Proof

A State Space Model Approach with Bayesian Shrinkage

Danny Klinenberg*

Last Updated: 2020-08-06

Abstract

This is my inference section/making a proof. I have spent most of June focusing on this section and it is driving me towards insanity.

1 Setup

Let $Y_{j,t}(0)$ and $Y_{j,t}(1)$ represent potential outcomes in the presence and absence of a treatment with $t = 1, ..., T_0 - 1, T_0, T_0 + 1, ...T$ and $j = 0, 1, ..., J$. Suppose only one unit, $j = 0$ is treated beginning in period $t = T_0$ and remains treated for all $t \geq T_0$. Suppose the other J units are unaffected by the treatment (i.e. no spillover effects). The researcher observes $Y_{j,t} = (1 - D_{j,t})y_{j,t}(0) + D_{j,t}y_{j,t}(1)$ where $D_{j,t} = I(j = 0, t \geq T_0)$. The goal is to create a fake, or synthetic,

Let $y_{0,t}(w_{0,t})$ denote the outcome of the treated unit in period t with treatment status $w_{0,t}$. Let $x_{i,t}(w_{i,t})$ denote the outcome of control unit i in period t with treatment status $w_{i,t}$ Assume the researcher observes $Y_{0,t} = (1 - w_{0,t})y_{0,t}(0) + w_{0,t}y_{0,t}(1)$. The researcher also observes $X_{j,t} = x_{j,t}(0)$ for all $j = 1, ..., J$ and $t = 1, ...T$, implying the control units are never treated. Define X to be the matrix of untreated units.

*University of California, Santa Barbara

2 Model Specification

This paper opts to estimate $y_{0,T_0+i}(0)$ using a state space model. This approach is widely used in time series analysis due to the flexibility in model specification. I define the model as:

$$y_{0,t}(0) = \sum_{j=1}^{J+1} \beta_{j,t} X_{j,t} + \epsilon_t \qquad \epsilon_t|\sigma^2 \sim N(0,\sigma^2) \tag{1}$$

$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t} \qquad \eta_{j,t} \sim N(0,\theta_j) \quad \forall j \tag{2}$$

$$\beta_{j,0} \sim N(\beta_j, \theta_j P_{jj}) \qquad \forall j \tag{3}$$

$$\pi(v) \tag{4}$$

where $X_{J+1,t} = 1$ for all t (intercept). The parameters of the model are $v = \{\sigma^2, \beta_1, ...\beta_{J+1}, \theta_1, ...\theta_{J+1}\}$ where $\pi(v)$ represents the prior distribution of the parameters. The hyperparameters are P_{jj} for $j \in \{1, 2, ..., J+1\}$.

In general state space form, the model is written as:

$$\begin{aligned} y_t &= X_t \iota_t + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0,\sigma^2) \quad \text{observation equation} \\ \iota_{t+1} &= T_t \iota_t + R_t \eta_t \quad \eta_t \sim \mathcal{N}(0,Q) \quad \text{state equation} \\ \iota_0 &\sim \mathcal{N}(a_0, P_0) \end{aligned}$$

where

$$T_t = I, \quad R_t = I,$$

$$P_0 = diag\left[\theta_1 P_{11}, ..., \theta_{J+1} P_{J+1,J+1}\right],$$

$$Q = diag\left[\theta_1, ..., \theta_{J+1}\right],$$

$$\iota_t = \begin{bmatrix} \beta_{1,t} \\ \vdots \\ \beta_{J+1,t} \end{bmatrix}, \quad \iota_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{J+1} \end{bmatrix}$$

where $diag(*)$ represents a diagonal matrix with the specified elements on the diagonal.

Notice that the evolution of $\beta_{j,t}$ can be changed by changing T_t. for example, setting $T_t = diag[\rho_1, \dots, \rho_{J+1}]$ creates an AR(1) process. In addition, adding off diagonal terms add interdependence between the evolution of the coefficients. Setting $\beta_{j,t} = \beta_j$ and allowing for a local linear time trend will yield the original estimator in Brodersen et al. (2015). The reader is referred to Durbin and Koopman (2012) for an advanced treatment of state space modeling.

3 Proof Motivation

Why This?

I chose to prove asymptotic unbiasedness because it is commonly proven in a synthetic control framework. In addition, Samartsidis et al. (2019) reference that the Causal Impact method is asymptotically unbiased but do not prove the result. All they say is:

> If the CIM of equations (3.7) is the true data-generating model and the prior on the vector of model parameters $\psi = [\beta_1, ...\beta_{n1}, \sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\zeta]^T$ assign nonzero probability to its true value, then the posterior distribution of ψ will converge to a point mass on its true value as $T_1 \to \infty$. Consequently, the posterior mean of $y_{n_1+1,t}(0)$ will converge to its true value, and so $\widehat{\tau}^{CIM}_{n_1+1,t}$ is an asymptotically unbiased estimate of $\tau_{n_1+1,t}$ (as $T_1 \to \infty$).

I then searched through the literature to find no formal proofs of Causal Impact being asymptotically unbiased. Proving a state space model is asymptotically unbaised formally would be a small but meaningful addition to the literature. I stop short of asymptotic normality because I don't need it for inference. My inference comes from the choice of priors.

Assumptions

I make the following assumptions:

Assumption 3.1. ϵ_t and $\eta_{j,t}$ are assumed independent of all other unknowns.

Assumption 3.2. $\pi(v)$ is unimodal.

Assumption 3.3. The likelihood function is bounded.

Assumption 3.4. Prior distribution includes the true value of the parameter.

Assumption 3.5. The model specification is the true DGP Write in math!!!

Assumption 3.6. Prior distribution includes the true value of the parameter.

What do I start with? The treatment effect is defined as $\alpha_{0,T_0+i} = y_{0,T_0+i}(1) - y_{0,T_0+i}(0)$ for $i = \{0, 1, ...T - T_0\}$. In words, I am interested in the treatment effect post-intervention. α_{0,T_0+i}, $y_{0,T_0+i}(1)$ and $y_{0,T_0+i}(0)$ are fixed values. I observe $y_{0,T_0+i}(1)$ and must estimate $y_{0,T_0+i}(0)$.

What am I proving: $\mathbb{E}[\hat{\alpha}_{0,T_0+i}] \to \alpha_{0,T_0+i}$ as $T_0 \to \infty$.

Where's the randomness? I get randomness from the choice of priors. Conditional on the data and the data generating process being correct, the choice of priors is the only source of variation.

How do I prove this? Show that $\mathbb{E}[\hat{y}_{0,T_0+i}(0) - y_{0,T_0+i}(0)] \to 0$.

I do this because:

$$\hat{\alpha}_{0,T_0+i} = y_{0,T_0+i}(1) - \hat{y}_{0,T_0+i}(0) = \alpha_{0,T_0+i} + y_{0,T_0+i}(0) - \hat{y}_{0,T_0+i}(0)$$

How do I prove that? Although $y_{0,T_0+i}(0)$ is thought as a fixed value, my modeling scheme sets $y_{0,T_0+i}(0)|Y_0, X \sim \mathcal{N}\left(\sum_{j=1}^{J+1} \beta_{j,T_0+i} x_{j,T_0+i}(0), \sigma^2\right)$.

I draw $\hat{y}_{0,T_0+i}(0) \sim \hat{f}(y_{0,T_0+i}(0)|Y_0, X)$. This distribution is called the posterior predictive distribution. The distribution can be rewritten as:

$$\hat{f}(y_{0,T_0+i}(0)|Y_0, X) = \int_{v' \in V} \hat{f}(y_{0,T_0+i}(0)|Y_0, X, v')\hat{Pr}(v'|Y_0, X)dv'$$

Notice: This is saying that the posterior predictive distribution is the average over all possible values of v'. This idea is known as Bayesian Model Averaging. For illustrative purposes, assume v is defined on a finite parameter space. If $\hat{Pr}(v' = v^0|Y_0, X) = 1$, then

$$\begin{aligned}\hat{f}(y_{0,T_0+i}(0)|Y_0, X) &= \sum_{v' \in V} \hat{f}(y_{0,T_0+i}(0)|Y_0, X, v')I(v' = v^0) \\ &= \hat{f}(y_{0,T_0+i}(0)|Y_0, X, v^0) \\ &\sim \mathcal{N}(\sum_{j=1}^{J+1} \beta_{j,T_0+i} x_{j,T_0+i}(0), \sigma^2)\end{aligned}$$

where I is an indicator function.

This directly leads to $\mathbb{E}[\hat{y}_{0,T_0+i}(0) - y_{0,T_0+i}(0)|Y_0, X] = 0$. Step 1 of the proof is to show that $\hat{Pr}(v' = v|Y_0, X) \to 1$ as $T_0 \to \infty$. If the parameter space for v is continuous, then step 1 is to show the mass of the posterior distribution of v collapses in the neighborhood of the true value of v.

In words, I am showing that the posterior distribution will collapse around the true values v^0 as more and more data is introduced. This is equivalent to saying the likelihood function dominates the choice of prior as $T_0 \to \infty$.

Step 2 is to show the posterior predictive distribution converges around $y_{0,T_0+i}(0)$ as $T_0 + i \to \infty$.

Step 3 is to show the average of the converged posterior perdictive distribution is $y_{0,T_0+i}(0)$.

4 Tools for Proof

From Gelman (2014)

Definition 4.1. The Kullback-Leibler Divergence of the proposed distribution $\hat{f}(y_{0,T_0+i}(0)|Y_0)$

with respect to the true distribution $f\left(y_{0,T_0+i}(0)|Y_0\right)$ is defined as:

$$KL_f(v) = \mathbb{E}_v\left(log\left(\frac{f\left(y_{0,T_0+i}(0)|Y_0\right)}{\hat{f}(y_{0,T_0+i}(0)|Y_0)}\right)\right)$$

Using the Kullback-Leibler Divergence, I can formally define the true parameter values, v^0, as:

$$v^0 = argmin_v KL_f(v)$$

Gelman (2014) also show $KL_f(v) \geq 0$ with equality if $f\left(y_{0,T_0+i}(0)|Y_0\right) = \hat{f}\left(y_{0,T_0+i}(0)|Y_0\right)$.

Proposition 4.0.1. Define $\mathcal{F}$ to be a finite family of likelihood models with $V = \{v : f(y_{0,T_0+i}(0)|Y_0, X) \in \mathcal{F}\}$ the parameter space. If V is a compact set, v^0 is defined above, and A is in the neighborhood of v^0, and $\hat{\pi}(v^0) > 0$, then $\hat{Pr}(v \in A|Y_0, X) \to 1$.

Remark. Gellman proves this statement when $\{Y_0, X\}$ are i.i.d. They also remark that this statement can still hold true if "there be 'replication' at some level, as, for example, if the data come in a time series whose correlations decay to zero". I tried proving this assuming ergodic stationarity but Dick brought up that assumption is invalid when the $\beta_{j,t}$ follow a random walk. I include the proposition to apply context for my thought process. I reprove the proposition without i.i.d. data in the proof.

5 The Actual Proof

Below, I type out the steps of the proof. I have my concerns highlighted in magenta:

Theorem: If all the above assumptions are met, then $\mathbb{E}[\hat{\alpha}_{0,t}] \to \mathbb{E}[\alpha_{0,t}]$ as $T_0 - 1 \to \infty$.

Proof. Step 1: Prove Proposition 4.0.1

Place a small neighborhood about each point in V with A being the only neighborhood to include v^0. Then cover V with a finite subset of these neighborhoods. Because V is assumed

compact, the finite subcovering can be obtained. I haven't worked with coverings before. Do I have to specify that for 2 neighborhoods A and B, $A \cap B = \{\}$?

Define B to be a neighborhood such that $v^0 \notin B$. Consider the following:

$$log\left(\frac{Pr(v \in B|Y_0, X)}{Pr(v \in A|Y_0, X)}\right) = log\left(\frac{\pi(v \in B|X)}{\pi(v \in A|X)}\right) + log\left(\frac{f(Y_0|v \in B, X)}{f(Y_0|v \in A, X)}\right) \tag{5}$$

$$= log\left(\frac{\pi(v \in B|X)}{\pi(v \in A|X)}\right) + log\left(\frac{\Pi_{t=1}^{T_0-1} f(y_{0,t}(0)|v \in B, X)}{\Pi_{t=1}^{T_0-1} f(y_{0,t}(0)|v \in A, X)}\right) \tag{6}$$

$$= log\left(\frac{\pi(v \in B|X)}{\pi(v \in A|X)}\right) + \sum_{t=1}^{T_0-1} log\left(\frac{f(y_{0,t}(0)|v \in B, X)}{f(y_{0,t}(0)|v \in A, X)}\right) \tag{7}$$

$$= log\left(\frac{\pi(v \in B|X)}{\pi(v \in A|X)}\right) + \frac{T_0 - 1}{T_0 - 1} \sum_{t=1}^{T_0-1} log\left(\frac{f(y_{0,t}(0)|v \in B, X)}{f(y_{0,t}(0)|v \in A, X)}\right) \tag{8}$$

Line (5) comes from Bayes Rule. Line (6) comes from the model specification in equations (1)-(4). Notice the model specification yields independent but non-identically distributed likelihoods. Line (7) is applying basic log rules and line (8) is multiplying the second term by 1.

Notice that as $T_0 - 1 \to \infty$:

$$\frac{1}{T_0-1}\sum_{t=1}^{T_0-1} log\left(\frac{f(y_{0,t}(0)|v\in B,X)}{f(y_{0,t}(0)|v\in A,X)}\right) \to \mathbb{E}\left(log\left(\frac{f(y_{0,t}(0)|v\in B,X)}{f(y_{0,t}(0)|v\in A,X)}\right)\right) \tag{9}$$

$$= \mathbb{E}\left(log\left(\frac{f(y_{0,t}(0)|v\in B,X)f(y_{0,t}(0)|v\in A,X)}{f(y_{0,t}(0)|v\in A,X)f(y_{0,t}(0)|v\in A,X)}\right)\right) \tag{10}$$

$$= \mathbb{E}\left(log(1) - log\left(\frac{f(y_{0,t}(0)|v\in A,X)}{f(y_{0,t}(0)|v\in B,X)}\right)\right) \tag{11}$$

$$= KL_f(v^0) - KL_f(v) \tag{12}$$

$$= -KL_f(v) \tag{13}$$

$$< 0 \tag{14}$$

Without ergodic stationarity, line 9 is INCORRECT. I am well aware this is a false statement. I cannot use the Law of Large Numbers because, conditional on my data and variables. I have independent non-identically distributed distributions. I am not sure what I have to assume to make this statement true. Line (10) is from multiplying by 1. Line (11) is from rearranging. Line (12) is the definition of the Kullback-Leibler Divergence. Line (13) shows that the first term in line (12) is 0 because log(1)=0. Line (14) is a property of the Kullback-Leibler Divergence.

Combining this result and line (8) yields:

$$log\left(\frac{Pr(v\in B|Y_0,X)}{Pr(v\in A|Y_0,X)}\right) \to -\infty$$

as $T_0 - 1 \to \infty$. Rearranging concludes $Pr(v \in B|Y_0, X) \to 0$ for all $v \neq v^0$. Therefore, $Pr(v \in A|Y_0, X) \to 1$. Equivalently, $v \to v^0$ as $T_0 - 1 \to \infty$.

Step 2:

$$lim_{T_0\to\infty}\hat{f}(y_{0,T_0+i}(0)|Y_0,X) = lim_{T_0\to\infty}\int_{v'\in V}\hat{f}(y_{0,T_0+i}(0)|Y_0,X,v')\hat{Pr}(v'|Y_0,X)dv' \quad (15)$$

$$= \int_{v'\in V} lim_{T_0\to\infty}\hat{f}(y_{0,T_0+i}(0)|Y_0,X,v')\hat{Pr}(v'|Y_0,X)dv' \quad (16)$$

$$= \int_{v'\in V}\hat{f}(y_{0,T_0+i}(0)|Y_0,X,v')lim_{T_0\to\infty}\hat{Pr}(v'|Y_0,X)dv' \quad (17)$$

$$= \hat{f}(y_{0,T_0+i}(0)|Y_0,X,v'\in A) \quad (18)$$

$$\sim \mathcal{N}(\sum_{j=1}^{J+1}\beta_{j,T_0+i}x_{j,T_0+i}(0),\sigma^2) \quad (19)$$

Line (15) is taking the limit of the posterior predictive distribution. Line (16) is bringing the limit into the integral. I am fairly certain this is illegal. Can we talk about situations where this is not illegal? Conditioned on Y_0, X, and v, $\hat{f}(y_{0,T_0+i}(0)|Y_0,X,v')$ does not depend on T_0; it can be treated as a constant with respect to the limit and pulled out. As of now, I am conditioning on the parameters only, NOT the states. This makes me a bit nervous my statement is incorrect. How do I think about states in an econometric setting? Line (18) comes from step 1. Line (19) is by definition.

step 3: $\mathbb{E}[\hat{\alpha}_{0,t}] \to \alpha_{0,t}$ as $T_0 - 1 \to \infty$

In order for $\mathbb{E}[\hat{\alpha}_{0,t}] \to \alpha_{0,t}$ as $T_0 - 1 \to \infty$, $\mathbb{E}\left[y_{0,T_0+i}(0) - \hat{y}_{0,T-0+i}(0)\right] \to 0$. By LIE,

$$\mathbb{E}\left[y_{0,T_0+i}(0) - \hat{y}_{0,T-0+i}(0)\right] = y_{0,T_0+i}(0) - \mathbb{E}\left[\hat{y}_{0,T-0+i}(0)\right]$$
$$= y_{0,T_0+i}(0) - \mathbb{E}\left[\mathbb{E}\left[\hat{y}_{0,T-0+i}(0)|Y_0,X\right]\right]$$

Take the limit on both sides leads to:

$$lim_{T_0+\to\infty}\mathbb{E}\left[y_{0,T_0+i}(0) - \hat{y}_{0,T-0+i}(0)\right] = lim_{T_0+i\to\infty}\left(y_{0,T_0+i}(0) - \mathbb{E}\left[\mathbb{E}\left[\hat{y}_{0,T-0+i}(0)|Y_0,X\right]\right]\right)$$
$$= y_{0,T_0+i}(0) - lim_{T_0+i\to\infty}\mathbb{E}\left[\mathbb{E}\left[\hat{y}_{0,T-0+i}(0)|Y_0,X\right]\right]$$

I stop here because of the limit and integral question. □

Things I need but am not sure how to say it - Moment conditions. I am unsure if I put my moment conditions on $\mathbb{E}(y_{0,T_0+i}(0)|Y_0, X)$ or $\mathbb{E}(y_{0,T_0+i}(0))$

- Stationarity or ergodicity? As mentioned before, I know that my LLN step is incorrect without further assumptions. I've reviewed the ergodic section of your notes from first year, but would like to discuss ways to actually implement this.

Citations

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics* 9 (1): 247–74. https://doi.org/10.1214/14-AOAS788.

Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods*. 2nd ed. Oxford Statistical Science Series 38. Oxford: Oxford University Press.

Gelman, Andrew. 2014. "Bayesian Data Analysis." In *Bayesian Data Analysis*, Third edition, 585–88. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press.

Samartsidis, Pantelis, Shaun R. Seaman, Anne M. Presanis, Matthew Hickman, and Daniela De Angelis. 2019. "Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units." *Statistical Science* 34 (3): 486–503. https://doi.org/10.1214/19-STS713.