(Empirical) Bayes Approaches to Parallel Trends

By Soonwoo Kwon and Jonathan Roth*

Researchers employing a difference-in-differences (DiD) design are often unsure about the validity of the parallel trends assumption. It is common to test for "pre-trends", yet such tests may be underpowered, and relying on them leads to statistical issues from pre-testing (Roth, 2022). Recent work by Manski and Pepper (2018) and Rambachan and Roth (2023, RR) has made progress on obtaining more credible inference when parallel trends may be violated by adopting a partial identification approach. In RR, for example, the researcher places bounds that restrict the possible values of the post-treatment violations of parallel trends δ_{post} given the identified pre-treatment violations δ_{pre}. The identified set for the treatment effect then corresponds to the worst-case bounds for δ_{post} given the observed δ_{pre}.

We instead consider a Bayesian approach where the researcher imposes a prior on the violations of parallel trends δ. The researcher then updates their posterior about δ_{post} given the observed estimate of δ_{pre}. This allows them to form posterior means and credible sets (CSs) for the treatment effect τ_{post}. The Bayesian approach allows the researcher to impose ex ante information about what violations of parallel trends may look like, and thus to potentially obtain more informative results than the partial identification approach using worst-case bounds. It also allows one to form point estimates in addition to confidence sets. For settings with many pre-treatment periods, we also consider empirical Bayes (EB) approaches, where the "prior" for the violations of parallel trends is calibrated using the pre-trends.

For more general work on Bayesian approaches in settings with partial identification, see Moon and Schorfheide (2012); Giacomini and Kitagawa (2021). For a related EB approach to parallel trends, see Leavitt (2020).

I. Set-up

Following RR, we consider a setting where the researcher observes a vector of event-study estimates $\hat{\beta} = (\hat{\beta}'_{pre}, \hat{\beta}'_{post})' \in \mathbb{R}^{\underline{T}+\bar{T}}$ corresponding to $\underline{T}$ pre-treatment and $\bar{T}$ post-treatment periods. Motivated by asymptotics based on the central limit theorem, we suppose that $\hat{\beta}$ is normally distributed with known variance, $\hat{\beta} \sim \mathcal{N}(\beta, \Sigma_{\hat{\beta}})$, where

$$\beta = \begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix} + \begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix}. \tag{1}$$

The vector τ corresponds to the treatment effect in each period (assumed to be zero prior to treatment, $\tau_{pre} = 0$), while δ corresponds to a vector of biases (e.g. violations of parallel trends). RR consider restrictions that impose that $\delta \in \Delta$. This enables partial identification of τ, with bounds corresponding to worst-case assumptions on the element $\delta \in \Delta$. In this paper, we alternatively consider Bayesian inference where the researcher places a prior on δ, as well as Empirical Bayes approaches where the prior on δ_{post} is calibrated using δ_{pre}.

II. Fully Bayesian Approach

We impose a prior $\pi_{\tau,\delta}(\cdot)$ over τ, δ.[1] From Bayes' rule, we have that

$$p(\tau, \delta \mid \hat{\beta}) \propto \ell(\hat{\beta} \mid \delta + \tau) \cdot \pi_{\tau,\delta}(\tau, \delta),$$

[*] Kwon: Brown University, soonwoo_kwon@brown.edu; Roth: Brown University, jonathanroth@brown.edu. We are grateful to Dmitry Arkhangelsky, Kirill Borusyak, Peter Hull, and Ashesh Rambachan for comments.

[1] For notational convenience, in this section we write τ for a vector of the form $(0, \tau'_{post})'$.

where $\ell(\hat{\beta} \mid \beta)$ denotes the normal likelihood of observing $\hat{\beta}$ given $\hat{\beta} \mid \beta \sim \mathcal{N}\left(\beta, \Sigma_{\hat{\beta}}\right)$. Consequently,

$$p(\tau \mid \hat{\beta}) = \int p(\tau, \delta \mid \hat{\beta})\, d\delta \\ \propto \int \ell(\hat{\beta} \mid \delta + \tau) \cdot \pi_{\tau,\delta}(\tau, \delta)\, d\delta$$

Thus, given a prior $\pi_{\tau,\delta}$ it is straightforward to compute the posterior $p(\tau \mid \hat{\beta})$.

In what follows, we will primarily consider the case where the researcher has an uninformative prior on $\tau \mid \delta$, so that $\pi_{\tau|\delta}(\tau \mid \delta) \propto 1$, in which case

$$p(\tau \mid \hat{\beta}) \propto \int \ell(\hat{\beta} \mid \delta + \tau) \cdot \pi_{\delta}(\delta)\, d\delta.$$

The following result characterizes the posterior mean for τ_{post} when the prior is uninformative.

PROPOSITION 1: *If the prior for τ is uninformative (i.e. $\pi_{\tau|\delta} \propto 1$), then*

$$E[\tau_{post} \mid \hat{\beta}] = \\ E[\beta_{post} \mid \hat{\beta}] - \underbrace{E[E[\delta_{post} \mid \delta_{pre} = \beta_{pre}] \mid \hat{\beta}]}_{=E[\delta_{post} \mid \hat{\beta}]}$$

Proposition 1 shows that the posterior mean for τ_{post} is simply the difference between the posterior for β_{post} and the posterior for δ_{post}. It shows further that the posterior for δ_{post} can be written as an iterated expectation, where the inner expectation is based on the conditional prior of δ_{post} given δ_{pre}, and the outer expectation is over the posterior for $\beta_{pre} \mid \hat{\beta}$.

It is worth noting that the expression for $E[\tau_{post} \mid \hat{\beta}]$ derived in Proposition 1 depends on the *conditional* prior on the post-treatment bias δ_{post} given the pre-trend δ_{pre}, regardless of the precision of the estimates $\hat{\beta}$. This reflects the well-known fact that in partially identified settings, the prior matters *even asymptotically.* Researchers adopting this approach must therefore be careful to choose a prior that reflects economic information about the possible violations of parallel trends.

Example: Gaussian Prior

Suppose we have a Gaussian prior for δ, $\delta \sim \mathcal{N}\left(\mu_{\delta}, V_{\delta}\right)$ and impose the uninformative prior for τ. A straightforward calculation using Bayes' rule shows that the posterior for τ_{post} is also Gaussian. To derive the posterior mean, note that the formula for the conditional mean of a Gaussian vector implies that

$$E[\delta_{post} \mid \delta_{pre}] = \mu_{\delta_{post}} + \Gamma_V'(\delta_{pre} - \mu_{\delta_{pre}}),$$

for $\Gamma_V = V_{\delta_{pre}}^{-1} V_{\delta_{pre},\delta_{post}}$. Applying Proposition 1,

$$E[\tau_{post} \mid \hat{\beta}] = \beta^*_{post} - \mu_{\delta_{post}} - \Gamma_V'(\beta^*_{pre} - \mu_{\delta_{pre}}),$$

where $\beta^* = E[\beta \mid \hat{\beta}]$ is the posterior mean for β.

One can further show that the posterior mean for β_{pre} is

$$\beta^*_{pre} = (\Sigma^{-1}_{\hat{\beta}_{pre}} + V^{-1}_{pre})^{-1}(\Sigma^{-1}_{\hat{\beta}_{pre}} \hat{\beta}_{pre} + V^{-1}_{pre} \mu_{\delta_{pre}}),$$

which "shrinks" the point-estimate $\hat{\beta}_{pre}$ towards the prior mean $\mu_{\delta_{pre}}$. Likewise, the posterior mean for β_{post} is

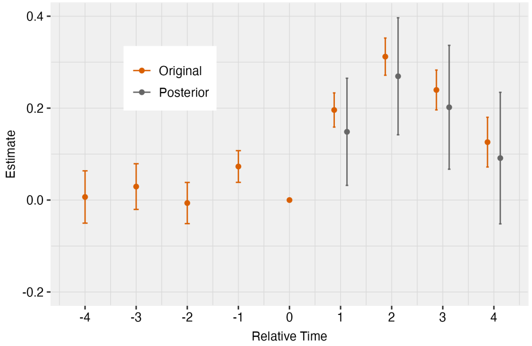
$$\beta^*_{post} = \hat{\beta}_{post} - \Gamma_\Sigma'(\hat{\beta}_{pre} - \beta^*_{pre})$$

for $\Gamma_\Sigma = \Sigma^{-1}_{\hat{\beta}_{pre}} \Sigma_{\hat{\beta}_{pre},\hat{\beta}_{post}}$. See the Online Appendix for detailed calculations, a formula for the posterior variance of τ_{post}, and an extension to the case with an uninformative Gaussian prior on τ_{post}.

Empirical Illustration

Benzarti and Carloni (2019, BZ) study the impacts of a reduction in the value-added tax on restaurants in France. They run a non-staggered DiD design comparing profits for restaurants to those of firms in other industries not affected by the tax change. The key concern with this approach is that there might be idiosyncratic economic factors affecting the profits of restaurants that do not affect other industries, which would lead to violations of parallel trends. We calibrate our prior on these violations using McGahan and Porter

(1999), who estimate an $AR(1)$ process for the industry-level component of firm profits (see the Online Appendix for detail). We take their estimates of the $AR(1)$ parameters and assume that the $AR(1)$ innovations are from a mean-zero Gaussian, which implies a Gaussian prior for the violations of parallel trends. The figure below shows the original OLS estimates and confidence intervals (CIs) from BZ, as well as posterior means and 95% CSs from our Bayesian approach.



The posterior CSs are wider than the OLS CIs, since the OLS CIs assume that parallel trends holds exactly, whereas our prior puts positive weight on violations of parallel trends. Nevertheless, the CSs are informative, excluding zero in 3 out of 4 post-treatment periods. The posterior means are also somewhat closer to zero than the OLS estimates. This is because δ_{pre} and δ_{post} are correlated under the imposed prior; thus, the primarily positive estimates for $\hat{\beta}_{pre}$ lead to a posterior that the post-treatment bias δ_{post} is positive.

III. Empirical Bayes Approaches

We saw in the previous section that the conditional prior $\delta_{post} \mid \delta_{pre}$ matters regardless of the precision of the event-study estimates $\hat{\beta}$. This conditional prior governs how violations of parallel trends evolve over time. In settings where we have many pre-treatment periods, and we think that the violations of parallel trends come from a *stationary* process, it might be attractive to learn the time-series dependence of violations of parallel trends from the pre-trends. This motivates an EB approach where the parameters of the time series process for violations of parallel trends are learned from the pre-trends, and posterior estimates are then calculated based on the prior implied by the estimated parameters.

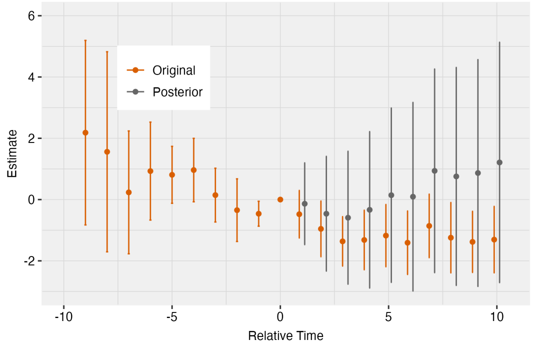
As a simple illustration, suppose that violations of parallel trends across consecutive periods are governed by $w_t := \delta_t - \delta_{t-1} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. In a simple non-staggered DiD, this corresponds to the case where the idiosyncratic factors differentially affecting the treated group follow a Gaussian random walk with drift. Let $w_{pre} = (w_{-\underline{T}+1}, ..., w_0)'$ collect the pre-treatment values of w_t. Analogously define the vector $\hat{w}_{pre}$ to collect the estimate of w_{pre} using $\hat{\beta}_{pre}$ instead of δ_{pre}. Then we have that $\hat{w}_{pre} \sim \mathcal{N}(\mu \cdot \mathbf{1}, \Sigma_w + \sigma^2 I)$, where $\mathbf{1}$ is the vector of ones, and $\Sigma_w = M\Sigma_{\hat{\beta}_{pre}} M'$ for M the matrix such that $\hat{w}_{pre} = M\hat{\beta}_{pre}$. The parameters μ and σ^2 can thus be estimated via maximum likelihood, which will be consistent (under mild regularity conditions on Σ_w) as the number of pre-treatment periods grows large, $\underline{T} \to \infty$. Since the assumption that $w_t \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ implies a normal prior for δ,[2] it is straightforward to calculate the posterior for τ_{post} using the prior implied by the estimates $\hat{\mu}, \hat{\sigma}$.

One caveat to this approach is that the consistency of the estimates for the prior depends on the number of pre-treatment periods $\underline{T}$ being large. In practice, the number of pre-treatment periods may be moderate—e.g., in our empirical application below, it is 9—in which case estimates based on this approach must be interpreted with some caution. We note that an alternative to the EB approach when the number of periods is moderate is to consider a hierarchical Bayes model, where one imposes a hyper-prior on the parameters μ, σ^2 and then updates their prior based on the observed estimate of $\hat{\beta}_{pre}$. This approach retains validity even when $\hat{\mu}, \hat{\sigma}$ are only imprecisely estimated, but of course requires the researcher to specify a prior on the hyper-parameters.

[2] We adopt the common normalization that $\delta_0 = 0$, which allows us to infer the distribution of δ from w.

Empirical Illustration

Lovenheim and Willén (2019) study how being exposed to laws that increase the power of teachers' unions as a student impacts earnings in adulthood. They use a two-way fixed effects event-study specification exploiting the differential timing of the passage of these laws.[3] The concern with the parallel trends assumption is that states passing these laws may have different secular trends in labor market outcomes. To address this, we suppose that $w_t \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, and estimate the parameters μ, σ using maximum likelihood based on the pre-trends.



Using female employment (in p.p.) as the outcome, we estimate $\hat{\mu} = -0.24, \hat{\sigma} = 0.61$, indicating a downward-sloping pre-trend and some variance around it. Because of the prior that the violation of parallel trends is downward sloping, the posterior means for the treatment effects are substantially closer to zero than the OLS estimates (see figure above). The CSs are also substantially wider than the OLS CIs, especially in later post-treatment periods, owing to uncertainty about the violations of parallel trends.

REFERENCES

Benzarti, Youssef, and Dorian Carloni. 2019. "Who Really Benefits from Consumption Tax Cuts? Evidence from a Large VAT Reform in France." *American Economic Journal: Economic Policy*, 11(1): 38–63.

Giacomini, Raffaella, and Toru Kitagawa. 2021. "Robust Bayesian Inference for Set-Identified Models." *Econometrica*, 89(4): 1519–1556.

Leavitt, Thomas. 2020. "Beyond Parallel Trends: Improvements on Estimation and Inference in the Difference-in-Differences Design." *Working paper.*

Lovenheim, Michael F., and Alexander Willén. 2019. "The Long-Run Effects of Teacher Collective Bargaining." *American Economic Journal: Economic Policy*, 11(3): 292–324.

Manski, Charles F., and John V. Pepper. 2018. "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions." *The Review of Economics and Statistics*, 100(2): 232–244.

McGahan, Anita M., and Michael E. Porter. 1999. "The Persistence of Shocks to Profitability." *The Review of Economics and Statistics*, 81(1): 143–153.

Moon, Hyungsik Roger, and Frank Schorfheide. 2012. "Bayesian and Frequentist Inference in Partially Identified Models." *Econometrica*, 80(2): 755–782.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *The Review of Economic Studies*, 90(5): 2555–2591.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics*, 235(2): 2218–2244.

[3] A recent literature surveyed in Roth et al. (2023) has shown that such specifications may be difficult to interpret under treatment effect heterogeneity; one could re-do the analysis here with the event-study from one of the estimators developed for this case.