behavior interacts poorly with my assumption that the white noise variance term σ^2 is constant over years. To see this, note that one would expect to see the magnitude of the random error to be more or less proportional to the absolute value of the variable.

To moderate this issue, I adjust the GDP per capita figures approximately for inflation and express them in constant 2003 US Dollars. To do so, I use the US GDP deflator time series which I record from World Bank's World Development Indicators [The World Bank, 2019].

5.3 Parameter Specification

To set the BSC framework up for computational estimation, I specify each of the model's hyperparameters.

5.3.1 White Noise

The prior distribution of the white noise parameter σ is Half-Cauchy, which only takes one scaling hyperparameter γ_σ. The Half-Cauchy distribution has an infamously fat tail, so the scaling parameter can be set with relatively little concern. I opt to set $\gamma_\sigma = 500$. This corresponds to saying that I believe, *a priori*, there to be $\frac{1}{2}$ probability that the country-year-specific noise term is drawn from a normal with a standard deviation less than USD 500.

5.3.2 Annual Fixed Effect

The annual fixed effect has a constant mean δ^μ for all years, which I set equal to zero. I suspect the annual fixed effect should never grow very large relative to the overall size of the typical society. However, the strength of that belief is moderated by the prospect that the annual fixed effect interact one way or another with the latent trends $\boldsymbol{L}$. To err on the side of ignorance, I thus set the

prior standard deviation $\delta^{sd} = 10000$.

5.3.3 Intercept

The intercept term indicates each country's mean income over the period 1960-2003. Setting it up requires specifying three hyperparameters. The unknown mean κ^{μ} is drawn from a gaussian with two hyperparameters: mean k_{μ} and standard deviation k_{sd}. The unknown standard deviation κ^{sd} is drawn from a Half-Cauchy with the scaling parameter γ_{κ}.

A priori, I believe the average of the country mean incomes should be lower than USD 30,000, a fairly typical Western per capita income around the turn of the millennium. I also believe that it should be higher than USD 6,000, a definite lower bound for most developed countries. To reflect this, I set $k_{\mu} = 18000$ and $k_{sd} = 6000$. I indicate similar uncertainty over the variance of country means by setting $\gamma_{\kappa} = 2,500$.

5.3.4 Treatment Effect

The year-country specific treatment effect α has a gaussian prior with the known mean α^{μ} and standard deviation α^{sd}. I set $\alpha^{\mu} = 0$ so as to not presuppose the sign of the effect. Also recall from section 4.5.4 how vital it is to set α^{sd} equal to some very large value. In other words, it is important to make this prior very uninformative. It should be almost flat over all even vaguely reasonable values. Preferring to err on the side of excessive flatness, I set $\alpha^{sd} = 30,000$.

5.3.5 Transformation Coefficients

Recall that the country-specific coefficient for each latent factor is drawn from the same gaussian prior with unknown mean and standard deviation. That mean itself has a gaussian hyperprior with mean b_{μ} and standard deviation b_{sd}, while the standard deviation has a Half-Cauchy hyperprior with the scaling

parameter γ_β. The scale of these coefficients is fundamentally linked to the scale of the latent factor trajectories with which they are multiplied to generate observed data. I set the priors of the latter to match the scale of the observed data. That requires coefficients in the unit neighborhood of zero, so that's where I fix the β-related priors: $b_\mu = 0$, $b_{sd} = 1$, and $\gamma_\beta = 1$.

5.3.6 Latent Factors

Recall from section 3.4 that the prior distribution of each latent factor trajectory is pinned around a frequentist estimate for one PCA component. This means that the latent factors L have a heavily informative, data-driven prior.

I carry out the PCA analysis using a pre-existing implementation based on singular-value decomposition from the `scipy` Python library. I fix the number of latent components at $M = 4$. The algorithm yields components centered around zero with variance similar in magnitude to those of the vectors in the observed dataset. The resulting components are stacked into a $T \times M$ matrix and the mean of the latent factor prior $\boldsymbol{P}$ is set equal to that matrix.

It is less clear what the appropriate standard deviation r_m is for each component. The frequentist PCA components vary in variance, so r_m should be related to the variance of the underlying component. I opt for direct proportionality where $r_m = \lambda \text{sd}_m^{pca}$. The remaining task then is to choose the λ multiplier.

The choice is important. If λ is too large, the prior will fail to pin each component to one model variable and the MCMC sampler will struggle to converge. If it is set too small, the model will further underestimate the uncertainty over the latent variable trajectories. In real terms I determined λ heuristically by attempting to run the sampler with a few obvious guesses. I concluded that $\lambda = 2$ is the largest integer for which the sampler consistently converges with my default chain size. Thus, my prior for a latent variable trajectory allows its value to vary by two standard deviations of the underlying PCA component's

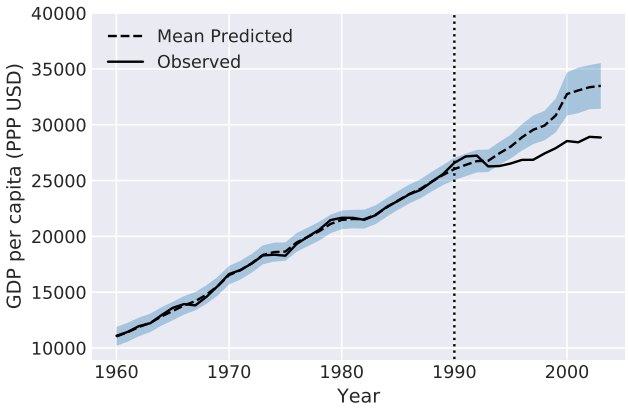
trajectory.

It should be noted that this specification of the L prior does not genuinely reflect my prior uncertainty over its value, but rather computational necessities. This has nontrivial implications for interpretation of my findings, and I return to them in more length in section 7.

5.4 Findings

Figure 10 summarizes the resulting BSC counterfactual estimate. It plots the mean BSC counterfactual trajectory estimate along with its 95% credible interval. The figure also includes the observed trajectory for comparison.

Figure 10: West German per capita GDP (2003 PPP USD)
Observation vs. BSC Counterfactual Estimate before and after Re-Unification



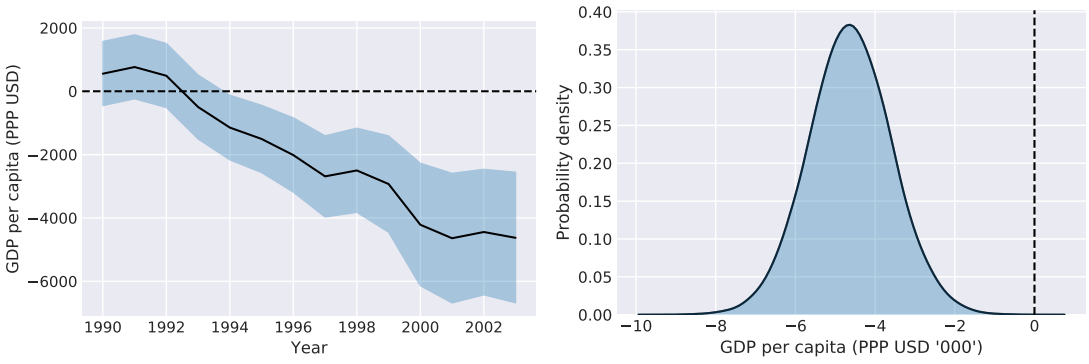
My findings are largely in line with those of [Abadie et al., 2015]. Namely, I find that the counterfactual growth trajectory doesn't much differ from the observed data in the first four years after the reform. Starting in 1994, however, the two trajectories begin to diverge substantially. By 2003, the observed GDP level is some USD 4,630, or 16.0%, below the mean predicted counterfactual value. This is equivalent to a fall in the average annual growth rate by 1.1

percentage points, from over 1.9% to just under 0.9%. This BSC-generated gap is slightly larger than the one predicted by SCM which is USD 3,360, or 11.7%, and corresponds to a 0.7 fall in the average growth rate.

Importantly, the observed data falls far outside the credible interval (CI) of the posterior predictive distribution from 1994 onwards. To be specific, the 95% CI of the cumulative treatment effect on per capita GDP by 2003 is USD 2,570 - 6,680. In other words, the BSC model indicates near-certain probability that the 1990 re-unification caused a substantial fall in West-German per capita GDP.

For an alternative visualization, we can also examine directly the posterior distribution of the treatment effect term. I provide two such graphs in Figure 11: the mean treatment effect estimate along with its 95%-CI on the left-hand side and the full posterior density of the cumulative treatment effect by 2003. The Figure further illustrates both how the treatment effect grew significant around 1994 and how strongly the model rejects the idea of a non-negative economic impact.

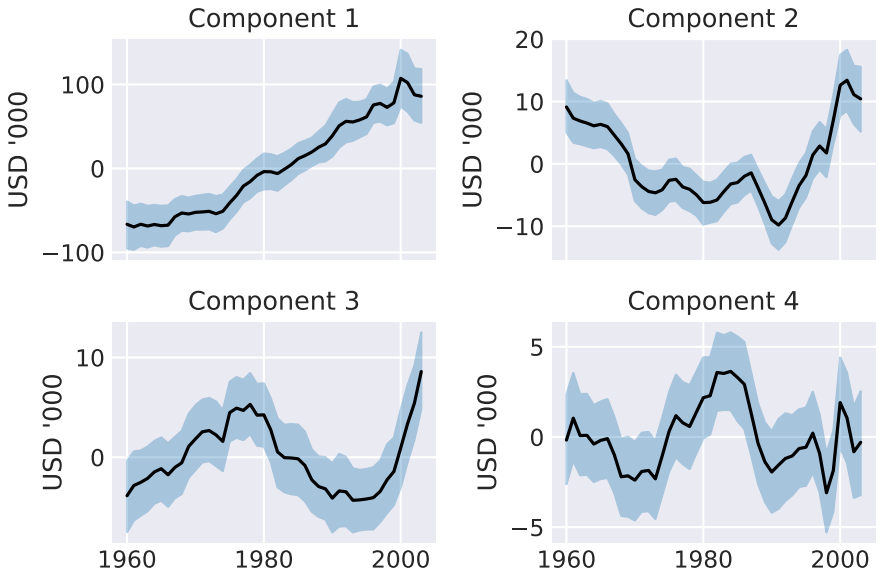
Figure 11: West German per capita GDP (2003 PPP USD)
Treatment Effect Posterior Distribution



The BSC model output also allows us to explore the posterior distributions for other interesting model variables. Consider, for instance, the distributions

for the latent factor trends, visualized in figure 12. Each mean trajectory is plotted along with its 95%-CI.

Figure 12: German Re-Unification
Latent Variable Posteriors

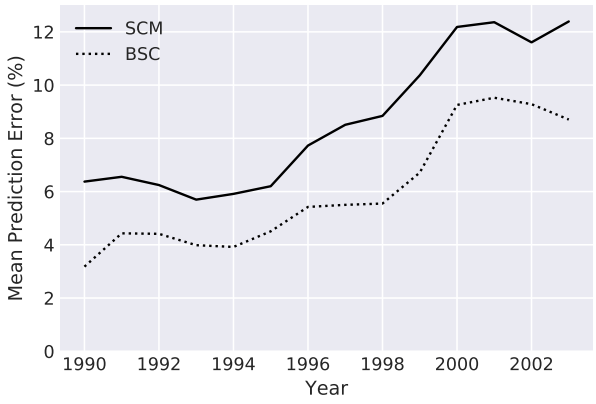


This visualization allows for qualitative interpretation of each latent factor. For instance, one can see that the first component reflects an overall growth trend that corresponds to global increase in per capita incomes over time. The other components are more difficult to read, but closer analysis would likely reveal correlation between these trajectories and important determinants of international growth trends, such as energy prices, military conflicts, and recessions.

5.5 Comparison to SCM

A re-labeling exercise provides an excellent opportunity to compare the BSC findings to those of SCM. To do so, I use each framework to predict in turn the observed post-treatment trajectory of each of the comparison societies. At

Figure 13: German Re-Unification, Accuracy Comparison



each run, I measure the distance between the prediction and the observation. To do so with BSC, I use the posterior predictive mean as my point estimate. I visualize the comparison society average of the error for each year and framework in figure 13. Overall, the findings show that the BSC exhibits greater predictive accuracy on this dataset.

5.6 Consistency Checking

BSC's ability to produce a full description of its prediction uncertainty has an important consequence: ability to test modeling assumptions. Recall that BSC, like any other statistical estimation strategy, is based on a set of strong modeling assumptions. Both the resulting point estimate and all measures of prediction uncertainty are valid only inasmuch as the modeling assumptions hold true. If the premises are violated, the results of the framework grow suspect in proportion to the scale of that violation.

In real terms these assumptions are bound to be more or less inaccurate. The important task is that of measuring how severe their violations are. One easy way to do so is to run a posterior predictive check. Namely, when the model

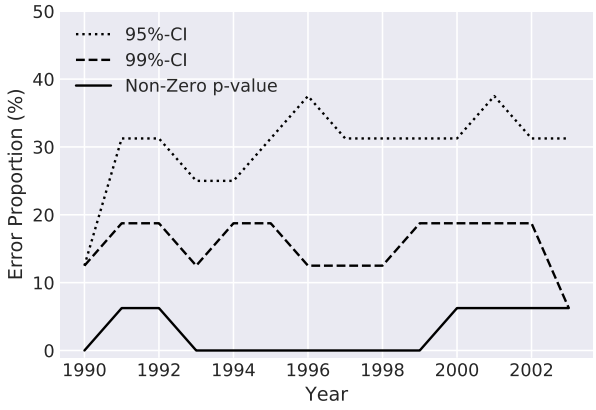
is used to predict observed trajectories in this dataset, the posterior predictive distribution should include the observed data. Posterior predictive checking is considered one of the best methods for model checking in the Bayesian context [Gelman and Shalizi, 2013]. Following the uncertainty quantification approach of [Abadie et al., 2015], [Xu, 2017], and [Ben-Michael et al., 2018] suggests an obvious way of doing so for BSC: run the model with re-labeling so as to predict in turn the "no German re-unification" counterfactual for each comparison society. If my modeling assumptions held true, the observed data would never fall far outside the spread of the resulting posterior predictive distribution.

Note also that this consistency check cannot be carried out for GSC, ASCM, or any other method that uses re-labeling to calculate confidence intervals. Instead, these methods artifically set their confidence bounds so as to include most of the observed data within the spread of uncertainty. This may hide warning signs of assumption violations, which is dangerous because both the point estimates and the confidence intervals are valid only to the extent that the assumptions hold up.

Figure 14 exhibits the results of this test. The dotted line indicates the share of the comparison societies for which the observed trajectory falls outside the 95%-CI of the BSC posterior predictive distribution. The dashed line plots the same measure for the 99%-CI. The dotted graph reflects the share of total prediction failures where the observed data is more extreme than any single draw from the estimated posterior predictive distribution or, put in othe words, receives the p-value of zero.

The share of predictions that falls within the 95%-CI starts at close to 90% in 1990, but soon falls to around two thirds and stabilizes at that level. The wider 99%-CI performs better, consistently capturing the observed data 80-90% of the time. Complete prediction failures are very infrequent, with only a handful

Figure 14: German Re-Unification, Posterior Predictive Check



of years seeing even one single occurrence. Perhaps surprisingly, none of the graphs demonstrates a clear upward trend over time.

This test clearly demonstrates a couple of points. First, the modeling assumptions are indeed violated. Second, the importance of those violations does not much depend on the timespan of the prediction, at least as long as it doesn't much exceed one decade.

At the same time, the results are not altogether hopeless. Even though the posterior predictive interval consistently includes the data less often than it should, it still does so most of the time. The 95%-CI succeeds more than two thirds of the time and the wider intervals perform better still. This finding does indicate that the modeling assumptions are somewhat accurate for the GDP per capita data in the OECD in 1960-2003. I discuss further this balance in section 7.

6 Application: California Tobacco Control Program

6.1 Background

Another important SCM paper examined the effect of a 1988 tobacco reform on California's smoking rate [Abadie et al., 2010]. The reform, known as Proposition 99, introduced sin tax hikes and other anti-smoking measures. California's smoking rate fell after the intervention but so did the national smoking rate and the rates of many other states. The authors of the 2010 study find that the reform's effect amounted to a 25% fall in cigarette sales. The study is famous for introducing the re-labeling based SCM significance testing. Its findings have been frequently re-analyzed [Ben-Michael et al., 2018]. I join in on this effort and study the same research question using BSC.

6.2 Data

The target variable for [Abadie et al., 2010] was the number of cigarette packs sold per capita according to tax data. As comparison societies they use a set of 38 other US states, or all such states that didn't introduce major tobacco controls of their own. The time period covered is 1970-2000, of which the years 1989-2000 are considered a treatment period for California. Again for consistency, I use the same selection of data and acquire it from a recent edition of the publication which the original authors used [Orzechowski and Walker, 2014]. As in section 5, I ignore the other covariates used for SCM, which in this case included income, age structure, cigarette prices, and beer consumption.

6.3 Parameter Specification

6.3.1 White Noise

Much of the US population was non-smokers throughout 1970-2000, and most smokers adjust their smoking rate relatively little year-to-year. Thus, I set a relatively conservative prior distribution on the white noise parameter σ. Namely, I set $\gamma_\sigma = 10$ to reflect that I imagine the white noise term to have a standard deviation less than ten packs per person with probability of one half.

6.3.2 Annual Fixed Effect

As in section 5, I set the mean of the annual fixed effect term equal to zero. To err on the side of ignorance, I nevertheless allow the prior standard deviation be quite large at $\delta^{sd} = 30$.

6.3.3 Intercept

I believe that the average of annual state smoking rates should be greater than zero but less than 365, or a daily pack for each person. To reflect this, I set $k_\mu = 180$, $k_{sd} = 90$, and $\gamma_\kappa = 90$.

6.3.4 Treatment Effect

As in section 5, I opt for a vastly uninformative treatment effect prior: $\alpha^\mu = 0$ and $\alpha^{sd} = 500$.

6.3.5 Transformation Coefficients and Latent Factors

I set transformation coefficient priors and latent variable priors exactly as in section 5. Namely, $b_\mu = 0$, $b_{sd} = 1$, and $\gamma_\beta = 1$ for the coefficients, and $\boldsymbol{P} = \mathbf{pca}$ and $\boldsymbol{r} = 2\mathbf{sd}^{pca}$ for the latent factors.

Unlike in section 5, I refrain from specifying the number of components M *a priori*. Instead, I carry out model selection to determine the most appropriate number, and ultimately implement the model wit $M = 6$. I present the details of this selection process in section 6.4.

6.4 Selecting the Number of Latent Factors

Recall that in section 5 I fixed the number of latent factors *a priori* at $M = 4$. That choice of M was ultimately arbitrary. The choice matters because increasing M helps the model explain more of the observed variance but also involves increasing the number of free model parameters, which introduces additional uncertainty. It seems reasonable to say that there should be some optimal M that finds the most desirable balance in this trade-off. The larger number of comparison societies in the California dataset emphasizes the importance of finding, or at least rigorously guessing, what that optimal value is.

One obvious way to select the best M is through formal model selection. Namely, we can run the model repeatedly with different values of M and record a measure of the model's performance at each round. The choice of that measure is not obvious but one popular option is a statistic known as the Watanabe-Akaike Information Criterion (WAIC). WAIC is known to be asymptotically equivalent to measuring the model's predictive accuracy with repeated cross-validation [Watanabe, 2010].

I implement this approach to selecting the optimal M for the California dataset. I begin with the *a priori* assertion that $M \in \{3, 4, 5, 6, 7, 8\}$, a set which I limit to be fairly small because larger values are computationally expensive. I run the model once for each value of M, using the parameterization from 6.3 and treating California as the target society. For each run, I calculate the WAIC measure using a pre-existing implementation from the `pymc3` Python library.

Figure 15: CA Prop. 99, WAIC Model Comparison

M	3	4	5	6	7	8
WAIC	7308	6834	6616	6538	6450	6326

The values are collected in table 15.

The WAIC value is decreasing in the number of latent components variables. Smaller WAIC indicates better predictive performance, so I choose the model with the largest number: $M = 8$.

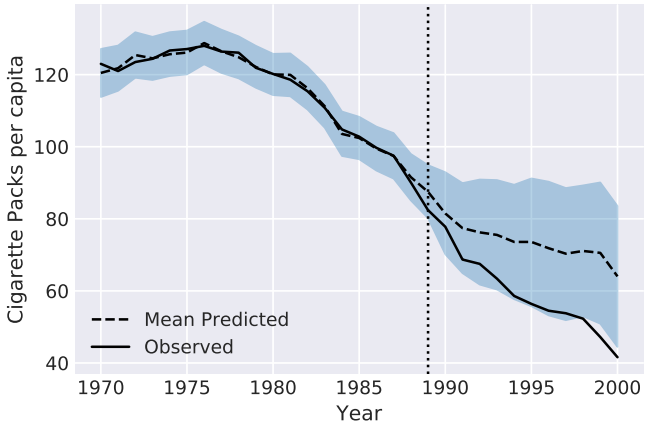
6.5 Findings

My core findings are captured in figure 16: it plots the mean predicted counterfactual trajectory and its 95%-CI along with the observed data. Like previous researchers, I find that the counterfactual trajectory falls slower than the observed smoking rate. The two trajectories don't diverge notably for the first couple of years after the reform. Starting around 1992, however, the gap begins to grow more substantial. By 2000, the predicted rate is 64.0 packs per person and almost 22.4 packs, or 54%, greater than the observed rate.

My findings are quite similar to those of [Abadie et al., 2015] when it comes to the scale of the treatment effect. I find that the the reform reduced smoking over the 1989-2000 period by 15.4 annual packs per person, or by 23%. The reported SCM estimate is slightly larger at approximately 25%. For further comparison, [Ben-Michael et al., 2018] report some predicted reform effects for the particular year 1997. The predictions are 26 for SCM and 20 or 13 packs per capita for two different Augmented SCM (ASCM) implementations. The BSC estimate is 16.5 packs, so substantially less than the SCM but halfway between the two ASCM estimates.

However, the observed post-treatment trajectory falls within the BSC 95%

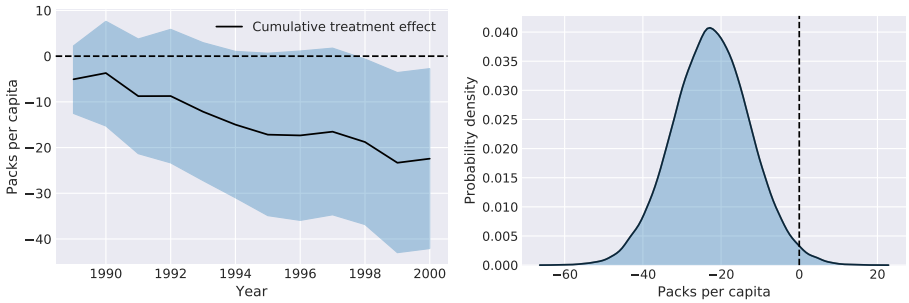
Figure 16: CA Proposition 99, Counterfactual and Observation



credible interval of my counterfactual estimate in almost all of the post-treatment years. This is clearly visible in figure 16 where the solid observation line remains well within the shaded CI region all the way into the late 90's. It is only for the three years starting in 1998 that the observed smoking rate is significantly lower than the prediction, and even then the observation is just barely outside the credible interval. In other words, for most of the studied period, BSC suggests that even without the reform there would have been over 5% probability of seeing a trajectory as far from the prediction or further still than the one we observe. That said, including the last three years, we conclude that the cumulative effect over the whole time period is indeed significant. The full posterior distribution of the cumulative effect is illustrated in Figure 17.

These significance findings differ somewhat from those of previous studies. The original SCM study [Abadie et al., 2015] found stronger evidence, with the cumulative treatment effect becoming significant as early as 1993. At the same time, [Ben-Michael et al., 2018] differs from both the original study and my BSC findings and concludes that the cumulative treatment effect was not

Figure 17: CA Proposition 99, Treatment Effect Posterior Distribtuion Mean with 95%-CI Over Time vs. Probability Density in 2000



even almost significant. They exhibit a frequentist confidence interval for each of their three tested frequentist SCM methods and show that the observation falls well within each of those intervals. (They use a two-standard error interval which corresponds to a 95.45%-CI but my significance findings are robust to adopting this wider interval.)

In conclusion, BSC's findings in the California tobacco control case are qualitatively similar to those of previous researchers. However, the prior frequentist methods disagree both with each other and with BSC on the exact point estimate for the treatment effect and on whether that effect was significant or not. These heterogenous findings demonstrate on one hand how much uncertainty counterfactual estimation often involves. On the other hand, they also emphasize the importance of the choice of methodology. If various synthetic controls always agreed, improving on them would make little difference. A borderline case like that of California's Proposition 99, however, demonstrates exactly how important it is to refine further counterfactual estimation methods.

7 Discussion

The Bayesian Synthetic Control framework has certain undeniable strengths as a statistical tool for the social sciences, but also some notable limitations. It can easily be extended to many more general types of problems, but there remains much space for further work and improvements on the framework. I discuss these considerations underneath.

7.1 Values

7.1.1 Uncertainty Quantification

Perhaps BSC's single most important improvement on prior work is its ability to produce full and probabilistically valid description of the uncertainty associated with its predictions. This is well exhibited in figures 10 and 16 where the mean counterfactual trajectory is surrounded by a full spread of other possible trajectories. In addition to visualizing this spread, we can describe it through credible intervals, the interquartile range, or any other measure appropriate for the research context.

This rich description of uncertainty makes BSC stand far apart from the original synthetic control methodology. SCM can at best yield a significance test for the sign of the error. In many impact assessment contexts, however, that does not suffice. To know that an intervention had an impact of the right sign is promising, but of little use in a cost-benefit analysis. BSC, instead, is perfectly suited for that type of assessment.

One particularly appealing feature of the resulting distribution of uncertainty is that it is in principle exact. It does not depend on asymptotic behavior such as infinite pre-treatment years or comparison socieities. Additionally, and in contrast to all prior frequentist work, the description of uncertainty is fully

model-based and avoids using up information generated by re-labeling. This property leaves re-labeling available for consistency checks. Consequently, and unlike the prior methods, BSC transparently reveals failures of the modeling assumptions.

7.1.2 Overfitting

I have previously pointed out how SCM, like many frequentist methods, is quite suspect to overfitting. Similarly, I've already noted (and explain further in the Appendix) that the issue is rarely of concern to Bayesian models. It does bear further emphasizing, though, that BSC is even better shielded from overfit than many other Bayesian models.

To see this, note that the SCM is supspect to overfit because the relatively large number of comparison societies allows the method to look for spurious correlations between them and the target society. BSC goes beyond the regular Bayesian approach to prevent this by first carrying out a dimensionality reduction from J to M. It is much more difficult to find a spuriously correlated linear combination of half a dozen than several dozen available covariates.

Finally, the hierarchical structure of BSC's prior distributions further moderate overfitting and extrapolation. Recall that the coefficients of all societies are drawn from the same distribution, and that distribution is estimated explicitly. This strongly regularizes against generating overfitted outlier estimates.

Admittedly, it is not easy to show empirically whether and how much one technque is more suspect to overfitting than another. However, we can take predictive performance for a tentative proxy of overfitting. The predictions of an overfitted model should be less accurate because they are distorted by random noise. As we saw in figure 14, BSC does indeed appear to produce more accurate predictions than SCM. This provides moderate empirical evidence for my claim of the overfitting concern.

7.1.3 Information Efficiency

Recall that the original SCM methodology discards certain types of information altogether. Namely, the method is unable to model negative correlation between societies. BSC, on the other hand, can utilize information on any linear inter-society trends. The loss of information associated with BSC is strictly tied to the dimensionality reduction used to estimate the latent trends. That loss, instead of prefering any particular type of information, extracts whatever patterns are most effective in explaining the observed variation. Further, SCM requires data on a number of covariates other than the variable of interest, while BSC analysis performs well on simple single-variate datasets.

7.1.4 Interpretability

An important part of BSC's appeal is that it is directly based on a causal model. In other words, it lays out a set of assumptions about causation in the real world and then estimates explicitly the associated variables. This again stands in fairly stark contrast to the original SCM. The SCM method motivates itself with a latent variable model, but those latent variables or their relation to the observed data is not estimated. Instead, the method estimates a relation between societies even while acknowledging that no causation runs between them in reality. SCM, then, is only motivated by an idea of causal stucture but cannot be interpreted as one. The same criticism applies directly to ASCM. Meanwhile GSC does estimate explicitly the underlying trends but fails to describe its uncertainty over their shapes.

7.2 Limitations

7.2.1 Assumption Violations

The BSC framework is ultimately based on a fairly strict set of assumptions. First, change in the variable of interest must be driven solely by inter-society trends along with a single clearly identified policy intervention. Second, the relation between those trends and the variable of interest must be linear. Third, the intervention has no effect on any of the comparison societies. Fourth, the white noise errors are independently drawn from a distribution that is constant over time and societies. In practice, these assumptions are frequently violated.

We saw evidence and implications of this type of behavior exhibited in figure 14. It shows that when BSC is run to predict a trajectory which by assumption was not affected by the German re-unification, the resulting credible interval does not consistently contain the observed data. By the end of the time period, a third of other societies were outside the 95% CI. Clearly, the model does not quite correctly capture the dynamics of per capita GDP growth in OECD countries.

It should be noted, though, that this limitation is shared by the previous frequentist tools, too. The shared characteristic of all SCM-related models is an assumption of latent, linearly combined causal factors. BSC's ability to describe its own uncertainty makes the violation of this assumption transparent. That SCM does not do so does not mean that its reliability is not compromised by those violations.

There is no obvious fix to this limitation. As long as it persists, researchers should be careful to run extensive checks for the validity of the linear factor model whenever using SCM-related tools. If figure 14 indicated an error rate that ascends slowly from 5% to 15%, we could take the model's predictions with

a fair degree of confidence. If the graph instead grew rapidly to the mid or high double digits, however, we should conclude that the predicted counterfactual trajectory is invalid. A pattern like the one currently seen suggests that the counterfactual estimate is useful, but should be interpreted with substantial reservations.

7.2.2 Number of Latent Variables

The structure and validity of the BSC probabilistic model in any particular practical context depends quite heavily on the number M of included latent variables. It is fairly obvious that my approach to choosing M in section 5, namely that of fixing it *a priori*, has few virtues beyond simplicity and ease of computation. The approach used in section 6, that of formal model comparison, is markedly more promising. However, even that approach is ultimately unsatisfying.

To see why, note that after the selection phase I fit the model with $M = 8$ as if I now knew for a fact 8 to be the correct number. In real terms, though, the true number of latent causal trends remains uncertain. The model output should reflect this uncertainty rather than hiding it as BSC currently does. A simple even if computionally expensive improvement would therefore be to give M an explicit prior and to include it in the model as an additional latent variable. Equivalently, the model could be run for several choices of M and the resulting posterior distributions could be combined using model averaging. Either approach would incorporate in the findings a non-trivial source of uncertainty that is currently hidden from the analyst.

7.2.3 Latent Variable Identification

The BSC implementation in its current form pins each of the latent factor trajectories into the neighborhood of one frequentist PCA component. Recall

that this is done to aid computation, not to reflect genuine prior information. This prevents the artificial multimodality resulting from rotational symmetry. However, it is possible the posterior distribution also has genuine multimodality. There could be a completely different set of trajectories that can be used to construct the observed dataset. The current informative prior structure prevents BSC from exploring such other modes.

Consequently, the model underestates our uncertainty over the shape of those latent trends. It is likely, even if not strictly necessary, that this also leads to some understating of uncertainty over the predicted counterfactual trajectory. Even where this understatement is small in scale, in marginal cases it might lead us to mistakenly place an observed trajectory outside a particular credible interval.

Unlike the issue described in section 7.2.1, this limitation is not inherent in the mathematical formulation of the Bayesian Synthetic Control. Instead it results from my preferred computational soluion. One simple improvement would be to run the model several times, each run with a different set of informative priors. For instance, another run could pin the latent factors using Independent Component Analysis (ICA) factors rather than PCA components. Results from the various runs could then be combined using a technique such as Bayesian model averaging.

More rigorously, the issue could be eradicated by removing the informative prior and allocating the sampler sufficient computational resources to explore the multimodal posterior distribution. Alternatively, the NUTS sampler could be replaced with another algorithm better prepared to deal with multimodality. In fact, the whole MCMC sampling approach could itself be replaced with another way of estimating the posterior, such as the variational inference framework.

7.2.4 Computational Resources

The BSC framework resembles many other Bayesian techniques in that its computationally much heavier than its frequentist relatives. A single run of the baseline BSC implementation as in sections 5 and 6 took 8-12 hours on a virtual Amazon Web Services Ubuntu machine equipped with substantial processing power. Tentative experimentation suggests that removing the informative latent variable priors might require upwards from 5 to 10-fold increase in computation time for sampler convergence even if $M = 4$. Furthermore, running the model for each of the comparison societies so as to produce the findings in figure 14 requires a J-fold increase in computation time. SCM and the related frequentist tools use less computation time by orders of magnitude.

7.3 Extensions

There are certain natural extensions that the BSC framework can easily accommodate. For instance, there is no reason why there should only be one target society. Rather, the indicator matrix D can be edited to mark any years in any number of societies as targets of the treatment effect. The BSC framework, therefore, could be used as a Bayesian alternative not only to the conventional SCM but also to other econometric techniques such at the canonical Difference-in-Differences method.

Similarly, BSC is also immediately able to accommodate missing comparison society data. The missing elements can be replaced with any reasonable dummy values and then marked as distorted in the treatment effect indicator matrix. Not only does this relax data quality requirements, but the model in fact outputs a rigorous estimate for the missing values as a side product. Indeed, BSC could even be used for the sole purppose of missing data imputation in any situation in the social sciences and beyond where latent linear variable structure can be

assumed.

BSC can also readily incorporate fairly substantial modifications of its underlying causal model. The inter-society relationship could be complemented with time-series behavior such as lagged outcomes or seasonality. We could relax the assumption of a linear latent variable structure and allow for nonlinear factor inputs. Fixed effect terms could be added or removed. Indeed, I included the annual fixed effect mostly for consistency with the causal structure of prior publications on synthetic controls. Over the long term, WAIC model comparison would likely form better grounds for variable inclusion than consistency with previous methodologies.

8 Conclusion

In this senior thesis I have laid out the design for a novel statistical tool for the social sciences, the Bayesian Synthetic Control. I have demonstrated that the BSC is well suited to study the kinds of research questions that have previously been investigated using the Synthetic Control Method and its frequentist extensions. Indeed, I have shown that BSC and SCM share a largely equivalent causal model. I have also demonstrated that the BSC framework lacks some of the flaws associated with SCM. The main improvement is quantification of uncertainty, which is accompanied by certain other strengths such as robustness to overfit concerns.

I have exhibited the framework's performance on two previously studied datasets and contrasted my findings with those of the original SCM methodology. The contrast showed that BSC, despite tending to agree with SCM's qualitative results, can further our understanding about the magnitude of the associated treatment effects, even to the point of casting doubt on the statistical significance of those findings. I have demonstrated that BSC outperforms SCM in a simple accuracy test when predicting untreated placebo trajectories. I have also showed that BSC's method of uncertainty quantification can yield evidence of model mis-specification where prior frequentist work fails to do so. Together these demonstrations show that the framework is ready for implementation and use in practical social scientific research contexts.

Nevertheless, BSC continues to have notable limitations. While these limitations are often shared by the related frequentist tools, their persistence emphasizes that researchers should exercise great caution when using the various synthetic control tools. It also exhibits the amount of further work that remains to be done in translating tools of causal inference into the Bayesian paradigm. The

study of treatment effects has long been dominated by the frequentist paradigm of statistics. This trend may be ripe for reversal and tools like BSC may play a part in doing so.

References

[Abadie et al., 2010] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

[Abadie et al., 2015] Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

[Abadie and Gardeazabal, 2003] Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132.

[Arnold and Stadelman-Steffen, 2017] Arnold, T. and Stadelman-Steffen, I. (2017). How federalism influences welfare spending: Belgium federalism reform through the perspective of the synthetic control method. *European Journal of Political Research*, 56(3):680–702.

[Aytuğ et al., 2017] Aytuğ, H., Kütük, M. M., Oduncu, A., and Togan, S. (2017). Twenty years of the euturkey customs union: A synthetic control method analysis. *Journal of Common Market Studies*, 55(3):419–431.

[Barlow, 2018] Barlow, P. (2018). Does trade liberalization reduce child mortality in low- and middle-income countries? a synthetic control analysis of 36 policy experiments, 1963-2005. *Social Science & Medicine*, 205:107–115.

[Ben-Michael et al., 2018] Ben-Michael, E., Feller, A., and Rothstein, J. (2018). The augmented synthetic control method. Arxiv open access.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer Science.

[Brodersen et al., 2015] Brodersen, K., Gallusser, F., Koehler, J., Remy, N., and Scott, S. (2015). Inferring causal impact using bayesian structural time-series models. *Annals Of Applied Statistics*, 9(1):247–274.

[Doudchenko and Imbens, 2017] Doudchenko, N. and Imbens, G. W. (2017). Difference-in-differences and synthetic control methods: A synthesis. Arxiv Open Access.

[Ferman and Pinto, 2016] Ferman, B. and Pinto, C. (2016). Revisiting the synthetic control estimator. Open Access.

[Gelman and Shalizi, 2013] Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

[Hoffman and Gelman, 2014] Hoffman, M. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal Of Machine Learning Research*, 15:1593–1623.

[Hsiao et al., 2012] Hsiao, C., Ching, S., and Wan, S. K. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong and mainland china. *Journal of Applied Econometrics*, 27(5):705–740.

[Karlsson and Pichler, 2015] Karlsson, M. and Pichler, S. (2015). Demographic consequences of hiv. *Journal of Population Economics*, 28(4):1097–1135.

[Orzechowski and Walker, 2014] Orzechowski and Walker (2014). *The Tax Burden on Tobacco. Historical Compilation*, 49 edition.

[Polson and Scott, 2012] Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

[Pompe et al., 2018] Pompe, E., Holmes, C., and Łatuszyński, K. (2018). A framework for adaptive mcmc targeting multimodal distributions. Arxiv open access.

[The World Bank, 2019] The World Bank (2019). World development indicators.

[Wan et al., 2018] Wan, S.-K., Xie, Y., and Hsiao, C. (2018). Panel data approach vs synthetic control method. *Economics Letters*, 164:121–123.

[Watanabe, 2010] Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal Of Machine Learning Research*, 11:3571–3594.

[Xu, 2017] Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Appendix: Review of Bayesian Data Analysis

Most work on synthetic controls has been carried out in the frequentist paradigm of statistics. Indeed, the same is more or less true of the broader field of causal inference in the social sciences. In this Appendix, I review the workflow of Bayesian data analysis, discuss briefly its strengths and weaknesses, and elaborate on the MCMC sampling approach to implementing it computationally.

Paradigm Differences

The core difference between the two paradigms of data analysis concerns the treatment of unknown parameters. In the frequentist workflow, only observed data is considered to consist of random variables. Model parameters, such as covariate coefficients in a regression, are considered unknown constants. They are not taken to be in any sense random.

This distinction roots directly to the frequentist interpretation of probability as a long-term frequency. Namely, the analyst supposes that the observed data was generated by some process involving randomness. Running that process again would yield different data. Running it repeatedly for long enough would generate data that converged to some population distribution. Our observation is one draw from this distribution and thus properly random. The unobserved parameters which define that process, on the other hand, do not themselves have a generating random process. They are simply constants, even if unknown ones. You cannot assign a probability distribution to the value of an unknown parameter any more than you can assign a probability distribution to the value of pi.

In contrast, a Bayesian analyst treats all unknown quantities as random variables, whether they be measurable outcomes or abstract model parameters.

This reflects a different understanding of probability. Namely, Bayesians consider probability to be a measure of uncertainty. Anything that is unobserved is also uncertain, and we can (or even should) describe it using probabilities. To do so, each model parameter is assigned a prior probability distribution to describe the analyst's *a priori* uncertainty over its value. This prior uncertainty is then updated using the observed data to derive a probability distribution to describe the analyst's *a posteriori* uncertainty over that value.

Estimation Goals

The two paradigms yield different approaches to data analysis. The frequentist paradigm has a particular limitation when it comes to discussing the values of an unknown parameter. It can only do so in terms of the likelihood, or the probability of seeing the observed data given those values: $p(\boldsymbol{y}|\boldsymbol{\theta})$. Point estimates are selected through some type of optimization, often by maximizing the likelihood. Construction of a confidence interval, whenever possible, is done by estimating how extreme the parameter could be without the likelihood falling very low.

In other words, for a parameter $\boldsymbol{\theta}$, observed data $\boldsymbol{y}$, and supposed probability function p, a frequentist analysis looks for a point estimate in this fashion:

$$\hat{\boldsymbol{\theta}}_{freq} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{y}|\boldsymbol{\theta}). \tag{51}$$

Bayesian analysis does not much concern itself with point estimates. Instead, it uses the Bayes theorem together with prior probabilities $p(\boldsymbol{y})$ and the likelihood function to derive a posterior probability distribution. Given full confidence in the model in use, the posterior offers a full description of the analyst's *a posteriori* uncertainty over parameter values. Formally, then, a Bayesian

analysis seeks this new probability function:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{52}$$

$$\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{53}$$

Note that this approach does not commit itself to a particular estimated value for the unknown parameters. Rather, the the output is a full distribution of probable values. Any predictions are made by integrating over all plausible parameter values, weighing each by its posterior probability.

We are also often interested in the posterior predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{y})$ of some unseen data $\boldsymbol{y}^*$. The posterior predictive describes the kinds of data we would expect to see given the data that we actually did observe. In the case of social scientific counterfactual estimation, we can specifically derive it for the unobserved trajectory we would have seen without the treatment effect. When individual draws of data are independent, the predictive has a fairly straightforward integral representation:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \int p(\boldsymbol{y}^*|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{54}$$

$$= \int p(\boldsymbol{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{55}$$

$$\propto \int p(\boldsymbol{y}^*|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{56}$$

In practice both 53 and 56 are too complex to be calculated exactly. However, each of the components $p(\boldsymbol{y}|\boldsymbol{\theta})$, $p(\boldsymbol{y}^*|\boldsymbol{\theta})$, and $p(\boldsymbol{\theta})$ tends to be easy to compute. In practice Bayesian posteriors are therefore estimated computationally. Perhaps the most well-established method of doing so is Markov Chain

Monte Carlo.

Markov Chain Monte Carlo

One often-used way to estimate a probability distribution is to draw samples from it. If we can draw a large sample from a distribution, we can derive robust estimates for most of its features. Means, standard deviations, and quintile functions are only some of the many sample statistics that converge to population values as the sample size grows. Samples make it particularly easy to answer questions such as "how probable is it that $\theta > 0$?" We can simply count the proportion of draws where that is so. With large enough sample size, the sample proportion will have converged to the population proportion.

MCMC refers to a group of algorithms designed to draw samples from complex probability distributions. To do so, an MCMC algorithm sends a sampler on a 'random walk' (Markov chain) around the parameter space. At each step, the sampler records its current location as a draw from the distribution and then jumps into a new stochastically chosen location in the parameter space. The exact method of determining where the jump goes varies by algorithm. However, the method always guarantees that the targeted posterior probability distribution is also the stationary distribution of the sampler's random walk.

Stationary distribution here refers to a type of limiting behavior. The share of draws recorded in any neighborhood of the parameter space converges to the share of probability mass allocated to that neighborhood by the stationary distribution. In other words, as the chain length grows, the collection of values recorded by the Markov chain converges into a sample drawn from the stationary distribution.

Therefore, given enough time and processing power, any MCMC algorithm is guaranteed to produce a random draw from the targeted posterior probability

distribution. The only required input is a method to compute that distribution's density at any point in the parameter space. In practice, though, available computation time is limited so the choice of an MCMC algorithm matters.

For a Markov chain to converge to the stationary distribution, it needs to spend plenty of time in all plausible neighborhoods of the parameter space. Therefore it is important that the sampler moves from one neighborhood to another fairly quickly. This requires it to take large enough steps during its random walk. A popular algorithm to help the sampler increase its step size is one known as Hamiltonian Monte Carlo (HMC).

Unlike many simpler algorithms, HMC uses information on the gradient of the underlying distribution to inform its next step. This allows it to move around the posterior more quickly, leading to faster convergence. Being a more complex algorithm, however, HMC is very sensitive to user-provided configuration specifications. This sensitivity motivated the recently quite popular No U-Turns Sampler (NUTS). The underlying structure of NUTS is merely that of HMC, but it carries out automatic, adaptive selection of appropriate sampling configurations. For full technical reference, I implemented the BSC model using pre-existing implementation of the NUTS sampler from the `pymc3` Python library. I ran the sampler with target acceptance rate of 0.9, maximum tree depth of 12, two simultaneous Markov chains, 5,000 tune-in steps per chain, and 20,000 sampling steps per chain. The results were checked for convergence using the Gelman-Rubin diagnostic and the lack of sampler divergences. Other Python libraries vital for my code included `numpy`, `pandas`, `theano`, `sklearn`, `matplotlib`, and `seaborn`.

Overfitting

Overfitting is a term used to describe a particular flaw in statistical analysis. In essence, an overfitted analysis yields results too specific to the sample it was trained on. It captures not only genuine patterns in the data but also random noise. An overfitted model will try to reconstruct that specific instance of white noise when used to make predictions on previously unseen data. Because random noise does not in fact reproduce itself in new data, overfitting generally worsens the predictive performance of a statistical analysis.

The roots of overfitting are largely in the optimization approach used in frequentist statistics. Recall that a frequentist analysis selects a single point estimate for $\boldsymbol{\theta}$. It is selected to be whatever value minimizes an error on the training data. Random noise, not only genuine patters, impacts the size of the error. Whenever there are many choices of $\boldsymbol{\theta}$ that produce fairly small errors, the exact choice among them is thus often determined by white noise. The threat of this grows as the number of the model's free parameters (eg. the number of covariates) increases.

Bayesian models, on the other hand, are by nature quite immune to parameter overfitting. Recall that a Bayesian analysis does not involve any optimization, nor does the model commit itself to any particular choice of parameter values. Instead, the estimate is constructed by integrating over all posssible parameter values. Each value is weighted by its posterior probability. Thus all values that produce a fairly small error in training data and have a fairly large priror probability also have a fair degree of impact on the prediction. This means predictions are based on averaging over many different values of random noise, which eliminates most of the threat of overfitting.

It should be noted that there are frequentist methods for constraining overfit. Some of the the famous ones include regularization terms and extensive cross-

validation. However, they do not fully diminish the overfit case for the Bayesian paradigm. This is in part because they all ultimately commit themselves to a particular choice of parameter values rather than aggregating over all or many plausible ones.

Prediction Uncertainty

Statistical models rarely if ever aim to yield exactly correct predictions. Instead, a prediction aims to be a good estimate of the unknown true value. As such, the prediction is uncertain. Quantifying that uncertainty is a core task for any serious statistical analysis. Such quantification usually includes an overall measure of confidence in the sign of the prediction (eg. a significance test) and either a measure of the expected magnitude of the error (eg. a standard error) or a description of the interval or region where the true value is expected to be (eg. a confidence interval).

As discussed before, a frequentist analysis cannot describe its uncertainty with probability statements related to the unknown parameters. Instead, it depends on well-studied formulae for significance tests and confidence intervals. These formulae, when applicable, are guaranteed to be consistent with the true paradigm values a specified proportion (usually 95%) of the time they are implemented on new datasets. However, there is no guarantee that the conventional formulae are applicable to any particular newly introduced estimator with an unknown distribution (such as the conventional SCM). Furthermore, frequentist confidence intervals are infamously convoluted to interpret. Unlike the name might suggest, the 95% confidence interval does not contain the true value with 95% probability. Rather, the associated confidence is strictly confidence in the method of computing the interval, not that in the true parameter value.

A Bayesian model, on the other hand, does not depend on the applicability of

particular confidence interval or standard error formulae. Recall that a Bayesian analysis estimates directly the posterior predictive distribution. In other words, the analysis output is a distribution that describes explicitly the analyst's full posterior uncertainty over the counterfactual trajectory. Once we know this distribution, we can readily discern measures like its standard deviation and any desired credible interval. The claims made about the paramter are probabilistic and thus trivially easy to interpret.

Measurement of prediction uncertainty, therefore, may occasionally present a major reason to prefer the Bayesian workflow. It more readily and reliably captures the spread of possible outcomes along with the respective probabilities.

Summary

The Bayesian and frequentist paradigms of statistics are characterized by substantially different workflows, strengths, and weaknesses. Previous work in synthetic controls has been overwhelminly dominated by the frequentist paradigm. Many of its more striking features, such as discarding of information and challenges in quantification of information, root directly to fundamental features of the underlying paradigm.

It seems clear that addressing the problem using the Bayesian workflow has substantial promise of improvement on past work. The issues of overfit and prediction uncertainty are of little to no concern for a robust Baysian modeling exercise. Interpretability, too, has proven an issue in the current frequentist approaches and may be best addressed with a paradigm shift.