Time-Varying Heterogeneous Treatment Effects in Event Studies*

Irene Botosaru
McMaster University

Laura Liu
University of Pittsburgh

First version: December 11, 2024
This version: September 17, 2025

Abstract

This paper examines the identification and estimation of heterogeneous treatment effects in event studies, emphasizing the importance of both lagged dependent variables and treatment effect heterogeneity. We show that omitting lagged dependent variables can induce omitted variable bias in the estimated time-varying treatment effects. We develop a novel semiparametric approach based on a short-T dynamic linear panel model with correlated random coefficients, where the time-varying heterogeneous treatment effects can be modeled by a time-series process to reduce dimensionality. We construct a two-step estimator employing quasi-maximum likelihood for common parameters and empirical Bayes for the heterogeneous treatment effects. The procedure is flexible, easy to implement, and achieves ratio optimality asymptotically. Our results also provide insights into common assumptions in the event study literature, such as no anticipation, homogeneous treatment effects across treatment timing cohorts, and state dependence structure.

Keywords: Event study, heterogeneous treatment effects, dynamic panel data, correlated random coefficients, empirical Bayes

JEL classification: C11, C14, C21, C23

*botosari@mcmaster.ca (Botosaru) and laura.liu@pitt.edu (Liu). We thank Stéphane Bonhomme, Simon Freyaldenhoven, Chris Muris, Jon Roth, and conference participants at CFE-CMStatistics for helpful comments and discussions. The authors are solely responsible for any remaining errors.

1 Introduction

Event study methods have been a cornerstone for tracing dynamic treatment effects in empirical research across economics, finance, public policy, and related fields. Indeed, between 2020 and 2024, over thirty papers employing event study or dynamic difference-in-differences were published in the *American Economic Review*. The most common implementation is via the two-way fixed-effects (TWFE) regression, which aligns units by event time rather than calendar time, allowing researchers to estimate dynamic responses to treatments and interventions, while controlling for unobserved heterogeneity that is constant over time within units (i.e., unit effects) and/or common across units within time (i.e., time effects). In practice, researchers often estimate

$$Y_{it} = \alpha_i + \gamma_t + \sum_{j=-L}^{J} D_{it}^{j} \delta_j + X_{it}'\beta + U_{it},$$

where D_{it}^{j} indicates that unit i is j periods from its event date, X_{it} are observed covariates, α_i and γ_t are unit and time fixed effects, and $\{\delta_j\}$ represent average treatment effects at different leads and lags. Typically, the covariates are assumed to be strictly exogenous, i.e., they are uncorrelated with the error term across all time periods, so that current, past, and future values of the covariates do not respond to shocks in the outcome equation. This framework is attractive for its intuitive interpretation and straightforward implementation. See also recent reviews by Freyaldenhoven, Hansen, Pérez, and Shapiro (2021) and Miller (2023).

Despite its widespread use, the standard two-way fixed effects (TWFE) estimator relies on strong assumptions that may not hold in empirical applications. In particular, by omitting lagged outcomes, it implicitly assumes that unit and time fixed effects are sufficient to eliminate all serial dependence in the residual. This assumption is often violated in settings where economic outcomes — such as consumption, employment, earnings, and investment — exhibit persistence due to habit formation, adjustment costs, or other dynamic mechanisms. When lagged outcomes are correlated with treatment timing, TWFE estimators conflate causal effects with residual dynamics. This can induce spurious pre-trends, bias post-treatment estimates, and lead to invalid inference, including misleading placebo tests and confidence intervals. Although dynamic panel methods are well developed, they remain underutilized in applied event study analyses.

Second, and of equal importance, is the potential heterogeneity in treatment effects. While the average treatment effect summarizes the mean response, distributional and welfare analyses often depend on the full distribution of treatment effects across units. For example, targeted subsidies may yield disproportionate benefits for specific demographic groups. Assuming homogeneous effects can mask such variation and lead to suboptimal or inequitable policy recommendations. Furthermore, treatment effects may vary systematically with observed covariates — such as pre-treatment outcomes or demographic characteristics — as well as unobserved unit-level attributes, including preferences or ability. Recognizing and modeling such heterogeneity is therefore essential for designing targeted interventions and for evaluating their distributional consequences.

In this paper, we introduce a semiparametric model for *time-varying heterogeneous treatment effects* (TV-HTE) that simultaneously tackles outcome dynamics and cross-unit heterogeneity. For example, we can model

$$Y_{it} = \rho_Y Y_{i,t-1} + \alpha_i + \gamma_t + \sum_{j=0}^{J} D_{it}^j \delta_{ij} + X_{it}'\beta + U_{it}, \quad U_{it} \overset{\text{iid}}{\sim} (0, \sigma_U^2),$$

where ρ_Y captures outcome persistence, and δ_{ij} is the *unit- and event-time-specific* treatment effect. To reduce dimensionality, we can impose an AR(p) process on the treatment effects. For $p = 1$, we can write

$$\delta_{ij} = \rho_\delta \delta_{i,j-1} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} (0, \sigma_\varepsilon^2), \quad j \geq 1,$$

with δ_{i0} unrestricted. This AR(1) specification parsimoniously captures persistence or decay in heterogeneous responses while allowing each unit to have a distinct initial effect δ_{i0}.

Interpreting $\lambda_i = (\alpha_i, \delta_{i0})'$ as *correlated random coefficients*, we permit their joint distribution to depend flexibly on the initial outcomes Y_{i0}, exogenous covariates X_i, and the treatment timing. Under the assumption of conditional strict exogeneity of treatment—that U_{it} is independent of treatment conditional on these covariates—and a mild non-vanishing characteristic function condition, we achieve nonparametric identification of both the common parameters $\theta = (\rho_Y, \rho_\delta, \beta, \sigma_U^2, \sigma_\varepsilon^2)'$ and the conditional distribution of the random coefficients λ_i.

Building on the identification result and further assuming Gaussianity on U_{it} and ε_{ij}, we develop a two-step estimation procedure that is straightforward to implement. In the first

step, we estimate the common parameters θ by quasi-maximum likelihood (QMLE). To do so, we assume a Gaussian form for the conditional distribution of the random coefficients λ_i, integrate them out of the joint likelihood, and obtain $\widehat{\theta}$ by maximizing the resulting marginal likelihood. We show that even when this Gaussian assumption is misspecified, the QMLE remains consistent and asymptotically normal.

In the second step, we recover unit-specific estimates of λ_i via *empirical Bayes*. Let $\widehat{\lambda}_i$ denote the MLE estimate of λ_i. One can show that $\widehat{\lambda}_i = \lambda_i + V_i$, where V_i has mean zero and a variance matrix estimated from the first-step output. Tweedie's formula then yields the posterior mean that combines this noisy MLE estimate with a correction term that depends on the derivative of the marginal density of the sufficient statistics. Intuitively, this correction shrinks the MLE estimate toward regions of higher density in the data, effectively combining information across units to improve the estimation accuracy.

By focusing on the derivative of the observed marginal density of the sufficient statistics $p(\widehat{\lambda}_i \mid Y_{i0}, X_i)$, we sidestep the challenging deconvolution problem to recover the underlying distribution of $\pi(\lambda \mid Y_0, X)$. The marginal density of the sufficient statistics can be estimated either parametrically or nonparametrically, and the resulting empirical Bayes estimator shrinks noisy unit-level estimates toward a data-driven prior and achieves ratio optimality, that is, its compound risk converges to the oracle risk that would be attained by an infeasible estimator with perfect knowledge of the true conditional random coefficient distribution.

This TV-HTE framework provides several advantages compared to the standard event study methods. Incorporating the lagged dependent variable eliminates omitted-variable bias due to persistence. Modeling heterogeneity through a time-series process captures the dynamics in treatment effects without high-dimensional estimation. The empirical Bayes step sharpens unit-level estimates in short panels, overcoming the many-means problem.

In addition to the above setup, our framework extends naturally to discrete or continuous treatments and to staggered adoption designs. We also allow for both strictly exogenous covariates, whose coefficients may be unit-specific or common, and predetermined covariates with common effects. The dynamics for Y_{it} and δ_{ij} can be generalized to AR(p) processes, e.g., AR(2) to capture oscillatory patterns, and the error structure can be generalized to allow for cross-sectional heteroskedasticity $\sigma^2_{U,i}$ or MA(q) process.

Moreover, our framework also sheds light on common assumptions in event study. For example, by examining the estimated means of the event-time coefficients in pre-treatment

periods ($j < 0$), we can formally test the no anticipation assumption. Also, by comparing these means across cohorts defined by treatment timing, we can assess the homogeneity of treatment effects. In addition, our dynamic panel structure with separate persistence parameters for the outcome ρ_Y and the treatment effects ρ_δ allows us to evaluate state dependence in both the underlying process and the policy response.

We assess the performance of our TV-HTE estimator through extensive Monte Carlo experiments and an empirical example on county-level unemployment during the 2008 Great Recession. In the Monte Carlo, our method nearly replicates the infeasible oracle in recovering the distribution of unit-specific effects under Gaussian, bimodal, and heavy-tailed distributions, and across dynamic response profiles ranging from monotonic decay to oscillatory paths. Our tests maintain correct size under the null and exhibit high power. In the empirical example, we find that the heterogeneous treatment effects are markedly non-Gaussian and irregularly distributed: county-level unemployment spikes range from roughly 0.5 to over 7 percentage points, far surpassing the average TWFE estimate, and dynamic trajectories differ across counties as well. Formal tests reject the random effects specification, the null of no correlation between heterogeneous effects and baseline heterogeneity, and the null of no state dependence, instead supporting our correlated random coefficients, time-varying analysis.

Related literature. Since the pioneering work by Ashenfelter (1978) on estimating the effects of training programs on earnings using a two-way fixed-effects model, empirical researchers have widely adopted panel data event study designs to quantify causal effects in economics. However, a growing literature recognizes that homogeneous effect assumptions can yield misleading estimates in staggered adoption settings, and recent work has fallen into three methodological strands. First, robust estimators for the mean treatment effect, such as de Chaisemartin and D'Haultfœuille (2023) and Borusyak, Jaravel, and Spiess (2024), rely on carefully constructed two-by-two comparisons or imputation-based counterfactuals to eliminate bias. Second, group-level approaches, such as Callaway and Sant'Anna (2021), Goodman-Bacon (2021), and de Chaisemartin and D'Haultfœuille (2023), estimate cohort- and period-specific treatment effects and aggregate them with convex weights or interaction weighted regressions to ensure no negative contributions. Finally, Arkhangelsky, Imbens, Lei, and Luo (2024) consider individual-level treatment effects via finite-mixture and latent-type models. In this paper, we also examine individual-level treatment effects and incorporate

an empirical Bayes approach to refine these estimates, thereby improving precision while flexibly accommodating time-varying heterogeneity. Our analysis also helps assess common assumptions underlying event study designs, such as those in Sun and Abraham (2021).

To accommodate outcome persistence and mitigate the Nickell bias in short panels, we draw on dynamic panel methods. Anderson and Hsiao (1982) propose first-differencing and using deeper lags as instruments to eliminate fixed effects. Arellano and Bond (1991) generalize this with a GMM estimator that exploits all available lagged levels, substantially improving efficiency in panels with small T. Blundell and Bond (1998)'s system GMM further addresses weak-instrument concerns when the autoregressive coefficient is high. Arellano and Bonhomme (2012) show that, under mild serial-correlation restrictions, one can identify moments—and even the full distribution—of random coefficients in a short panel. Alvarez and Arellano (2022) develop robust QMLE for dynamic panels that remain valid under heteroskedasticity and arbitrary serial correlation, demonstrating that random-effects likelihood methods can outperform GMM when distributional assumptions approximately hold. In this paper, we similarly estimate the common autoregressive parameters via QMLE in the first step, and the time dynamics of the heterogeneous treatment effects are further modeled by time-series processes to reduce dimensionality.

Our second step employs an empirical Bayes estimator to recover unit-specific treatment trajectories. Robbins (1951) introduces empirical Bayes as a compound decision problem, yielding shrinkage rules that minimize average risk without knowing the prior distribution. With exponential family likelihood, Tweedie's formula links posterior means to the derivatives of the marginal density of sufficient statistics, enabling nonparametric π-modeling empirical Bayes (Efron, 2011). Brown and Greenshtein (2009) and Jiang and Zhang (2009) establish that maximum-likelihood empirical Bayes estimators for normal-means problems achieve asymptotic minimaxity or ratio optimality. Gu and Koenker (2017) and Liu, Moon, and Schorfheide (2020) show substantial gains in estimation and forecasting accuracy by efficiently combining information across cross-sectional units. In this paper, we employ both parametric and nonparametric empirical Bayes to obtain posterior mean estimates of unit-specific treatment trajectories, optimally balancing individual signal and noise, and establish their ratio optimality.

The remainder of this paper is organized as follows. Section 2 introduces the model and discusses the identification of time-varying heterogeneous treatment effects. Section 3 presents our two-step estimation method and establishes its asymptotic properties, including

ratio optimality. Section 4 extends our estimator to various contexts and discusses tests for common event study assumptions. Section 5 conducts Monte Carlo experiments to examine the finite-sample properties of our estimators. Section 6 employs our panel data estimator to analyze how the Great Recession in 2008 affected local labor markets. Finally, Section 7 concludes. Appendix A provides the proofs for all propositions and theorems, and the online appendix contains additional tables and figures.

2 Simple model and identification

2.1 Importance of lagged dependent variables

Economic series tend to be persistent over time. For example, consumption adjusts gradually as habits evolve, and wages move slowly amid contract and adjustment frictions. When such built-in persistence coincides with event timing, the dummy variables in a TWFE regression absorb not only the true effect of the intervention but also the persistence present in the data. As a result, what appear as treatment effects may also reflect the persistence of past outcomes, giving rise to spurious pre-trends, distorted post-treatment estimates, and misleading inference in placebo tests and confidence intervals.

A simple, yet revealing, illustration shows why excluding lagged dependent variables from an event study regression generates omitted variable bias in the estimated treatment effect path. Consider a panel with five periods ($t = 0, 1, 2, 3, 4$) and a common treatment occurring at $t = 2$, so that $D_{it}^j = \mathbf{1}\{t - j = 2\}$. Suppose the true DGP is an AR(1) model with persistence ρ_Y and a treatment effect path $(\delta_0, \delta_1, \delta_2)$,

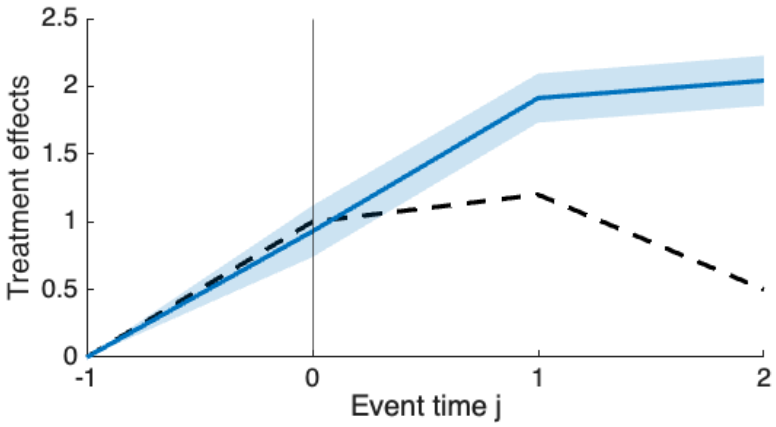
$$Y_{it} = \rho_Y Y_{i,t-1} + \sum_{j=0}^{2} D_{it}^j \delta_j + U_{it},$$

and let $\mathbb{E}[Y_{i0}] = 0$ for simplicity. In contrast, the naive event study regression omits dynamics and simply fits

$$Y_{it} = \sum_{j=0}^{2} D_{it}^j \widetilde{\delta}_j + \widetilde{U}_{it}.$$

Because the true outcomes are serially correlated, each indicator D_{it}^j is correlated with the

Figure 1: Omitted variable bias - toy example



Notes: The black dashed line shows the true treatment effect path $(\delta_0, \delta_1, \delta_2) = (1, 1.2, 0.5)$, while the blue solid line shows the estimated treatment effect path of $\{\widetilde{\delta}_j\}$ from a naive event study regression without lagged dependent variables. The blue band shows the 95% confidence interval.

omitted lag $Y_{i,t-1}$, producing bias in $\widetilde{\delta}_j$. One can show analytically that

$$\text{Bias}(\widetilde{\delta}_j) = \rho_Y \mathbb{E}\left[D_{it}^j Y_{i,t-1}\right] = \rho_Y \mathbb{E}[Y_{i,j+1}] = \begin{cases} 0, & j = 0, \\ \rho_Y \delta_0, & j = 1, \\ \rho_Y \delta_1 + \rho_Y^2 \delta_0, & j = 2. \end{cases}$$

Thus, even if the true effect at $j = 0$ is identified without bias, biases accumulate at longer horizons, distorting the entire treatment path.

Figure 1 contrasts the true effects (black dashed) with the biased estimates (blue solid) for $\rho_Y = 0.8$ and $(\delta_0, \delta_1, \delta_2) = (1, 1.2, 0.5)$ in a simulated sample of $N = 100$, and their differences are statistically significant. This toy example highlights the necessity of explicitly modeling lagged dynamics in event study designs. By incorporating $Y_{i,t-1}$, researchers can control for outcome persistence and recover unbiased estimates of the time-varying treatment effects.

2.2 Dynamic panel with time-varying het. treatment effects

We now introduce a simple dynamic panel framework that accommodates both persistence in the outcome and heterogeneous treatment effects across units and event time horizons. To highlight the main intuition, we focus on a simple model that drops time fixed effects and other covariates, and adopts a common treatment timing in this section. More general cases are discussed in subsequent sections.

Let $i = 1, \ldots, N$ index cross-sectional units and $t = 0, \ldots, T$ denote time periods. We consider a large N, fixed T setup, which is natural for many event study applications where the number of treated and control units is large but the available pre- and post-treatment windows are of limited length. For simplicity, each unit undergoes a single treatment at a common period t_0. We define the event time indicator $D_{it}^j = \mathbf{1}\{t - j = t_0\}, \quad j = 0, 1, \ldots, J$, so that $D_{it}^j = 1$ when unit i is in the jth period after treatment. Our baseline outcome equation augments a standard dynamic panel with these event time dummies

$$Y_{it} = \rho_Y Y_{i,t-1} + \alpha_i + \sum_{j=0}^{J} D_{it}^j \delta_{ij} + U_{it}, \quad U_{it} \overset{\text{iid}}{\sim} (0, \sigma_U^2). \tag{1}$$

Here, ρ_Y captures first-order persistence in the outcome, while the unit-specific intercept α_i controls for time-invariant heterogeneity. The term δ_{ij} is the treatment effect for unit i at event time j, allowing each unit to respond differently and dynamically to the intervention.

Because freely estimating the full matrix $\{\delta_{ij}\}$ would involve $(J + 1) \times N$ parameters, we can incorporate a simple time series structure on the heterogeneous effects to reduce the dimensionality.[1] For example, for $j \geq 1$ we assume an AR(1) process

$$\delta_{ij} = \rho_\delta \delta_{i,j-1} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} (0, \sigma_\varepsilon^2). \tag{2}$$

The persistence parameter ρ_δ governs the decay or oscillation of treatment effects over successive periods, while the variance σ_ε^2 captures unit-specific shocks to the response path. Only the initial effect δ_{i0} remains freely heterogeneous, enabling each unit to have its own starting point for the dynamic treatment response.

To capture potential correlations between initial outcomes, individual heterogeneity, and initial treatment effects, we let

$$\lambda_i = (\alpha_i, \delta_{i0})', \quad \lambda_i \mid Y_{i0} \sim \pi(\lambda_i \mid Y_{i0}),$$

where $\pi(\lambda \mid Y_0)$ is an unrestricted conditional density. This correlated random coefficients specification allows α_i and δ_{i0} to depend flexibly on the initial outcome Y_{i0} (and, in extensions,

[1] The assumed time series structure for δ_{ij} is testable in the data. For example, one can obtain preliminary estimates of the individual effect trajectories by orthogonal forward differencing of Arellano and Bover (1995), and then subject these series to standard time-series diagnostics to assess whether an AR(p) process provides an adequate fit.

on additional exogenous covariates). Moreover, by allowing for correlation between the baseline heterogeneity α_i and the initial treatment effects δ_{i0}, the framework can capture meaningful heterogeneity in treatment effects that standard event study methods might overlook.

Collecting the parameters into the vector $\theta = (\rho_Y, \rho_\delta, \sigma_U^2, \sigma_\varepsilon^2)'$ with true value θ_0, we aim to recover θ, the conditional distribution of λ_i, and posterior mean estimates of λ_i.

2.3 Identification

We now formalize the conditions under which both the common parameters θ and the conditional distribution of the unit-specific coefficients λ_i are nonparametrically identified.

Assumption 2.1 (Model) *Consider the simple model given by* (1) *and* (2) *with common treatment period* t_0.

(a) (Y_{i0}, λ_i) *are i.i.d. across* i.

(b) $U_{it} \perp (Y_{i,0:t-1}, \lambda_i)$, $\varepsilon_{ij} \perp (\delta_{i,1:j-1}, Y_{i0}, \lambda_i)$, *and* $U_{it} \perp \varepsilon_{ij}$, *for all* i, t, *and* j.

Condition (b) implies that the combined error terms $\check{U}_{i,1:T}(\rho_\delta)$ in (3) and hence the noise $V_i(\rho_\delta)$ in (5) below are independent of λ_i conditional on Y_{i0}, a key requirement for the deconvolution exercise.

Remark 2.1 (Conditional exogeneity in treatment) Under a common treatment timing, our baseline specification implicitly imposes conditional exogeneity of treatment: the innovation U_{it} is assumed independent of the event time indicators D_{it}^j (or, in a more general case with different treatment timings, independent once we condition on observed covariates). This condition ensures that the design matrix $W_i(\rho_\delta)$ for heterogeneous coefficients in (4) below is exogenous, so that the deconvolution step yields valid identification results.

It is useful to contrast this with the classic parallel trends assumption, which typically requires no outcome persistence ($\rho_Y = 0$) and $\mathbb{E}[U_{it}^{(0)} \mid \{D_{ij}^j\}] = 0$, where $U_{it}^{(0)}$ denotes the potential error under no treatment. Here we relax the parallel trends assumption by allowing $\rho_Y \neq 0$.[2] Although our conditional exogeneity assumption is stronger than standard parallel

[2]Under our model, the transformed outcome $Y_{it} - \rho_Y Y_{i,t-1}$ satisfies a conditional parallel trend assumption once we control for exogenous covariates, as discussed in Wooldridge (2021).

trends in terms of its assumption on the error terms, it affords us the flexibility to estimate richer heterogeneous treatment effect trajectories.

Moreover, by framing (α_i, δ_{i0}) as correlated random coefficients, we naturally accommodate *selection on unobservables*, where treatment timing can correlate with observed covariates, latent heterogeneity including heterogeneous treatment effects, as well as time fixed effects in the general model.

Combining the simple dynamic panel data model (1) and the AR(1) process (2), we obtain

$$Y_{it} - \rho_Y Y_{i,t-1} = \alpha_i + \left(\sum_{j=0}^{J} \rho_\delta^j D_{it}^j \right) \delta_{i0} + \underbrace{\left(U_{it} + \sum_{j=0}^{J} \sum_{k=1}^{j} \rho_\delta^{j-k} D_{it}^j \varepsilon_{ik} \right)}_{\equiv \check{U}_{it}(\rho_\delta)}. \tag{3}$$

where $\check{U}_{i,1:T}(\rho_\delta)$ is a mean-zero vector with covariance matrix $\Sigma_{\check{U}}(\theta)$. Next, define the $T \times 2$ design matrix $W_i(\rho_\delta)$ by

$$W_i(\rho_\delta) = \begin{pmatrix} 1 & \sum_{j=0}^{J} \rho_\delta^j D_{i1}^j \\ 1 & \sum_{j=0}^{J} \rho_\delta^j D_{i2}^j \\ \vdots & \vdots \\ 1 & \sum_{j=0}^{J} \rho_\delta^j D_{iT}^j \end{pmatrix}, \tag{4}$$

and let $W_{it}(\rho_\delta)$ be its t-th row.[3] The model can then be written compactly as

$$Y_{it} - \rho_Y Y_{i,t-1} = W_{it}(\rho_\delta)' \lambda_i + \check{U}_{it}(\rho_\delta).$$

Given $\rho = (\rho_\delta, \rho_Y)'$, the OLS/MLE estimator of the latent coefficient vector λ_i is

$$\widehat{\lambda}_i(\rho) = W_i(\rho_\delta)^+ \left(Y_{i,1:T} - \rho_Y Y_{i,0:T-1} \right) = \lambda_i + V_i(\rho_\delta), \tag{5}$$

where $W_i(\rho_\delta)^+ = \left(W_i(\rho_\delta)' W_i(\rho_\delta) \right)^{-1} W_i(\rho_\delta)'$ and $V_i(\rho_\delta) = W_i(\rho_\delta)^+ \check{U}_{i,1:T}(\rho_\delta)$, which is mean-zero and has covariance matrix $\Sigma_{V,i}(\theta) = W_i(\rho_\delta)^+ \Sigma_{\check{U}}(\theta) [W_i(\rho_\delta)^+]'$. Thus, $\widehat{\lambda}_i(\rho)$ is a sufficient statistic for λ_i with noise $V_i(\rho_\delta)$.

[3] In our simple setup with common treatment timing t_0, the design matrix $W_i(\rho_\delta)$ is deterministic and homogeneous across all units, so there is no need to condition on it in the assumptions and derivations, thereby simplifying the exposition.

Assumption 2.2 (Distributions)

(a) The characteristic functions of $\lambda_i \mid Y_{i0}$, U_{it}, and ε_{ij} are non-vanishing almost everywhere.

(b) The characteristic functions of U_{it} and ε_{ij} are twice differentiable.

(c) $Var(\delta_{i0}) > 0$ and $Var(Y_{i0}) > 0$.

Conditions (a) and (b) guarantee that the convolution in (5) can be inverted via characteristic function methods, thereby recovering the conditional distribution of $\lambda_i \mid Y_{i0}$. Condition (c) ensures cross-sectional variation in both the initial treatment effects and initial outcomes, guaranteeing that the moment conditions for identifying ρ_δ and ρ_Y are non-degenerate.

Assumption 2.3 (Rank condition) *$t_0 \geq 3$, and $T - t_0 \geq J \geq 1$.*

Since $\check{U}_{i,1:T}(\rho_\delta)$ is an MA(J) process in the error terms $\{U_{it}, \varepsilon_{ij}\}$, we require sufficient pre-treatment variation to disentangle these shocks from the treatment effect dynamics. In the simple common timing design, this amounts to imposing $t_0 \geq 3$, which helps satisfy the rank conditions in Arellano and Bonhomme (2012). For general cases with different treatment timings and additional covariates, we can extend to a more general rank condition on the expanded design matrix. $T - t_0 \geq J \geq 1$ ensure that there are enough post-treatment observations to identify the full sequence of dynamic treatment effects.

Theorem 2.1 (Nonparametric identification) *Under Assumptions 2.1–2.3, the common parameters θ and the conditional density $\pi(\lambda_i \mid Y_{i0})$ are identified.*

First, we can identify the autoregressive parameters ρ from moment conditions. Second, the identification of the conditional density $\pi(\lambda_i \mid Y_{i0})$ relies on the sufficient statistics representation (5). Taking characteristic functions on both sides transforms the convolution in the time domain into a product in the frequency domain, so one obtains on the right hand side a product of the characteristic functions of the latent coefficients $\lambda_i \mid Y_{i0}$ and the noise term $V_i(\rho)$. Under the non-vanishing characteristic functions, this product can be deconvolved to recover both distributions. Our proof builds on the deconvolution argument of Arellano and Bonhomme (2012) and Liu (2023) for correlated random coefficients panels and extends it to the dynamic event study framework.

Algorithm 1 Semiparametric TV-HTE estimator

Input: Panel data $\{Y_{it}\}_{i=1,\dots,N}^{t=0,\dots,T}$, treatment timing t_0, horizon J.

Output: Estimates of common parameters $\widehat{\theta}$ and unit-level parameters $\{\widetilde{\lambda}_i\}$.

Step 1: QMLE for common parameters. Maximize the marginal quasi-log-likelihood

$$\ell_N(\theta, b_0, b_1, \Sigma_\lambda) = \sum_{i=1}^{N} \log \phi \left(Y_{i,1:T};\, \mu_i(\theta, b_0, b_1),\, \Omega_i(\theta, \Sigma_\lambda) \right),$$

where $\mu_i(\cdot)$ and $\Omega_i(\cdot)$ are given in (6), to obtain $\left(\widehat{\theta}, \widehat{b}_0, \widehat{b}_1, \widehat{\Sigma}_\lambda\right)$.

Step 2: Empirical Bayes for unit-specific parameters

1. Build $T \times 2$ matrix $\widehat{W}_i = W_i(\widehat{\rho}_\delta)$ with rows $\left[1, \sum_{j=0}^{J} \widehat{\rho}_\delta^j D_{it}^j\right]$, and $\widehat{W}_i^+ = \left(\widehat{W}_i' \widehat{W}_i\right)^{-1} \widehat{W}_i'$.
2. Compute OLS/MLE estimate and noise covariance

$$\widehat{\lambda}_i = \widehat{W}_i^+ \left(y_{i,1:T} - \widehat{\rho}_Y y_{i,0:T-1} \right), \quad \widehat{\Sigma}_{V,i} = \widehat{W}_i^+ \Sigma_{\check{U}}(\widehat{\theta}) \widehat{W}_i^{+\prime}.$$

3. Estimate marginal density of the sufficient statistics $p(\widehat{\lambda}_i \mid Y_{i0})$ either parametrically or nonparametrically.
4. Apply Tweedie's formula:

$$\widetilde{\lambda}_i = \widehat{\lambda}_i + \widehat{\Sigma}_{V,i} \nabla_{\widehat{\lambda}_i} \log \widehat{p}\left(\widehat{\lambda}_i \mid y_{i0}\right).$$

3 Estimation and asymptotics

3.1 Two-step estimation

Building on the identification results and further assuming Gaussianity on U_{it} and ε_{ij}, we implement a simple two-step estimator that first estimates the common parameter and then recovers the unit-specific parameters, as summarized in Algorithm 1.

In the first step, we estimate the common parameters θ by QMLE, treating the latent coefficients $\lambda_i \mid Y_{i0}$ as if they followed a Gaussian regression model

$$\lambda_i \mid Y_{i0} \sim N\left(b_0 + b_1 Y_{i0},\, \Sigma_\lambda\right).$$

Even though this correlated random coefficients distribution may be misspecified, maximizing the resulting marginal likelihood over θ and the nuisance parameters $(b_0, b_1, \Sigma_\lambda)$ yields consistent and asymptotically normal estimates for θ. In practice, the Gaussian prior and likelihood imply conjugacy, yielding a closed-form marginal likelihood

$$Y_{i,1:T} \sim \mathcal{N}\left(\mu_i(\theta, b_0, b_1), \Omega_i(\theta, \Sigma_\lambda)\right),$$

where

$$\begin{aligned} \mu_i(\theta, b_0, b_1) &= A(\rho_Y) Y_{i0} + \widetilde{W}(\rho_Y, \rho_\delta)\left(b_0 + b_1 Y_{i0}\right), \\ \Omega_i(\theta, \Sigma_\lambda) &= B(\rho_Y)\Sigma_{\check{U}}(\theta)B(\rho_Y)' + \widetilde{W}(\rho_Y, \rho_\delta)\Sigma_\lambda \widetilde{W}(\rho_Y, \rho_\delta)', \end{aligned} \tag{6}$$

where $A(\rho_Y) = (\rho_Y, \rho_Y^2, \ldots, \rho_Y^T)'$ captures initial condition propagation, $B(\rho_Y)$ is the $T \times T$ lower triangular matrix with (s,t)-th element ρ_Y^{s-t} for $s \geq t$ (zero otherwise), and $\widetilde{W}(\rho_Y, \rho_\delta) = B(\rho_Y)W(\rho_\delta)$ transforms the treatment design matrix. We can efficiently maximize this marginal likelihood using standard numerical optimization routines.

In the second step, the sufficient statistic $\widehat{\lambda}_i(\rho)$ has been derived in Section 2.3: see equation (5). For the empirical Bayes estimator, we exploit Tweedie's formula (Robbins, 1951; Efron, 2011) to compute the posterior mean of each unit's random coefficients λ_i,

$$\mathbb{E}\left[\lambda_i \mid Y_{i,0:T}, t_0, \rho, p\right] = \widehat{\lambda}_i(\rho) + \Sigma_{V,i}(\theta)\frac{\partial}{\partial \widehat{\lambda}_i(\rho)} \log p\left(\widehat{\lambda}_i(\rho) \mid Y_{i0}\right). \tag{7}$$

The first term is the OLS/MLE estimate and the sufficient statistic $\widehat{\lambda}_i(\rho)$, while the second term is a Bayes correction that depends on the derivative of the marginal density of the sufficient statistics $\widehat{\lambda}_i(\rho) \mid Y_{i0}$. The correction term adapts to the local shape of the marginal density of $\widehat{\lambda}_i(\rho) \mid Y_{i0}$: a positive derivative indicates the estimate falls below the mode so we shrink upward, while a negative derivative indicates it lies above the mode so we shrink downward. Moreover, steeper slopes, i.e., higher density concentration, yield larger corrections, whereas flatter regions induce milder shrinkage.

With fixed T in event studies, the unit-specific parameters λ_i cannot be consistently estimated; instead, the empirical Bayes estimator helps efficiently combine information across all units to shrink and refine these estimates, thereby reducing the overall compound risk. Crucially, Tweedie's formula circumvents the challenge to deconvolve the latent coefficient

density $\pi(\lambda_i \mid Y_{i0})$; one only needs to estimate the marginal density of the observable quantities $\left(\widehat{\lambda}_i(\rho), Y_{i0}\right)$.[4] In practice, this marginal can be fit parametrically, such as plugging in the Gaussian form implied by the QMLE, or nonparametrically via kernel or mixture methods. The former is easier to implement, while the latter helps reveal richer heterogeneity patterns. The resulting empirical Bayes estimator shrinks the noisy OLS/MLE $\widehat{\lambda}_i(\rho)$ toward a data-driven prior and attains ratio optimality, i.e., its compound risk is asymptotically equivalent to the oracle risk, where one knows the true conditional distribution of λ_i.

3.2 Asymptotics for QMLE

We now establish that the QMLE in the first step is consistent and asymptotically normal.

Assumption 3.1 *(Estimation)*

(a) U_{it} *and* ε_{ij} *follow Gaussian distributions with* $\sigma_U^2, \sigma_\varepsilon^2 > 0$.

(b) (λ_i, Y_{i0}) *have finite fourth moment.*

This Gaussianity condition (a) is imposed for the two-step estimator, not for identification. Nonparametric identification in Theorem 2.1 only requires a non-vanishing characteristic function of the composite noise, regardless of its exact distribution. In more general specifications with additional covariates, we need only conditional Gaussianity of $\{U_{it}, \varepsilon_{ij}\}$ given those covariates. Furthermore, if one forgoes the AR(p) dimension reduction and instead directly estimates the full vector of $\{\delta_{ij}\}$, the normality of ε_{ij} can also be dispensed with. However, when employing the AR-based reduction, where $V_i(\rho)$ is a linear combination of U_{it} and ε_{ij}, we require that this composite noise lie in an exponential family, such as Gaussian, to obtain the Tweedie's formula for the empirical Bayes estimator.

Let $\eta = (\theta', b_0', b_1', \operatorname{vech}(\Sigma_\lambda)')'$ collect both the common parameters and the Gaussian prior parameters, and η_0 be the pseudo-true value of η. For the prior parameters, $b_{0,0}$ and $b_{1,0}$ are those that minimize the Kullback-Leibler distance between the true conditional distribution of $\lambda_i \mid Y_{i0}$ and the working Gaussian regression. Equivalently, $b_{1,0}$ is the best linear predictor coefficient of λ_i on Y_{i0} and $b_{0,0} = \mathbb{E}[\lambda_i] - b_{1,0}\mathbb{E}[Y_{i0}]$, while $\Sigma_{\lambda,0}$ is the corresponding residual covariance.

[4] Since the conditional and joint log densities differ only by a constant that drops out under differentiation, we can work with $\log p\left(\widehat{\lambda}_i, Y_{i0}\right)$ instead of $\log p\left(\widehat{\lambda}_i \mid Y_{i0}\right)$ in practice.

Theorem 3.1 (QMLE) *Under Assumptions 2.1-2.3 and 3.1,*

$$\widehat{\eta} \xrightarrow{p} \eta_0, \quad \sqrt{N}\left(\widehat{\eta} - \eta_0\right) \xrightarrow{d} \mathcal{N}\left(0,\, H(\eta_0)^{-1} G(\eta_0) H(\eta_0)^{-1}\right),$$

where

$$H(\eta_0) = -\mathbb{E}\left[\nabla_\eta^2 \ell_i(\eta_0)\right], \quad G(\eta_0) = \mathbb{E}\left[\nabla_\eta \ell_i(\eta_0) \nabla_\eta \ell_i(\eta_0)'\right],$$

and ℓ_i is the marginal quasi-log-likelihood of $Y_{i,1:T}$. The asymptotic variance of $\widehat{\theta}$ is obtained by taking the corresponding sub-block of this sandwich matrix.

The intuition is in line with standard M-estimation arguments applied to a pseudo-likelihood: the identification and moment conditions ensure a unique maximizer and uniform convergence of the score, while smoothness guarantees a valid Taylor expansion of the log-likelihood. The resulting sandwich-form variance reflects potential misspecification of the prior. Note that the there is no Nickell bias for the marginal likelihood after integrating out λ_i, although there is for the conditional likelihood: see also the robust QMLE discussion in Alvarez and Arellano (2022).

3.3 Ratio optimality for empirical Bayes

In this subsection, we show that the empirical Bayes estimator in the second step achieves oracle risk performance.

Define the risk for any estimator $\widetilde{\lambda}_{1:N}$ and the oracle risk as follows:

$$R_N(\widetilde{\lambda}_{1:N};\, \theta_0, \pi_0) = \mathbb{E}_{\theta_0,\pi_0}\left[\sum_{i=1}^N \|\widetilde{\lambda}_i - \lambda_i\|^2\right], \quad R_N^{\text{oracle}}(\theta_0, \pi_0) = \mathbb{E}_{\theta_0,\pi_0}\left[\sum_{i=1}^N \text{Var}_{\theta_0,\pi_0}(\lambda_i \mid Y_{i,0:T})\right],$$

where the subscripts (θ_0, π_0) indicate that the expectation and variance are under the true data generating law $\mathbb{P}_{\theta_0,\pi_0}$. θ_0 and π_0 are unknown to the econometrician but fixed in the DGP. Let the leave-one-out kernel estimator be

$$\widehat{p}_{(-i)}(\widehat{\lambda}_i(\rho), y_{i0}) = \frac{1}{N-1}\sum_{j\neq i} \frac{1}{B_N^d} \phi\left(\tfrac{\widehat{\lambda}_j(\rho) - \widehat{\lambda}_i(\rho)}{B_N}\right) \phi\left(\tfrac{Y_{j0} - y_{i0}}{B_N}\right),$$

with $d = \dim(\widehat{\lambda}) + 1$, and the empirical Bayes estimator for λ_i be

$$\widetilde{\lambda}_i = \left[\widehat{\lambda}_i(\widehat{\rho}) + \left(\widehat{\Sigma}_{V,i} + B_N^2 I_{\dim(\widehat{\lambda}_i)} \right) \frac{\partial}{\partial \widehat{\lambda}_i(\widehat{\rho})} \log \widehat{p}_{(-i)} \left(\widehat{\lambda}_i(\widehat{\rho}) \mid Y_{i0} \right) \right]_{C_N}, \tag{8}$$

where $\widehat{\Sigma}_{V,i}$ is given in Algorithm 1, and $[\cdot]_{C_N}$ means truncate the vector inside to lie within the Euclidean ball of radius C_N.

We adopt Assumptions 3.2–3.6 of Liu, Moon, and Schorfheide (2020), restated as Assumptions A.1–A.5 in Appendix A.3. First, exponential tails for (λ_i, Y_{i0}) ensure that the probability mass trimmed away at $\|\lambda_i\| > C_N$ vanishes as $N \to \infty$. Second, trimming and bandwidth rates (C_N, C_N', B_N) balance kernel bias and variance. Third, smoothness of $\pi(Y_{i0} \mid \lambda_i)$ prevents sharp spikes in the distribution of Y_{i0}. Together, these conditions ensure that the leave-one-out density $\widehat{p}_{(-i)}$ is consistent. Fourth, posterior-mean truncation ensures that the empirical Bayes procedure remains stable by preventing outlier units with extreme estimates from dominating the overall performance, thereby maintaining uniform control over the risk across all possible priors. Finally, $\sqrt{N}$-consistency of the common parameters $\widehat{\theta}$ follows from the QMLE result in Theorem 3.1.

Theorem 3.2 (Ratio optimality) *Let θ_0 denote the unknown true parameter, treated as fixed in the DGP. Under Assumptions 2.1–2.3, 3.1, and A.1–A.5, the empirical Bayes estimator $\widetilde{\lambda}_{1:N}$ in (8) achieves ε_0-ratio optimality uniformly over $\pi_0 \in \Pi$: for any $\varepsilon_0 > 0$,*

$$\limsup_{N \to \infty} \sup_{\pi_0 \in \Pi} \frac{R_N(\widetilde{\lambda}_{1:N};\, \theta_0, \pi_0) - R_N^{\text{oracle}}(\theta_0, \pi_0)}{N \mathbb{E}_{\theta_0, \pi_0}[\, Var_{\theta_0, \pi_0}(\lambda_i \mid Y_{i,0:T})] + N^{\varepsilon_0}} \leq 0.$$

In a decision theoretic framework for compound risk, our event study estimator attains ratio optimality, meaning that its overall risk converges to the infeasible oracle benchmark up to vanishing terms. In other words, the mean squared error of our empirical Bayes shrinkage estimator is asymptotically equivalent to the minimum possible risk one would achieve if the true distribution of $\lambda_i \mid Y_{i0}$ were known. Our analysis builds on the foundational work of Brown and Greenshtein (2009) on compound decision problems, the refinements by Jiang and Zhang (2009), and the recent dynamic panel extension of Liu, Moon, and Schorfheide (2020).

4 Extensions and tests

4.1 Extensions

Beyond the baseline specification in (1) and (2), our proposed method accommodates various extensions to address richer policy questions and realistic data features. First, one can generalize the treatment indicator D_{it}^j to discrete or continuous dosages Z_{it}^j, accommodate staggered adoption designs by allowing treatment timing to vary across units, incorporate time fixed effects γ_t further controls for common shocks, and estimate δ_{ij} for $j \in \{-L, \ldots, -1\}$ to partially check for the no anticipation assumption.

Second, additional covariates X_{it} can be woven into both the QMLE and empirical Bayes steps. For strictly exogenous controls X_{it}^O, their coefficients can be either common or unit-specific, whereas for predetermined covariates X_{it}^P, they can only have common coefficients to ensure identification. These covariate extensions allow researchers to flexibly adjust for observed confounders while still exploiting the shrinkage benefits of empirical Bayes.

Third, the dynamic structure itself can be enriched. Both the outcome process Y_{it} and the treatment effect sequence δ_{ij} may follow AR(p) dynamics; in particular, AR(2) specifications capture potential non-monotonic or oscillatory responses that simple AR(1) models miss. Moreover, the error term U_{it} can be generalized to admit cross-sectional heteroskedasticity $\sigma_{U,i}^2$ (see for example, Chen (2022) and Liu (2023)) or temporal dependence via MA(q) processes, improving finite sample inference under complex serial correlation patterns.

Finally, our empirical Bayes prior can conditional on various observables: one can consider $\pi(\lambda_i \mid C_i)$, where conditioning variables C_i can include the initial outcome Y_{i0}, treatment timing and size D_{it}^j or Z_{it}^j, whole time series paths of strictly exogenous covariates $X_{i,0:T}^O$, and initial values of predetermined covariates X_{i0}^P. Under a *conditional strict exogeneity assumption*, namely, the error terms U_{it} is independent of the treatment conditional on $(X_{i,0:T}^O,\, X_{i0}^P)$, these extensions preserve identification and capture richer sources of heterogeneity across units.

4.2 Tests

Our analysis not only delivers flexible estimates of treatment effect heterogeneity but also provides a unified toolkit for formally testing model specifications and key event study assumptions.

In terms of model specification, first, we can examine whether we have random coefficients, where λ_i is uncorrelated with Y_{i0}, against correlated random coefficients ($H_0\colon b_1 = 0$).[5] We can test whether there is no correlation between heterogeneous effects and individual heterogeneity, where λ_i is uncorrelated with Y_{i0} and δ_{ij} is uncorrelated with α_i conditional on Y_{i0} ($H_0\colon b_1 = 0, \Sigma_{\lambda,12} = 0$). Third, we can check the absence of state dependence in δ_{ij} ($H_0\colon \rho_{\delta 1} = \rho_{\delta 2} = 0$). See Table 2 for the size and power of these tests in our simulation study, and Table 4 for their performance in the county-level recession and unemployment application.

In terms of common event study assumptions, first, as discussed in Remark 2.1, the parallel trends assumption, such as Assumption 1 in Sun and Abraham (2021), amounts to zero persistence in Y_{it} absent treatment ($H_0\colon \rho_Y = 0$). Second, the no anticipation assumption, such as Assumption 2 in Sun and Abraham (2021), requires that $\mathbb{E}[\delta_{ij}] = 0$ for $j < 0$, which can be tested by verifying that pre-treatment event time coefficients have zero mean. Third, the homogeneous treatment effects assumption, such as Assumption 3 in Sun and Abraham (2021), implies identical mean treatment paths across cohorts defined by treatment timing, which can be assessed by comparing the estimated means of δ_{ij} across these cohorts.

5 Monte Carlo simulations

5.1 Alternative estimators and DGPs

Alternative estimators. In our simulation study, we evaluate two broad groups of estimators for time-varying treatment effects in event studies: the homogeneous treatment effect estimators and the heterogeneous treatment effect ones. For simplicity, we focus below on the basic setup of Section 2.2 without time fixed effects and additional covariates, and extensions to the generalized model in Section 4.1 can be carried out in a similar manner.

The first group comprises the traditional TWFE without any lagged outcome and an augmented version with an AR(1) term. The baseline *TWFE* regresses the observed outcome

[5] Uncorrelation is a necessary but not sufficient condition for independence, making this a more conservative test.

Y_{it} on event time dummies and unit fixed effects,

$$Y_{it} = \sum_{j=-L}^{J} D_{it}^{j} \delta_j + \alpha_i + U_{it},$$

normalizing the pre-treatment period by setting $\delta_{-1} = 0$. While straightforward, omitting dynamics can lead to omitted variable bias when outcomes are serially correlated. To mitigate this bias, we introduce an augmented *TWFE+AR(1)* estimator, which includes a lagged outcome $Y_{i,t-1}$ as an additional regressor,

$$Y_{it} = \rho_Y Y_{i,t-1} + \sum_{j=-L}^{J} D_{it}^{j} \delta_j + \alpha_i + U_{it}, \quad \text{normalizing } \delta_{-1} = 0,$$

while still consider a common effect δ_j across units.

The second class of estimators allows for unit-specific dynamic responses as in (1). We consider the following four heterogeneous treatment effect estimators, which differ in how they recover the marginal density of the sufficient statistics $p(\widehat{\lambda} \mid Y_0)$ in Tweedie's formula (7). The *oracle* estimator knows the true distribution and the true common parameters, and thus attains the infeasible optimum to which we benchmark our feasible estimator. The *parametric* estimator adopts a parametric form of the distribution, typically Gaussian, which is in line with the QMLE and easy to implement. The nonparametric estimator models the distribution via *kernel* or *mixture* and offers flexibility to uncover complex heterogeneity patterns at the cost of longer computation time and higher variance.[6] Our main focus is on the parametric and nonparametric approaches.

DGPs. We simulate panel data according to a dynamic event study model in (1), and the treatment effect sequence $\{\delta_{ij}\}_{j=0}^{m}$ follows an AR(p) process,

$$\delta_{ij} = \sum_{k=1}^{p} \rho_{\delta p} \delta_{i,j-p} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2),$$

for $j = p, \ldots, J$, with initial draws δ_{i0}.

In our baseline design, we set the cross-sectional sample size to $N = 1000$, the time series

[6]For the kernel estimator, we use a Gaussian kernel with bandwidth chosen by Silverman's rule of thumb, which performs well in our simulations and empirical application, although more advanced bandwidth selection methods could further improve its estimation accuracy.

dimension to $T = 10$, the treatment onset to $t_0 = 5$, and the maximum event horizon to $J = 5$. The common parameters are $\rho_Y = 0.8$, $\sigma_U^2 = 1/T$, and $\sigma_\epsilon^2 = 1/T$.

For the distribution of unit-specific parameters $\pi(\lambda \mid Y0)$, we take into account the following four aspects that capture different heterogeneity and state dependence patterns. First, we explore both normal and non-normal distributions. Second, we examine both a random coefficients (RC) setup with $\lambda_i \perp Y_{i0}$ and a correlated random coefficients (CRC) setup with $\lambda_i \not\perp Y_{i0}$. Third, we investigate scenarios where α_i and δ_i are either independent or correlated conditional on Y_{i0}. Finally, we consider both AR(1) and AR(2) for state dependence in the treatment effect dynamics. For the AR(2), we specify four cases: $(\rho_{\delta,1}, \rho_{\delta,2}) =$ (0, 0) in Case 1 for no state dependence, (0.3, 0) in Case 2 for pure AR(1), (0.5, 0.2) in Case 3 for a monotonic decay, (0.75, -0.25) in Case 4 for an oscillation response, all with initial means $\mathbb{E}[\delta_{i0}] = 3$ and $\mathbb{E}[\delta_{i1}] = 1.5$. For each experimental setup, we execute $N_{\text{sim}} = 100$ Monte Carlo simulations.

5.2 Results

In the main text, we focus on the common parameter estimates, joint distribution of the individual heterogeneity, time-varying treatment effects, and tests, for the specifications with non-normal distribution, correlated random coefficients, $\alpha_i \not\perp \delta_i \mid Y_{i0}$, and $\delta_{ij} \sim \text{AR}(2)$. For detailed results across all model specifications, please refer to the online appendix. The main messages are similar across all specifications.

Table 1 reports the bias, standard error, and RMSE of the QMLE for the common parameters. Standard errors are computed using the robust QMLE variance formula from Theorem 3.1. Across all four cases, the QMLE exhibits small bias and variance with RMSE below 0.05 for every parameter.

Figures 2 and 3 plot the joint distribution of the empirical Bayes estimates $\widetilde{\lambda}_i = (\widetilde{\alpha}_i, \widetilde{\delta}_{i0}, \widetilde{\delta}_{i1})$ via their pairwise marginal heatmaps, for the random coefficients and correlated random coefficients designs, respectively.[7] The rows correspond to $(\widetilde{\alpha}_i, \widetilde{\delta}_{i0})$, $(\widetilde{\alpha}_i, \widetilde{\delta}_{i1})$, and $(\widetilde{\delta}_{i0}, \widetilde{\delta}_{i1})$, from top to bottom, and the columns show the oracle, parametric, kernel, and mixture empirical Bayes estimators, from left to right. All three feasible empirical Bayes estimators produce very similar heatmaps that closely track the oracle benchmark and successfully capture the

[7]Note that the distribution $p(\widetilde{\lambda})$ differs from $\pi(\lambda)$. The former is based on the empirical Bayes posterior means and embeds information from each unit's observed sequence.

Table 1: Common parameter estimates by QMLE - Monte Carlo

	Case 1			Case 2		
	Bias	SD	RMSE	Bias	SD	RMSE
ρ_Y	0.000	0.002	0.002	0.001	0.003	0.003
$\rho_{\delta 1}$	0.000	0.014	0.014	0.023	0.022	0.032
$\rho_{\delta 2}$	0.000	0.008	0.008	-0.012	0.012	0.017
σ_U^2	0.000	0.003	0.003	0.000	0.003	0.003
σ_ϵ^2	-0.001	0.005	0.005	0.003	0.005	0.006
	Case 3			Case 4		
	Bias	SD	RMSE	Bias	SD	RMSE
ρ_Y	0.003	0.005	0.006	0.004	0.003	0.005
$\rho_{\delta 1}$	0.037	0.024	0.043	0.027	0.016	0.031
$\rho_{\delta 2}$	-0.028	0.014	0.031	-0.015	0.008	0.017
σ_U^2	-0.001	0.002	0.003	-0.002	0.002	0.003
σ_ϵ^2	0.019	0.006	0.020	0.038	0.006	0.038

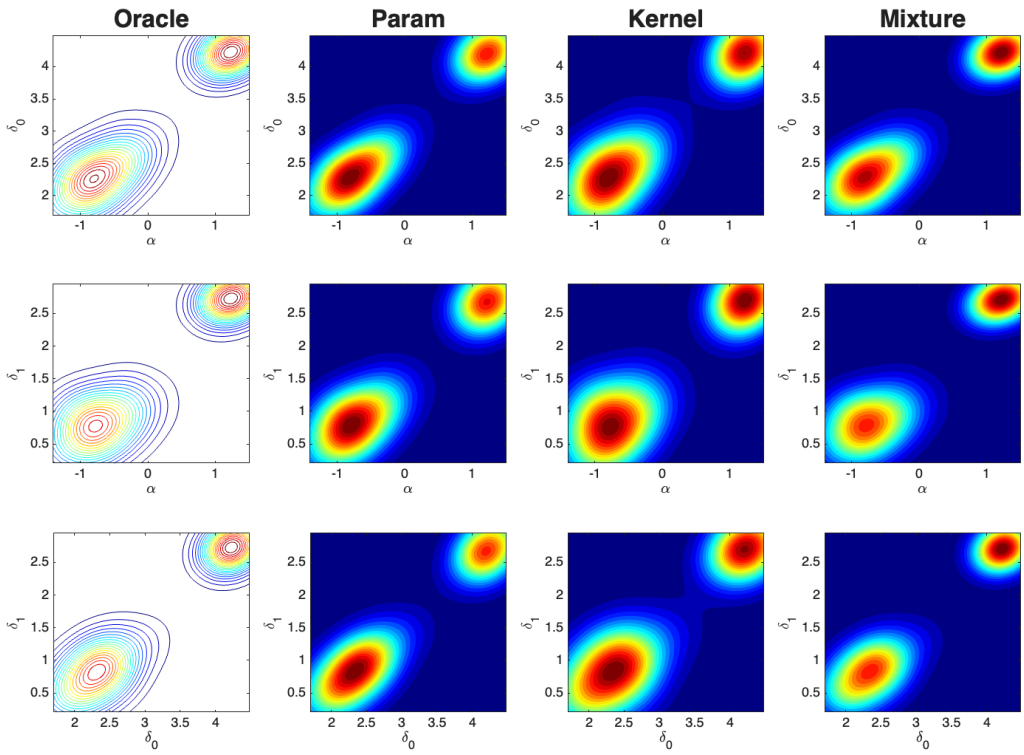
Notes: DGP: Non-normal, CRC, $\alpha_i \not\perp \delta_i \mid Y_{i0}$, $\delta_{ij} \sim$ AR(2). $(\rho_{\delta,1}, \rho_{\delta,2}) = (0, 0)$ in Case 1, (0.3, 0) in Case 2, (0.5, 0.2) in Case 3, (0.75, -0.25) in Case 4. Initial means: $\mathbb{E}[\delta_{i0}] = 3$, $\mathbb{E}[\delta_{i1}] = 1.5$.

bimodal pattern in Figure 2 and the heavy tail behavior in Figure 3. Quantitatively, the mixture estimator achieves the lowest RMSE for λ_i, with roughly a 5–10% improvement over both the parametric and kernel approaches. The parametric estimator shows a slightly larger bias due to its misspecified Gaussian prior, and the kernel estimator exhibits slightly higher variance due to its nonparametric setup.

Figure 4 displays the estimated heterogeneous dynamic treatment effect paths across event time for four DGP scenarios. From top to bottom, the rows show Cases 1–4: no state dependence, pure AR(1), monotonic AR(2), and oscillatory AR(2). From left to right, the columns present the infeasible optimum oracle estimator, followed by the parametric, kernel, and mixture empirical Bayes estimators, as well as the homogeneous TWFE and TWFE+AR(1) estimators. In each graph, the thin lines depict the heterogeneous dynamic responses of the individual units. As before, all feasible empirical Bayes estimators yield trajectories nearly indistinguishable from the oracle benchmark and accurately recover each DGP's dynamic patterns, whether simple exponential decay, gradual tapering, or sign-changing oscillation, thereby recovering substantial dynamic heterogeneity across units.

The last two columns are the homogeneous estimators. The baseline TWFE estimator fails to account for the dynamics in the outcome, and produces substantial misspecification bias with larger and more persistent estimated effects. For the augmented TWFE+AR(1), its

Figure 2: Joint distribution of $\widetilde{\lambda}_i$ - Monte Carlo, random coefficients



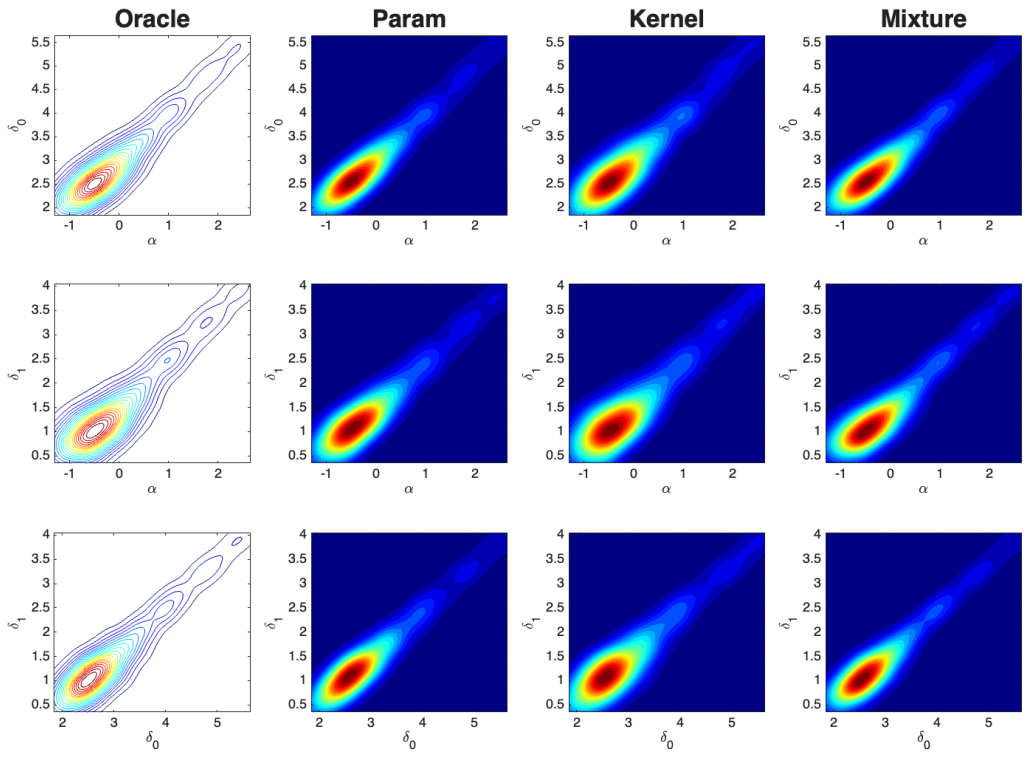
Notes: DGP: Non-normal, $\alpha_i \not\perp \delta_i \mid Y_{i0}$, $\delta_{ij} \sim$ AR(2). Case 3: $(\rho_{\delta,1}, \rho_{\delta,2}) = (0.5,\ 0.2)$. Initial means: $\mathbb{E}[\delta_{i0}] = 3$, $\mathbb{E}[\delta_{i1}] = 1.5$.

estimated mean path aligns closely with the true mean pattern, but it is not able to capture the cross-unit dispersion, and its 95% confidence bands are too narrow to reflect underlying heterogeneity. In contrast, our empirical Bayes estimators efficiently combine information across all units and flexibly adapt to each unit's own response profile, and thus deliver good estimates of the average treatment path and effectively capture the heterogeneity in dynamics.

Table 2 reports the rejection rates over 100 simulations for three tests regarding the heterogeneity pattern. As described in Section 4.2, Test 1 checks for random versus correlated random coefficients, Test 2 for the joint independence of δ_{ij} against (α_i, Y_{i0}), and Test 3 for the state dependence in the treatment effect processes.

The table is partitioned into three blocks. The left block reports rejection rates under a random coefficients DGP in which $\alpha_i \perp \delta_i \mid Y_{i0}$, satisfying the null hypotheses of Tests 1 and 2. The middle block corresponds to a random coefficients DGP with $\alpha_i \not\perp \delta_i \mid Y_{i0}$, which satisfies Test 1's null but violates Test 2's. The right block is based on a correlated

Figure 3: Joint distribution of $\widetilde{\lambda}_i$ - Monte Carlo, correlated random coefficients



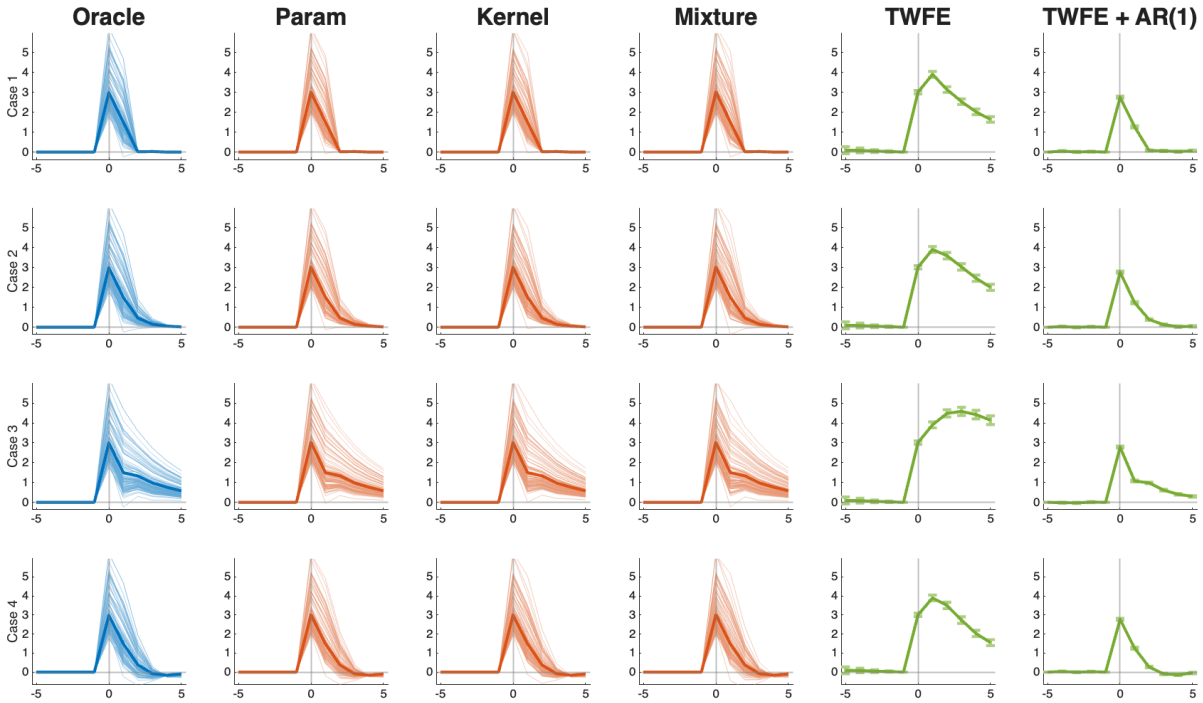
Notes: DGP: Non-normal, $\alpha_i \not\perp \delta_i \mid Y_{i0}$, $\delta_{ij} \sim$ AR(2). Case 3: $(\rho_{\delta,1}, \rho_{\delta,2}) = (0.5,\ 0.2)$. Initial means: $\mathbb{E}[\delta_{i0}] = 3$, $\mathbb{E}[\delta_{i1}] = 1.5$.

random coefficients DGP with $\alpha_i \not\perp \delta_i \mid Y_{i0}$, violating the nulls of both Tests 1 and 2. Within each block, columns give results for Cases 1–4: no AR, AR(1), monotonic AR(2), oscillatory AR(2), where Case 1 conforms to Test 3's null and Cases 2–4 lie under its alternative.

Together, the blue entries indicate the size of the tests, while the black entries show their power. Under the null hypotheses, all tests maintain size close to the nominal 5 % level, with rejection rates between 0.04 and 0.06.[8] Under the alternative hypotheses, the power is 1.00, possibly due to the relatively large sample size with $N = 1000$ and $T = 10$. Therefore, these tests provide a reliable means of diagnosing the heterogeneity pattern and state dependence structure. In particular, these tests allow us to assess whether treatment effect dynamics are driven primarily by unobserved baseline heterogeneity or by the initial treatment impact.

[8]One observed size of 0.02 likely reflects Monte Carlo noise with only 100 repetitions.

Figure 4: Event study with time-varying treatment effects - Monte Carlo



Notes: DGP: Non-normal, CRC, $\alpha_i \not\perp \delta_i \mid Y_{i0}$, $\delta_{ij} \sim$ AR(2). $(\rho_{\delta,1}, \rho_{\delta,2}) = (0, 0)$ in Case 1, (0.3, 0) in Case 2, (0.5, 0.2) in Case 3, (0.75, -0.25) in Case 4. $\mathbb{E}[\delta_{i0}] = 3$, $\mathbb{E}[\delta_{i1}] = 1.5$. TWFE and TWFE+AR(1): bars indicate 95% CI, clustered s.e. by unit.

6 Empirical example: recession and unemployment

6.1 Data and sample

Understanding how recessions shape local labor markets is crucial for designing targeted policy responses. The 2008 Great Recession led to a nationwide spike in unemployment, peaking at nearly 10% in October 2009, and ushered in a protracted recovery that saw the national rate fall back to pre-crisis levels only by late 2015.[9] However, aggregate figures mask substantial variation across regions: some counties experienced sharp spikes, while others bore delayed and milder losses. For example, Yagan (2019) documents long-lasting employment and earnings losses for harder-hit areas, and Hershbein and Stuart (2020) further show that those areas also experienced persistent population declines.

In this empirical example, we exploit county-level unemployment data to map these heterogeneous responses over time. Our outcome, Y_{it}, is the annual unemployment rate for

[9]See the BLS website, such as `https://www.bls.gov/spotlight/2012/recession/pdf/recession_bls_spotlight.pdf` and `https://www.bls.gov/news.release/archives/empsit_01082016.pdf`

Table 2: Rejection rates of tests - Monte Carlo

	RC, $\alpha_i \perp \delta_i \mid Y_{i0}$				RC, $\alpha_i \not\perp \delta_i \mid Y_{i0}$				CRC, $\alpha_i \not\perp \delta_i \mid Y_{i0}$			
Case	1	2	3	4	1	2	3	4	1	2	3	4
Test 1	0.04	0.04	0.04	0.06	0.06	0.06	0.05	0.04	1.00	1.00	1.00	1.00
Test 2	0.05	0.06	0.04	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test 3	0.04	1.00	1.00	1.00	0.03	1.00	1.00	1.00	0.02	1.00	1.00	1.00

Notes: DGP: Non-normal, $\delta_{ij} \sim$ AR(2). $(\rho_{\delta,1}, \rho_{\delta,2}) = (0, 0)$ in Case 1, (0.3, 0) in Case 2, (0.5, 0.2) in Case 3, (0.75, -0.25) in Case 4. Initial means: $\mathbb{E}[\delta_{i0}] = 3$, $\mathbb{E}[\delta_{i1}] = 1.5$. Blue entries: size; black entries: power. Based on robust s.e.

Table 3: Common parameter estimates by QMLE - recession and unemployment example

	Est.	SD		Est.	SD
ρ_Y	0.845	(0.010)	σ_U^2	0.431	(0.103)
$\rho_{\delta 1}$	0.306	(0.011)	σ_ϵ^2	0.276	(0.094)
$\rho_{\delta 2}$	-0.061	(0.011)			

county i in year t. We define the onset of the Great Recession as 2008, assigning it to period $t_0 = 5$ within a ten year window. The sample spans 2003–2013 ($T = 10$) across $N = 3142$ U.S. counties, capturing five pre- and five post-recession years. The county-level not seasonally adjusted unemployment rates are obtained from the Bureau of Labor Statistics (BLS) website, and we aggregate the monthly data to an annual frequency by time averaging.

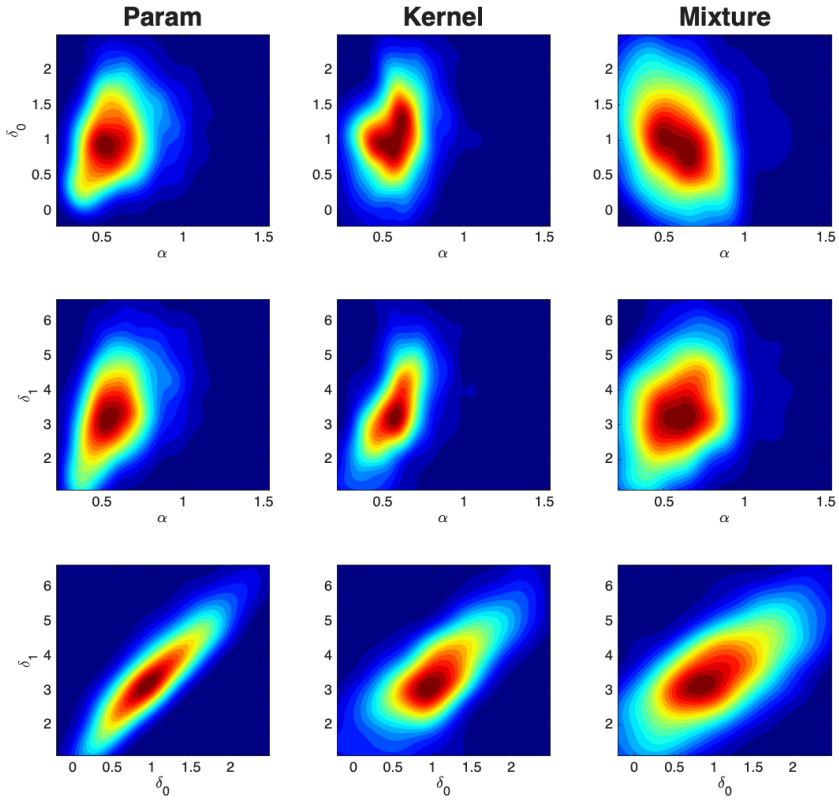
This panel event study analysis allows us to estimate county-specific dynamic effects while controlling for unobserved heterogeneity and serial dependence, thereby shedding light on both the immediate and persistent impacts of the recession across diverse local economies.

6.2 Results

In this section, we focus on the estimator under the AR(2) specification for δ_{ij}. Analogous results for the AR(1) case and models with time fixed effects are provided in the online appendix.

In Table 3 for common parameter estimates, the estimated persistence in the unemployment rate is high and significant with $\widehat{\rho}_Y = 0.845$, so the omitted variable bias could be substantial for the traditional TWFE regression. The AR(2) dynamics of the recessionary effect are likewise significant with $\widehat{\rho}_{\delta 1} = 0.306$ and $\widehat{\rho}_{\delta 2} = -0.061$, indicating a damped

Figure 5: Joint distribution of $\widetilde{\lambda}_i$ - recession and unemployment example

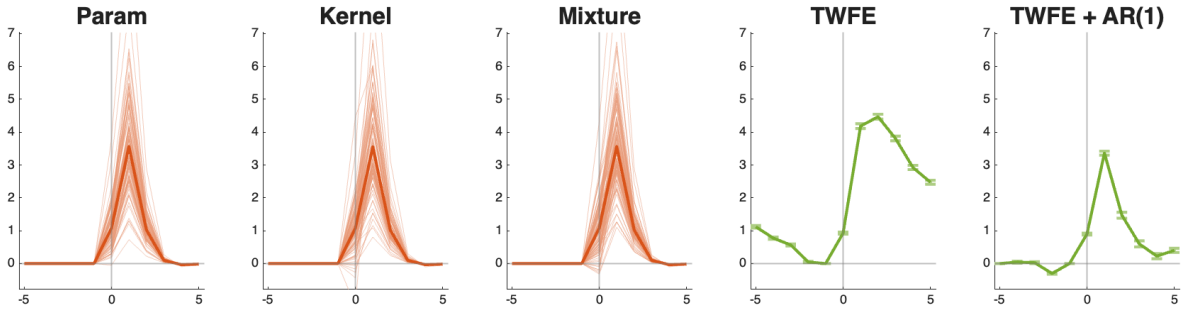


oscillatory decay in local labor market responses.

Figure 5 presents the heatmaps of the joint distributions of empirical Bayes posterior means across the parametric, kernel, and mixture estimators. All three estimators yield qualitatively similar density shapes. The heatmaps also reveal strong non-Gaussian heterogeneity with asymmetric mass and possible heavy tails rather than simple elliptical contours. In the first two rows, counties with higher baseline heterogeneity α_i tend to exhibit larger initial effects $(\delta_{i0}, \delta_{i1})$, indicating that areas already suffering from high unemployment were hit hardest by the recession. The third row shows a strong positive correlation between δ_{i0} and δ_{i1}, reflecting persistent temporal dynamics in treatment responses. These irregular patterns underscore the value of the flexible empirical Bayes methods for jointly modeling $(\alpha_i, \delta_{i0}, \delta_{i1})$ and uncovering the rich heterogeneity across counties.

Figure 6 plots county-specific event study estimates of the time-varying treatment effects. As seen in the joint distributions, all three empirical Bayes estimators produce qualitatively similar trajectories. The individual curves reveal stark heterogeneity: some counties suffered a dramatic jump in unemployment of over 7 percentage points in 2009, others experienced

Figure 6: Event study w. time-varying treatment effects - recession & unemployment example



Notes: TWFE and TWFE+AR(1): bars indicate 95% CI, clustered s.e. by unit.

Table 4: Tests - recession and unemployment example

	Test stat	Crit. val.	Reject?
Test 1	672.6	5.99	Y
Test 2	766.7	9.49	Y
Test 3	1069.2	5.99	Y

Notes: Based on QMLE estimates with robust s.e. Test 1: $H_0: b_1 = 0$; Test 2: $H_0: b_1 = 0, \Sigma_{\lambda,12} = 0$; Test 3: $H_0: \rho_{\delta 1} = \rho_{\delta 2} = 0$. Critical values: 5% level.

only modest rises of around 0.5 points, and a few even registered slight declines in the initial recession year 2008. These spikes and the varied post-2008 decay profiles far exceed the average effect implied by the TWFE model. In particular, the baseline TWFE yields pre-2008 coefficients that are significantly different from zero, indicating substantial omitted variable bias from ignoring the serial dependence of unemployment.

Finally, Table 4 formally tests three key modeling assumptions: see Section 4.2 for a more detailed description of the tests. The rejections of all three tests reveal several key features of the Great Recession's impact on U.S. local labor markets. First, rejecting the pure random coefficients null (Test 1) shows that the unobserved heterogeneity, including the treatment effects, is not idiosyncratic but instead systematically related to county characteristics: places with higher pre-crisis unemployment were hit especially hard. Second, the rejection of the joint independence null (Test 2) confirms a strong link between baseline heterogeneity and dynamic responses, indicating that local labor market resilience or vulnerability cannot be treated as exogenous. Finally, ruling out the no state dependence null (Test 3) demonstrates that the recessionary impact on local labor markets is not a one-off hit but unfolds dynamically, with early effects shaping subsequent recovery or further distress.

Together, these results highlight the inadequacy of homogeneous static TWFE specifications and validate the need for our dynamic heterogeneous panel framework.

7 Conclusion

In summary, our paper makes three key contributions. First, we demonstrate how omitting predetermined variables can severely bias event study estimates, and we introduce a semiparametric dynamic panel model with correlated random coefficients that simultaneously captures outcome persistence and treatment effect heterogeneity. Second, we develop a two-step estimator—QMLE for common parameters followed by an empirical Bayes correction for unit-specific effects—that is easy to implement and achieves oracle risk performance. Finally, our analysis offers new insights into standard event study assumptions, including no anticipation, homogeneous treatment effects across treatment timing cohorts, and state dependence structure, making it easier to diagnose and address potential violations in empirical research.

The potential applications of our method extend to any setting with short panel data where we are interested in the dynamics of the heterogeneous treatment effects. In corporate finance, it can revisit classic event studies of earnings announcements, mergers, or regulatory changes, allowing for firm-level persistence and heterogeneous responses. In public policy, it can evaluate staggered social program roll-outs, uncovering differential impacts across communities or demographic groups. Likewise, research in health, education, environmental policy, labor markets, and macroprudential regulation can potentially benefit by using our semiparametric, shrinkage-based estimator to produce more accurate estimates of how treatment effects evolve over time.

References

Alvarez, J., and M. Arellano (2022): "Robust Likelihood Estimation of Dynamic Panel Data Models," *Journal of Econometrics*, 226(1), 21–61.

Anderson, T. W., and C. Hsiao (1982): "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 18(1), 47–82.

Arellano, M., and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58(2), 277–297.

Arellano, M., and S. Bonhomme (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *Review of Economic Studies*, 79(3), 987–1020.

Arellano, M., and O. Bover (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 68(1), 29–51.

Arkhangelsky, D., G. W. Imbens, L. Lei, and X. Luo (2024): "Design-Robust Two-Way-Fixed-Effects Regression for Panel Data," *Quantitative Economics*, 15(4), 999–1034.

Ashenfelter, O. C. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60(1), 47–57.

Blundell, R., and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87(1), 115–143.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 91(6), 3253–3285.

Brown, L. D., and E. Greenshtein (2009): "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Normal Means," *Annals of Statistics*, 37(4), 1684–1704.

Callaway, B., and P. H. Sant'Anna (2021): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225(2), 200–230.

Chen, J. (2022): "Empirical Bayes when estimation precision predicts parameters," *arXiv preprint arXiv:2212.14444*.

de Chaisemartin, C., and X. D'Haultfœuille (2023): "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey," *The Econometrics Journal*, 26(3), C1–C30.

Efron, B. (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106(496), 1602–1614.

Freyaldenhoven, S., C. Hansen, J. Pérez, and J. M. Shapiro (2021): "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design," Discussion Paper 29170, National Bureau of Economic Research.

Goodman-Bacon, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 225(2), 254–277.

Gu, J., and R. Koenker (2017): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Business & Economic Statistics*, 35(1), 1–16.

Hershbein, B., and B. A. Stuart (2020): "Recessions and local labor market hysteresis," .

Jiang, W., and C.-H. Zhang (2009): "General Maximum Likelihood Empirical Bayes Estimation of Normal Means," *Annals of Statistics*, 37(4), 1647–1684.

Liu, L. (2023): "Density forecasts in panel data models: A semiparametric bayesian perspective," *Journal of Business & Economic Statistics*, 41(2), 349–363.

Liu, L., H. R. Moon, and F. Schorfheide (2020): "Forecasting With Dynamic Panel Data Models," *Econometrica*, 88(1), 171–201.

Miller, D. L. (2023): "An Introductory Guide to Event Study Models," *Journal of Economic Perspectives*, 37(2), 203–230.

Robbins, H. (1951): "Asymptotically Subminimax Solutions of Compound Decision Problems," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 131–148.

Sun, L., and S. Abraham (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225(2), 175–199.

Wooldridge, J. M. (2021): "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," Discussion Paper SSRN 3906345, Department of Economics, Michigan State University, 77 pages; posted August 18, 2021.

Yagan, D. (2019): "Employment hysteresis from the great recession," *Journal of Political Economy*, 127(5), 2505–2558.

Appendix:

Time-Varying Heterogeneous Treatment Effects in Event Studies

Irene Botosaru Laura Liu

September 17, 2025

A Proofs

A.1 Identification

Proof of Theorem 2.1. We prove identification in two steps, building on the approach of Arellano and Bonhomme (2012). First, we establish identification of the parameters $\rho = (\rho_Y, \rho_\delta)'$. Second, given identified ρ, we show that the conditional density $\pi(\lambda_i \mid Y_{i0})$ is identified via characteristic function deconvolution.

Step 1: Identification of common parameters ρ. Under Assumption 2.1 for model setup, we identify ρ via the following moment conditions.

First, for the autoregressive parameter ρ_Y, under Assumption 2.3, $t_0 \geq 3$ provides at least two pre-treatment periods, and the moment condition for ρ_Y is

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{t_0-1}(Y_{it} - \rho_Y Y_{i,t-1} - \overline{Y}_t + \rho_Y \overline{Y}_{t-1})Y_{i,t-1}\right] = 0,$$

where $\overline{Y}_t = N^{-1}\sum_{i=1}^{N} Y_{it}$ helps remove the individual levels α_i. Assumption 2.2(c) ensures $\mathrm{Var}(Y_{i0}) > 0$, and thus this moment condition is non-degenerate.

Second, for treatment effect persistence ρ_δ, using treatment and post-treatment periods $t \geq t_0$, we exploit the autoregressive structure of δ_{ij}. Let $\widetilde{Y}_{it} = Y_{it} - \rho_Y Y_{i,t-1}$ denote the transformed outcome. The moment condition is:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=t_0+1}^{T}\widetilde{Y}_{it}\widetilde{Y}_{i,t-1}\right] = \rho_\delta \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=t_0+1}^{T}\widetilde{Y}_{i,t-1}^2\right].$$

Under Assumption 2.3, the condition $T - t_0 \geq J \geq 1$ ensures sufficient post-treatment

observations. Moreover, Assumption 2.2(c) ensures $\text{Var}(\delta_{i0}) > 0$, and thus this moment condition is non-degenerate.

Step 2: Identification of $\pi(\lambda_i \mid Y_{i0})$ given identified ρ. Having identified ρ in Step 1, we now verify the conditions of Theorem 2 in Arellano and Bonhomme (2012) for deconvolving the conditional density $\pi(\lambda_i \mid Y_{i0})$. The true composite error $\check{U}_{i,1:T}(\rho_{\delta,0})$ has the MA(J) structure

$$\check{U}_{it}(\rho_{\delta,0}) = U_{it} + \sum_{j=0}^{J} \sum_{k=1}^{j} \rho_{\delta,0}^{j-k} D_{it}^{j} \varepsilon_{ik}.$$

First, for their Assumption 1 (Mean independence) and Assumption 3 (Conditional independence), our simple model with common treatment timing t_0 together with Assumption 2.1(b) ensures that $\mathbb{E}[\check{U}_{it}(\rho_{\delta,0}) \mid \lambda_i, Y_{i0}] = 0$ and $\check{U}_{it}(\rho_{\delta,0}) \perp \lambda_i \mid Y_{i0}$. Since $W_i(\rho_{\delta,0})$ is deterministic and identical across units, we omit it from the conditioning set.

Second, for their Assumption 4 (Non-vanishing characteristic functions), our Assumption 2.2(a) directly imposes that the characteristic functions of $\lambda_i \mid Y_{i0}$, U_{it}, and ε_{ij} are non-vanishing almost everywhere, which extends to $\check{U}_{it}(\rho_0)$.

Third, for their Assumption 5 (MA structure), the key insight is that our composite error involves exactly $m = 2$ fundamental variance components from U_{it} and ε_{ij}, given the model structure in (1) and (2). Following from Assumption 2.2(a,b), the hessian of the log characteristic function of $\check{U}_{it}(\rho_0)$ exists almost everywhere. The hessian can be decomposed as

$$\text{vec}\left(\frac{\partial^2 \log \Psi_{\check{U}_{i,1:T}(\rho_0)}(\tau)}{\partial \tau \partial \tau'}\right) = \mathcal{S}\omega(\tau),$$

for $\tau \in \mathbb{R}^T$, where $\omega(\tau) = (\omega_U(\tau), \omega_\varepsilon(\tau))'$ with

$$\omega_U(\tau) = \frac{\partial^2 \log \Psi_U(\tau)}{\partial \tau^2}, \quad \omega_\varepsilon(\tau) = \frac{\partial^2 \log \Psi_\varepsilon(\tau)}{\partial \tau^2}.$$

The selection matrix $\mathcal{S} = S(\{D_{it}^j\}, \rho_{\delta,0})$ encodes the treatment pattern and MA lag structure.

Fourth, for their rank condition in equation (24), $\text{rank}(M_i\mathcal{S}) = m = 2$, where $M_i = \mathcal{I}_{T^2} - (W_i \otimes W_i)[(W_i \otimes W_i)'(W_i \otimes W_i)]^{-1}(W_i \otimes W_i)'$ projects out the design matrix effect. Here we suppress the dependence on ρ_0 for notational simplicity. To illustrate, consider the

minimal case $T = 4$, $t_0 = 3$, $J = 1$. The variance-covariance matrix of $\check{U}_{i,1:4}$ is

$$\Sigma_{\check{U}} = \begin{pmatrix} \sigma_U^2 & 0 & 0 & 0 \\ 0 & \sigma_U^2 & 0 & 0 \\ 0 & 0 & \sigma_U^2 & 0 \\ 0 & 0 & 0 & \sigma_U^2 + \sigma_\varepsilon^2 \end{pmatrix},$$

and the selection matrix $\mathcal{S}$ is 16×2 and encodes how $(\sigma_U^2, \sigma_\varepsilon^2)$ contribute to $\text{vec}(\Sigma_{\check{U}})$. The design matrix is

$$W_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 + \rho_{\delta,0} \end{pmatrix}.$$

One can verify that $\text{rank}(M_i\mathcal{S}) = 2$ and satisfying the identification condition.

More generally, under Assumption 2.3, $t_0 \geq 3$ provides sufficient pre-treatment and treatment periods to satisfy the degrees of freedom bound $m = 2 \leq \frac{t_0(t_0+1)}{2} - \frac{d_\lambda(d_\lambda+1)}{2}$ where $d_\lambda = 2$: see Remark 3 and equation (27) in Arellano and Bonhomme (2012). Then, the projection matrix M_i removes the variation attributable to the heterogeneous parameters λ_i, leaving sufficient variation from the two variance components $(\sigma_U^2, \sigma_\varepsilon^2)$ to achieve identification, and the rank condition $\text{rank}(M_i\mathcal{S}) = 2$ holds.

Unlike standard applications, our design matrix $W_i(\rho)$ depends on unknown ρ. Our two-step approach resolves this because the identification of ρ in Step 1 uses only the covariance structure of the data and does not require knowledge of $\pi(\lambda_i \mid Y_{i0})$.

Finally, we have the sufficient statistic representation

$$\widehat{\lambda}_i(\rho_0) = W_i(\rho_0)^+(Y_{i,1:T} - \rho_{Y,0}Y_{i,0:T-1}) = \lambda_i + V_i(\rho_0),$$

where $V_i(\rho_0) = W_i(\rho_0)^+\check{U}_{i,1:T}(\rho_0)$ is the projection noise. Since the conditions of Theorem 2 in Arellano and Bonhomme (2012) have been verified above, characteristic function deconvolution yields

$$\Psi_{\lambda_i|Y_{i0}}(\tau \mid Y_{i0}) = \frac{\Psi_{\widehat{\lambda}_i(\rho_0)|Y_{i0}}(\tau \mid Y_{i0})}{\Psi_{V_i(\rho_0)}(\tau)},$$

for $\tau \in \mathbb{R}^2$, and the conditional density is recovered via inverse Fourier transform. ■

A.2 QMLE

Proof of Theorem 3.1. As defined in the main text, $\theta = (\rho_Y, \rho_\delta, \sigma_U^2, \sigma_\varepsilon^2)'$ denotes the common parameters, and $\eta = (\theta', b_0', b_1', \mathrm{vech}(\Sigma_\lambda)')'$ collects both the common parameters and Gaussian random effects parameters. Recall that the marginal quasi-log-likelihood is

$$\ell_N(\eta) = -\frac{N}{2}\log|\Omega(\eta)| - \frac{1}{2}\sum_{i=1}^{N}\left(Y_{i,1:T} - \mu_i(\eta)\right)'\Omega(\eta)^{-1}\left(Y_{i,1:T} - \mu_i(\eta)\right), \tag{A.1}$$

where

$$\begin{aligned}
\mu_i(\eta) &= \mu_i(\rho_Y, \rho_\delta, b_0, b_1) = A(\rho_Y)Y_{i0} + \widetilde{W}(\rho_Y, \rho_\delta)(b_0 + b_1 Y_{i0}),\\
\Omega(\eta) &= \Omega(\rho_Y, \rho_\delta, \sigma_U^2, \sigma_\varepsilon^2, \Sigma_\lambda) = B(\rho_Y)\Sigma_{\check{U}}(\rho_\delta, \sigma_U^2, \sigma_\varepsilon^2)B(\rho_Y)' + \widetilde{W}(\rho_Y, \rho_\delta)\Sigma_\lambda\widetilde{W}(\rho_Y, \rho_\delta)',
\end{aligned}$$

and

$$\begin{aligned}
A(\rho_Y) &= (\rho_Y, \rho_Y^2, \rho_Y^3, \cdots, \rho_Y^T)',\\
B(\rho_Y) &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0\\ \rho_Y & 1 & 0 & \cdots & 0\\ \rho_Y^2 & \rho_Y & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_Y^{T-1} & \rho_Y^{T-2} & \rho_Y^{T-3} & \cdots & 1 \end{pmatrix},\\
\widetilde{W}(\rho_Y, \rho_\delta) &= B(\rho_Y)W(\rho_\delta).
\end{aligned}$$

Let $s = \partial\ell_N/\partial\eta$ denote the score. We now show that under correct conditional mean and covariance, the QMLE satisfies $\mathbb{E}\left[s(\eta_0) \mid Y_{i0}\right] = 0$ at the true parameter values.

(i) Random effects mean parameters b_0 and b_1. These derivatives only involve the mean.

$$\begin{aligned}
s_{b_0} &= \frac{\partial\ell_N}{\partial b_0} = \sum_{i=1}^{N}\widetilde{W}(\rho_Y, \rho_\delta)'\Omega(\eta)^{-1}(Y_{i,1:T} - \mu_i(\eta)),\\
s_{b_1} &= \frac{\partial\ell_N}{\partial b_1} = \sum_{i=1}^{N}\widetilde{W}(\rho_Y, \rho_\delta)'\Omega(\eta)^{-1}(Y_{i,1:T} - \mu_i(\eta))Y_{i0}.
\end{aligned}$$

Since $\mathbb{E}[Y_{i,1:T} - \mu_i(\eta_0) \mid Y_{i0}] = 0$, we have $\mathbb{E}[s_{b_0}(\eta_0) \mid Y_{i0}] = 0$ and $\mathbb{E}[s_{b_1}(\eta_0) \mid Y_{i0}] = 0$.

(ii) Covariance parameters $\theta_\sigma = (\sigma_U^2, \sigma_\varepsilon^2, \mathbf{vech}(\Sigma_\lambda)')'$. These derivatives only involve the covariance matrix. There are five parameters in θ_σ. For $k = 1, \ldots, 5$,

$$\begin{aligned} s_{\theta_{\sigma,k}} &= \frac{\partial \ell_N}{\partial \theta_{\sigma,k}} \\ &= -\frac{N}{2}\mathrm{tr}\left[\Omega(\eta)^{-1}\frac{\partial \Omega}{\partial \theta_{\sigma,k}}(\eta)\right] + \frac{1}{2}\sum_{i=1}^{N}(Y_{i,1:T} - \mu_i(\eta))'\Omega(\eta)^{-1}\frac{\partial \Omega}{\partial \theta_{\sigma,k}}(\eta)\Omega(\eta)^{-1}(Y_{i,1:T} - \mu_i(\eta)). \end{aligned}$$

As $\mathbb{E}[x'Ax] = \mathrm{tr}(A\mathrm{Var}(x))$ for $x \sim (0, \mathrm{Var}(x))$ and $\mathrm{Var}(Y_{i,1:T} - \mu_i(\eta_0) \mid Y_{i0}) = \Omega(\eta_0)$, the second term cancels out the first term, and we have $\mathbb{E}[s_{\theta_{\sigma,k}}(\eta_0) \mid Y_{i0}] = 0$.

(iii) Dynamic parameters ρ_δ and ρ_Y. These derivatives combine both the mean and covariance matrix. For $\rho_k \in \{\rho_\delta, \rho_Y\}$,

$$\begin{aligned} s_{\rho_k} = \frac{\partial \ell_N}{\partial \rho_k} = &\underbrace{\sum_{i=1}^{N}\frac{\partial \widetilde{W}(\rho_Y, \rho_\delta)}{\partial \rho_k}\Omega(\eta)^{-1}\left(Y_{i,1:T} - \mu_i(\eta)\right)(b_0 + b_1 Y_{i0})}_{(1)} + \underbrace{-\frac{N}{2}\mathrm{tr}\left[\Omega(\eta)^{-1}\frac{\partial \Omega}{\partial \rho_k}(\eta)\right]}_{(2)} \\ &\underbrace{+\frac{1}{2}\sum_{i=1}^{N}(Y_{i,1:T} - \mu_i(\eta))'\Omega(\eta)^{-1}\frac{\partial \Omega}{\partial \rho_k}(\eta)\Omega(\eta)^{-1}(Y_{i,1:T} - \mu_i(\eta))}_{(3)}, \end{aligned}$$

where the (1) is from the mean and $\mathbb{E}[(1) \mid Y_{i0}] = 0$ by a similar argument as in part (i), and the (2) and (3) are from the covariance matrix and $\mathbb{E}[(2) + (3) \mid Y_{i0}] = 0$ by a similar argument as in part (ii). Note that for ρ_Y, there is Nickell bias for conditional likelihood, but not for the marginal likelihood here.

Combining parts (i)–(iii), every component of the quasi-score $s(\eta)$ has zero expectation under the true DGP, as long as the first two conditional moments are correctly specified. Finally, under Assumptions 2.1–2.3 and 3.1, the strictly concave quasi-log-likelihood and pointwise LLN yield consistency by the argmax theorem, and a Taylor expansion of the score around the true parameter together with the CLT establishes asymptotic normality. ■

A.3 Ratio optimality

We adopt Assumptions 3.2–3.6 of Liu, Moon, and Schorfheide (2020), restated as in our setting as follows. First, define the slowly diverging sequence as follows.

Definition A.1 (Slowly diverging sequences)

(a) $A_N(\pi) = o_{u.\pi}(N^\epsilon)$ *for some* $\epsilon > 0$*, if there exists a sequence* $\eta_N \to 0$ *that does not depend on* $\pi \in \Pi$ *such that* $N^{-\epsilon} A_N(\pi) \leq \eta_N$*.*

(b) $A_N(\pi) = o(N^+)$*, if for every* $\epsilon > 0$*, there exists a sequence* $\eta_N(\epsilon) \to 0$ *such that* $N^{-\epsilon} A_N(\pi) \leq \eta_N(\epsilon)$*.*

(c) $A_N(\pi) = o_{u.\pi}(N^+)$*, if for every* $\epsilon > 0$*, there exists a sequence* $\eta_N(\epsilon) \to 0$ *that does not depend on* $\pi \in \Pi$ *such that* $N^{-\epsilon} A_N(\pi) \leq \eta_N(\epsilon)$*.*

Intuitively, (a) holds for some ϵ and uniformly in π, (b) holds for every ϵ but only pointwise in π, and (c) holds for every ϵ uniformly in π.

Assumption A.1 (Trimming and bandwidth)

(a) *The truncation sequence* C_N *satisfies* $C_N = o(N^+)$ *and* $C_N \geq (2 \log N)/M_2$*.*

(b) *The truncation sequence* C'_N *satisfies* $C'_N = C_N + \sqrt{(2\sigma^2 \log N)/T}$*.*

(c) *The bandwidth sequence* B_N *is bounded by* $\underline{B}_N \leq B_N \leq \overline{B}_N$*, where* $1/\underline{B}_N^2 = o(N^+)$*,* $\overline{B}_N(C_N + C'_N) = o(1)$*, and the bounds do not depend on the observed data or* $\pi_0 \in \Pi$*.*

Assumption A.2 (CRC distribution: tails) *There exist constants* $0 < M_1, M_2, M_3, M_4 < \infty$ *such that for the true distribution* $\pi_0 \in \Pi$*:*

(a) $\int_{\|\lambda\| \geq C} \pi_0(\lambda) d\lambda \leq M_1 e^{-M_2(C - M_3)}$*, and* $\int \|\lambda\|^4 \pi_0(\lambda) d\lambda \leq M_4$*.*

(b) $\int_{|y_0| \geq C} \pi_0(y_0) dy_0 \leq M_1 e^{-M_2(C - M_3)}$*, and* $\int y_0^4 \pi_0(y_0) dy_0 \leq M_4$*.*

To estimate the unknown prior nonparametrically, we trim off very large λ_i so our kernel estimates do not explode in the tails, but let the trimming threshold C_N grow slowly with N. The exponential tail bound on the prior guarantees little mass beyond C_N. Meanwhile, the kernel bandwidth B_N shrinks just fast enough to capture local features of the prior, but not so fast that variance dominates bias. Together, these conditions balance trimming and smoothing so the leave-one-out density $\widehat{p}_{(-i)}$ is consistent.

Assumption A.3 (CRC distribution: boundedness and smoothness) *The conditional density* $\pi_0(y_0 \mid \lambda)$ *is uniformly bounded and*

$$\sup_{|y_0| \leq C'_N,\, \|\lambda\| \leq C_N} \left| \frac{1}{B_N} \int \phi\left(\tfrac{y - y_0}{B_N}\right) \pi_0(y \mid \lambda) dy \Big/ \pi_0(y_0 \mid \lambda) - 1 \right| = o(1),$$

where sequences C_N, C'_N, and B_N satisfy Assumption A.1.

We need the conditional density $\pi_0(y_0 \mid \lambda)$ to be smooth on the trimmed region, so that convolving it with our Gaussian kernel does not distort its shape substantially. This prevents spikes or point mass priors on $Y_{i0} \mid \lambda_i$, ensuring the leave-one-out smoothing step yields a valid approximation to the true prior.

The posterior mean function and the joint sampling distribution of the sufficient statistic and the initial condition take the form

$$m(\widehat{\lambda}, y_0; \pi_0) = \widehat{\lambda} + \Sigma_V(\theta_0) \frac{\partial}{\partial \widehat{\lambda}} \log p(\widehat{\lambda}, y_0; \pi_0),$$
$$p(\widehat{\lambda}, y_0; \pi_0) = \int \frac{1}{\sqrt{\det(\Sigma_V(\theta_0))}} \phi\left(\Sigma_V(\theta_0)^{-1/2}(\widehat{\lambda} - \lambda)\right) \pi_0(\lambda, y_0) d\lambda.$$

Also define the following $*$-counterparts by convolving the prior $\pi_0(\lambda, y_0)$ with a Gaussian kernel with bandwidth B_N. These $*$-objects are the population targets of the expected leave-one-out kernel estimator

$$m_*(\widehat{\lambda}, y_0; \pi_0, B_N)$$
$$= \widehat{\lambda} + \left(\Sigma_V(\theta_0) + B_N^2 I\right) \frac{\partial}{\partial \widehat{\lambda}} \log p_*(\widehat{\lambda}, y_0; \pi_0, B_N),$$
$$p_*(\widehat{\lambda}, y_0; \pi_0, B_N)$$
$$= \frac{1}{B_N^d} \int \frac{1}{\sqrt{\det(\Sigma_V(\theta_0) + B_N^2 I)}} \phi\left((\Sigma_V(\theta_0) + B_N^2 I)^{-1/2}(\widehat{\lambda} - \lambda)\right) \phi\left(\frac{y_0 - \tilde{y}_0}{B_N}\right) \pi_0(\lambda, \tilde{y}_0) d\lambda d\tilde{y}_0.$$

Assumption A.4 (Posterior mean functions) *Let C_N be a sequence satisfying Assumption A.1. The posterior mean functions satisfy:*

(a) $N \iint \left\| m(\widehat{\lambda}, y_0; \pi_0) \right\|^2 \mathbf{1}\left\{ \left\| m(\widehat{\lambda}, y_0; \pi_0) \right\| \geq C_N \right\} p(\widehat{\lambda}, y_0; \pi_0) d\widehat{\lambda} dy_0 = o_{u.\pi_0}(N^+),$

(b) $N \iint \left\| m_*(\widehat{\lambda}, y_0; \pi_0, B_N) \right\|^2 \mathbf{1}\left\{ \left\| m_*(\widehat{\lambda}, y_0; \pi_0, B_N) \right\| \geq C_N \right\} p(\widehat{\lambda}, y_0; \pi_0) d\widehat{\lambda} dy_0 = o_{u.\pi_0}(N^+),$

(c) $N \iint \left\| m(\widehat{\lambda}, y_0; \pi_0) \right\|^2 \mathbf{1}\left\{ \left\| m(\widehat{\lambda}, y_0; \pi_0) \right\| \geq C_N \right\} p_*(\widehat{\lambda}, y_0; \pi_0, B_N) d\widehat{\lambda} dy_0 = o_{u.\pi_0}(N^+).$

This assumption guarantees that outside a slowly growing ball of radius C_N, the contribution to the overall risk is negligible. In other words, only a vanishing fraction of units have such extreme estimates that they could undermine our uniform risk bound. We check this not

only for the posterior mean m and density p, but also for the variance inflated versions (m^*, p^*) that arise from adding the kernel variance B_N^2.

Assumption A.5 (Rates for $\widehat{\theta}$) *The estimator for the common parameters satisfies*

$$\mathbb{E}_{\theta_0,\pi_0}\left[\left|\sqrt{N}(\widehat{\rho}_Y - \rho_{Y,0})\right|^4\right] = o_{u.\pi_0}(N^+), \quad \mathbb{E}_{\theta_0,\pi_0}\left[\left|\sqrt{N}(\widehat{\sigma}_U^2 - \sigma_{U,0}^2)\right|^2\right] = o_{u.\pi_0}(N^+),$$

and similarly for $\rho_\delta, \sigma_\varepsilon^2$.

Finally, we require our estimator of the common parameters to converge at the usual $\sqrt{N}$-rate with sufficiently thin tails. This ensures that plugging $\widehat{\theta}$ into our empirical Bayes update does not introduce any first-order errors in the risk comparison against the oracle. By Theorem 3.1, our QMLE estimator attains the required $\sqrt{N}$-rate and thus fulfills this assumption.

Proof of Theorem 3.2. In the simple model under Assumption 2.3 (rank condition), the common treatment timing design $W_i(\rho_\delta)$ in (4) is deterministic and satisfies $W_i(\rho_\delta)'W_i(\rho_\delta)$ invertible with finite eigenvalues. Hence, the Moore-Penrose inverse $W_i^+(\rho_\delta) = (W_i(\rho_\delta)'W_i(\rho_\delta))^{-1}W_i(\rho_\delta)'$ exists and the sufficient statistic $\widehat{\lambda}_i(\rho) = W_i^+(\rho_\delta)\,(y_{i,1:T} - \widehat{\rho}_Y y_{i,0:T-1})$ in (5) is well defined. Following from (3), the covariance of the stacked innovations $\check{\Sigma}_U(\theta_0)$ is positive semidefinite. Then, the projection noise covariance $\Sigma_{V,i}(\theta_0) = W_i^+(\rho_\delta)\check{\Sigma}_U(\theta_0)\left[W_i^+(\rho_\delta)\right]'$ is well defined with finite eigenvalues.

Since $W_i(\rho_\delta)$ is deterministic and common across i in the simple model, we follow the proof strategy in Liu, Moon, and Schorfheide (2020), which instead focuses on individual forecasts. Under Assumptions A.1–A.5 governing trimming/bandwidth, CRC tails/smoothness, posterior mean functions, and $\sqrt{N}$-rates for the common parameters, we obtain the ratio optimality for the jointly estimated individual effects α_i and heterogeneous treatment effects δ_{i0}. ∎

Remark A.1 (Extension: rich controls C_i) Consider the extension in Section 4.1 with a conditional prior $\pi(\lambda_i \mid C_i)$, where $C_i = \left(Y_{i0}, Z_{i,1:T}^{0:J}, X_{i,0:T}^O, X_{i0}^P\right)$, $Z_{i,1:T}^{0:J}$ collects treatment timing and size (w.l.o.g. we consider continuous treatment here), $X_{i,0:T}^O$ are strictly exogenous covariate paths, and X_{i0}^P are initial values of predetermined covariates. Now $W_i(\rho_\delta) = W(\rho_\delta, C_i)$ and $\Sigma_{V,i}(\theta) = \Sigma_V(\theta, C_i)$ are functions of C_i.

Assume that $W(\rho_{\delta 0}, C_i)$ has full column rank with the eigenvalues uniformly bounded away from zero over trimmed C_i. Following from the continuity of $W(\rho_\delta, C_i)$ in ρ_δ uniformly

over trimmed C_i, there exists a compact neighborhood $\rho_{\delta 0} \in \Theta_\rho$ and a constant $0 < c < \infty$ such that

$$\inf_{\rho_\delta \in \Theta_\rho, \text{ trimmed } C_i} \lambda_{\min}\left(W(\rho_\delta, C_i)' W(\rho_\delta, C_i)\right) \geq c,$$

so $W^+(\rho_\delta, C_i)$ is well-defined uniformly over $\rho_\delta \in \Theta_\rho$ and trimmed C_i. Similarly, the covariance mapping $\Sigma_V(\theta, C_i)$ is smooth in θ uniformly over a compact neighborhood of θ_0 and trimmed C_i.

With this in place, replace Y_{i0} by C_i throughout Assumptions A.1–A.5. The Tweedie step and the ratio optimality argument then carry over verbatim, now conditional on C_i.