Department of Economics

Athens University of Economics and Business

WORKING PAPER no. 20-2025

On robust Bayesian causal inference

Angelos Alexopoulos and Nikolaos Demiris

November 2025

The Working Papers in this series circulate mainly for early presentation and discussion, as well as for the information of the Academic Community and all interested in our current research activity.

The authors assume full responsibility for the accuracy of their paper as well as for the opinions expressed therein.

On robust Bayesian causal inference

Angelos Alexopoulos[*]   Nikolaos Demiris[†]

Abstract

This paper develops a Bayesian framework for robust causal inference from longitudinal observational data. Many contemporary methods rely on structural assumptions, such as factor models, to adjust for unobserved confounding, but they can lead to biased causal estimands when mis-specified. We focus on directly estimating time–unit–specific causal effects and use generalised Bayesian inference to quantify model mis-specification and adjust for it, while retaining interpretable posterior inference. We select the learning rate ω based on a proper scoring rule that jointly evaluates point and interval accuracy of the causal estimand, thus providing a coherent, decision-theoretic foundation for tuning ω. Simulation studies and applications to real data demonstrate improved calibration, sharpness, and robustness in estimating causal effects.

Keywords: Causal Inference, Model mis-specification, Bayesian, Robust

1 Introduction

Problem setting This paper is concerned with causal inference that is robust to model mis-specification. We posit a Bayesian approach to the generic statistical problem of potentially correcting for using the wrong model and then use this method for causal inference. Our main focus lies upon robust causal inference from time-series observational data on multiple units. Such data are common in many scientific disciplines where randomized experiments are often infeasible. We suggest to (i) start with a model that is considered the most appropriate for tackling this problem, (ii) allow for this model to be wrong, (iii) aim to understand the level of mis-specification by focusing on the appropriate estimand and (iv) derive suitably corrected causal effects. Although model mis-specification is common in causal inference, inference in such settings represents a general statistical problem, appearing for example when approximations are derived due to computational reasons, like in Variational Bayes or Composite Likelihood methods.

Recent methods in causal inference adjust for unobserved heterogeneity and confounding via data-generating processes such as factor models, see for example Athey et al. [2021] and Xu [2017]. A proposal for positing models of this kind (up to transformations) as a general data-generating process may be found in Whiteley et al. [2025]. While such models appear like a reasonable baseline they remain subject to mis-specification due to issues like missing confounders, incorrect number of latent factors or unnecessary linearity assumptions. Such mis-specification can lead to biased or misleading causal estimates.

Methodological approach In this paper we set the estimand of interest to be the causal effect(s) and aim to understand the degree of model mis-specification with respect to the accuracy of estimating this causal effect. We focus on factor analysis-type models which are commonly used to evaluate real-world interventions while still adjusting for known confounders. We adopt the generalized Bayesian inference paradigm and replace the likelihood with a loss function while retaining Bayesian updating and access to the posterior of the (generally asymmetric, e.g. Claxton et al. [2015]) decision space. Estimation accuracy is assessed in terms of calibration and sharpness. Note that focused inference is not new in statistics, e.g. Claeskens and Hjort [2003], but is integral in generalised/robust Bayesian analysis, see for example Bissiri et al. [2016] and Fong et al. [2024]. We select the learning rate ω via a proper scoring rule that quantifies the joint accuracy of the posterior mean and credible interval of the causal estimand. The propriety of the scoring rule implies it favours posteriors whose summaries are correctly calibrated under the true data-generating process, providing a transparent and theoretically suitable criterion for tuning ω. This approach aligns the choice of ω directly with inferential accuracy rather than predictive or marginal likelihood objectives.

The proposed method has several advantages. It enables model-robust causal learning by adaptively adjusting influence of the two model components, the prior and the likelihood. It introduces (semi-)automatic bias and variance corrections based on the degree of model mis-specification. In particular, the posterior behavior adapts by either (i) shifting toward the prior when uncertainty is high or (ii) concentrating more closely to the data

---

[*]Department of Economics, Athens University of Economics and Business, Greece. Email: `angelos@aueb.gr`
[†]Department of Statistics, Athens University of Economics and Business, Greece. Email: `nikos@aueb.gr`

when appropriate. Although motivated by factor models, the method is broadly applicable to a range of causal models and other imperfect likelihood settings.

Experiments The method is studied in two simulation settings. First, in linear-type models where we find that our method naturally reduces to standard Bayesian inference when the model is well-specified and adapts appropriately as necessitated by the distinct types of model mis-specification. The second class of simulation experiments focus on Factor models where we also observe that selecting the optimal learning rate enables suitable location and scale corrections of the causal treatment effects. In summary, we demonstrate that the proposed method provides robustness and improvement over current alternatives.

We evaluate the proposed robust Bayesian causal inference (RBCI) framework using three real-world panel datasets. First, we revisit the California tobacco control program, a benchmark example in synthetic control evaluation [Abadie et al., 2010]. Second, we analyze the effect of a spatially targeted industrial policy on French regional employment [Gobillon and Magnac, 2016]. Third, we study a digital tax enforcement intervention in the Greek energy market introduced by the Independent Authority for Public Revenue (IAPR). Across all applications, we compare against state-of-the-art matrix completion methods and find that accounting for model mis-specification leads to more accurate causal estimates and substantially improved uncertainty quantification.

1.1 Connections to the relevant literature

A vast literature exists on accounting for model mis-specification in general inference settings. Early work includes Huber [1967] and also White [1982] who showed that the asymptotic density of the maximum likelihood estimator in mis-specified models is Gaussian with sandwich covariance matrix and centered on the Kullback–Leibler divergence minimising a pseudo-true parameter value. Muller [2013] studied the asymptotic behavior of the posterior in mis-specified parametric models and showed that generally it can have lower asymptotic frequentist risk.

A common way to account for model mis-specification in the causal inference literature is via the propensity score, see Rosenbaum and Rubin [1983]. This approach essentially generates a second option for obtaining consistency (in addition to using the correct outcome regression model), in the case where one specifies a correct propensity score model, thus leading to the 'doubly robust' terminology. A generic framework for causal inference via Inverse Probability weights has been developed by Robins and colleagues and is typically termed G-estimation; a detailed account may be found in Hernan and Robins [2020].

The propensity score approach has been criticised from the Bayesian viewpoint as an unnatural way to characterise a data generating mechanism [Li et al., 2023]. An excellent review of Bayesian approaches to causal inference may be found in Stephens et al. [2023] who explain the difficulty of casting the propensity score within Bayesian inference. The paper covers some (seemingly unnatural for Bayesians) ways developed in the literature for circumventing such issues, including cutting feedback and two-stage inference. The authors then develop a decision-theoretic approach to incorporate the propensity score within a Bayesian non-parametric (Dirichlet process-based) framework and use the Bayesian bootstrap. Antonelli et al. [2022] developed a Bayesian approach to doubly robust estimation with favourable frequentist properties by estimating a propensity score and an outcome model. Ray and Van de Vaart [2020] proposed to model the propensity score non-parametrically and showed that standard Gaussian process priors satisfy a semi-parametric Bernstein–von Mises theorem. Applications of Bayesian causal inference may be found in numerous diverse fields, including on algorithmic decisions (e.g. Jia et al. [2025] and Imai et al. [2023]).

A parallel strand of fruitful, mostly non-Bayesian, recent developments may be found in the intersection of (bio)statistics, machine learning and econometrics, see for example Chernozhukov et al. [2018], Lewis and Syrgkanis [2021] and also van der Laan and Rose [2018] where a detailed account of the targeted minimum loss estimation may be found. Recent studies include Dorn et al. [2025], Ghosh and Rothenhäusler [2025] who present robust confidence intervals for the Average Treatment Effect (ATE), Jin and Syrgkanis [2025] who show optimality of the celebrated and widely used doubly robust estimators for both the ATE and the Average Treatment Effect on the Treated (ATT) and Bruns-Smith et al. [2025] who demonstrate that a type of doubly robust estimators termed augmented balancing weights may achieve covariate balance directly instead of inverting the propensity score, see also Resa and Zubizarreta [2020].

Our approach broadly follows a distinct line of research originating in Bayesian methods that aim for inference which is robust (semi-parametric-type), Bayesian and does not follow the usual Dirichlet process or Gaussian process-based Bayesian methods. Early such examples may be found in Seaman and Richardson [2004] and Rice [2008] while Bissiri et al. [2016] develop a general approach that we present in detail in Section 2. This method allows for model mis-specification and its degree is estimated at Lyddon et al. [2019] by matching the covariance matrix to its asymptotic limit while Syring and Martin [2019] estimate it by calibrating the coverage of the credible intervals, see also Holmes and Walker [2017]. The predictive performance of such methods is investigated in McLatchie et al. [2025]. Our method may be seen as an alternative to these methods since we are concerned with a non-asymptotic viewpoint and we aim to understand the degree of model mis-specification with respect to a focused task, estimating the causal estimand of interest.

Outline of the paper The remainder of the paper is organised as follows. In section 2 we describe the methodological part of this paper and section 3 the causal learning specifics. Section 4 contains the simulation

experiments demonstrating the robustness and accuracy of the proposed method. Section 5 consists of three real world applications and the paper concludes with a discussion at section 6.

2 Robust Bayesian causal inference

This section introduces our RBCI framework and the proposed criterion for selecting the learning rate ω. The method combines generalized Bayesian inference, which allows for potential model mis-specification, with a proper scoring rule designed to evaluate the posterior accuracy of the causal estimand. The score targets the two aspects most relevant to applied causal inference, the bias of the posterior mean and calibration of credible intervals, and provides a decision-theoretic foundation for choosing ω that ensures well-calibrated and sharp posterior inference.

2.1 Generalised Bayesian inference

The generalised Bayesian inference framework provides a versatile extension to classical Bayesian updating in cases where the assumed model (and likelihood) is potentially mis-specified. In the standard Bayesian paradigm, posterior inference relies on a correctly specified likelihood $p(y_{1:n} \mid \theta)$ to update prior beliefs about the parameter θ after observing data $y_{1:n}$. When this likelihood is incorrect, a situation that is ubiquitous in causal inference with complex observational data, the resulting posterior distribution may be biased or misleading. Generalized Bayesian inference replaces the likelihood with a suitably chosen loss function $L(\theta, y_{1:n})$ that measures the discrepancy between model predictions and observed data [Bissiri et al., 2016]. The resulting posterior distribution, often termed Gibbs posterior, is defined as

$$\pi_\omega(\theta \mid y_{1:n}) \propto \pi(\theta) \exp\{-\omega L(\theta, y_{1:n})\}, \tag{1}$$

where $\pi(\theta)$ denotes the prior distribution and $\omega > 0$ is a learning rate parameter that controls the weight of the loss relative to the prior.

This formulation retains the Bayesian machinery for uncertainty quantification and decision-making but relaxes the need for a fully specified generative model. In particular, the choice of L allows the analyst to target specific aspects of the model, such as the fit of a specific estimand, while ω acts as a tuning (hyper)parameter to mitigate the impact of model mis-specification. In well-specified models, $\omega = 1$ recovers the standard Bayesian posterior while under mis-specification alternative values of ω can yield better calibrated or sharper posteriors [Syring and Martin, 2019]. In the next subsection we describe how this framework may be used for causal inference.

2.2 A Robust Bayesian approach to causal inference

Causal inference from observational data involves estimating causal estimands, such as average treatment effects or unit-specific intervention effects, in the presence of confounding and other complexities. Existing Bayesian approaches typically proceed by specifying a data-generating model and performing posterior inference under the assumption that the model is correct, see for example [Saarela et al., 2016]. However, deviations from the assumed model, e.g., unmeasured confounders, or omitted dynamics can induce bias in the resulting causal estimates.

We aim to account for such mis-specification by (i) start with a model that may reasonably be considered the state of the art, (ii) use generalised Bayesian inference to allow for this model to be incorrect, (iii) assess the degree of mis-specification by focusing on the causal estimand τ itself and (iv) derive suitably corrected causal effects. In essence, rather than attempting to fully recover the data-generating mechanism and then estimate τ as a by-product, we take a somewhat less ambitious approach and aim to directly calibrate the posterior distribution of τ, using a loss function tailored to causal estimation accuracy. This shift of focus, from the correct model to robust estimand learning, is a central characteristic of the generalized Bayesian approach and we suggest it is suitable for causal inference.

2.2.1 Selecting the loss function

Since our aim is to understand the degree of model mis-specification, we select a model-based type of loss function that encompasses parametric models and can incorporate covariates. A natural approach is to adopt the negative log-likelihood of the selected *baseline model* as the loss function, scaled by a learning rate ω, leading to a *power posterior* (or tempered likelihood). Hence, our primary formulation corresponds to

$$L(\theta, y_{1:n}) = -\log p(y_{1:n} \mid \theta),$$

yielding the Gibbs posterior in (1). This choice preserves interpretability since it reduces to standard Bayesian inference when $\omega = 1$ and allows for robust adaptation; when $\omega < 1$, the posterior distribution inflates uncertainty to account for potential model mis-specification while $\omega > 1$ leads to sharpening around the data. The choice of

the likelihood function is problem-specific, effectively defined by the state of the art model for the problem at hand. The term ω is essentially a hyper-parameter that may not naturally be estimated by the data and the next section is concerned with its selection.

Interestingly, alternative loss functions that target bias reduction or variance control more directly, e.g., doubly-robust or orthogonal losses, could also be used in principle. However, these generally break conjugacy and lead to substantially higher computational cost. For this reason, we retain the likelihood-based loss and rely on ω to provide a computationally efficient form of robustness.

2.2.2 Estimating the optimal learning rate

Our inference procedure defines a grid of learning-rate values $\{\omega_1, \dots, \omega_B\}$ and selects the value that optimizes a decision-theoretic criterion based on the accuracy of estimating the causal effect τ. For each ω, let $\Pi_\omega(\tau \mid y_{1:n})$ denote the marginal posterior (or posterior predictive) distribution of τ induced by the generalized posterior in (1). Denote its expectation and central (i.e. equal-tailed) $(1-\alpha)$ quantiles by

$$\widehat{\tau}_\omega = \mathbb{E}_{\Pi_\omega}[\tau], \qquad (\ell_\omega, u_\omega) = \left(F_{\Pi_\omega}^{-1}(\alpha/2),\ F_{\Pi_\omega}^{-1}(1-\alpha/2)\right),$$

respectively, where F_{Π_ω} is the cumulative distribution function (CDF) of $\Pi_\omega(\tau \mid y_{1:n})$. We assess the quality of each posterior through a score that combines point and interval accuracy:

$$S(F_{\Pi_\omega}, \tau) = (\widehat{\tau}_\omega - \tau)^2 + S_{\text{int}}(\ell_\omega, u_\omega; \tau), \tag{2}$$

where S_{int} is the interval score of Gneiting and Raftery [2007],

$$S_{\text{int}}(\ell, u; \tau) = (u - \ell) + \frac{2}{\alpha}(\tau - u)\mathbf{1}\{\tau > u\} + \frac{2}{\alpha}(\ell - \tau)\mathbf{1}\{\tau < \ell\}.$$

The first term in (2) penalizes bias of the posterior mean, while the second measures the width and calibration of the central $(1-\alpha)$ credible interval. One may interpret $S(F_{\Pi_\omega}, \tau)$ as the specialization, at $G = F_{\Pi_\omega}$, of a general scoring rule $S(G, \tau)$ defined on distributions G for τ; see in the supplementary material for details, including a proof of the following Proposition.

Proposition 1. *Let G_0 denote the true distribution of τ. For any distribution G, the expected score satisfies*

$$\mathbb{E}_{\tau \sim G_0}\big[S(G, \tau)\big] \ \geq\ \mathbb{E}_{\tau \sim G_0}\big[S(G_0, \tau)\big],$$

with equality whenever G and G_0 share the same mean and central $(1-\alpha)$ quantiles.

Hence, under the true data-generating process, the score is minimised in expectation when the posterior reports the correct mean and central $(1-\alpha)$ interval of τ. The rule is not strictly proper for the full predictive distribution since distinct posteriors with the same mean and central interval may achieve the same expected score. This suffices in our setting, since our focus lies on accurate point estimation and well-calibrated intervals for causal effects rather than full distributional elicitation. Therefore, working with a proper rather than a strictly proper scoring rule does not compromise the validity of our evaluation criterion. If one wishes to enforce strict propriety for the full predictive distribution, $S(F_{\Pi_\omega}, \tau)$ could be replaced or augmented by a strictly proper score such as the continuous ranked probability score (CRPS); see in the supplementary material for a discussion.

Proposition 2. *Let $R_n(\omega) = \mathbb{E}[S(F_{\Pi_\omega}, \tau)]$ denote the expected score under the true data-generating process. In the correctly specified Gaussian model $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, with estimand of interest $\tau = \mu$, $R_n(\omega)$ is uniquely minimized at $\omega = 1$. The result holds exactly for a flat (improper) prior on τ when σ^2 is known and asymptotically for any regular conjugate prior on τ and σ^2 when σ^2 is unknown.*

Intuitively, under correct model specification, the generalized posterior reduces to the standard Bayesian posterior, and the expected score $R_n(\omega) = \mathbb{E}[S(F_{\Pi_\omega}, \tau)]$ attains its unique minimum at $\omega = 1$. This can be verified analytically in the normal case, where the posterior mean is the sample mean and the posterior variance scales as $\sigma^2/(\omega n)$, implying that the expected score is minimized when the posterior uncertainty matches the true sampling variance. A detailed proof is provided in the supplementary material. The optimal learning rate is then chosen as

$$\omega^\star = \arg\min_{\omega \in \{\omega_1, \dots, \omega_B\}} S(F_{\Pi_\omega}, \tau),$$

balancing sharpness and calibration of the posterior distribution of τ against the accuracy of its mean. For comparison, one may consider the mean-squared-error criterion

$$S_{\text{MSE}}(\omega, \tau) = \left(\mathbb{E}_{\Pi_\omega}[\tau] - \tau\right)^2 + \text{Var}_{\Pi_\omega}(\tau), \tag{3}$$

which decomposes uncertainty into squared bias and variance. In simple Gaussian cases, the interval width and posterior variance are linked via

$$u_\omega - \ell_\omega = 2z_{1-\alpha/2}\sqrt{\text{Var}_{\Pi_\omega}(\tau)} = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{\omega n}},$$

for some σ, but in general $S_{\text{int}}(\ell_\omega, u_\omega; \tau)$ and $\text{Var}_{\Pi_\omega}(\tau)$ capture different aspects of calibration and sharpness.

The optimization of ω is carried out by evaluating the score (2) (or, alternatively, (3)) over a grid of candidate values and selecting the minimizer. The procedure is summarized below. Importantly, since the score in (2) is proper, this procedure provides a theoretically grounded and interpretable way of tuning ω to achieve robust, well-calibrated inference for causal effects.

Algorithm 1 Selecting the optimal learning rate $\omega^\star$

1: Specify a grid of candidate values $\{\omega_1, \ldots, \omega_B\}$.
2: **for** $b = 1, \ldots, B$ **do**
3:      Compute the generalized posterior $\Pi_{\omega_b}(\tau \mid y_{1:n})$.
4:      Extract $\widehat{\tau}_{\omega_b}$, ℓ_b, and u_{ω_b}.
5:      Evaluate the score $S(F_{\Pi_{\omega_b}}, \tau)$ from (2).
6: **end for**
7: Select $\omega^\star = \arg\min_{\omega_b} S(F_{\Pi_{\omega_b}}, \tau)$.

---

3 Causal Inference

The goal of causal inference is to estimate the effect of a treatment or intervention on an outcome of interest while accounting for confounding and other sources of bias. The standard formalisation is the *potential outcomes framework* [Rubin, 1974]. For each unit i, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively. The individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$ and because we only ever observe one of these potential outcomes, causal inference is fundamentally a missing data problem. A key estimand of interest is the ATE, $\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)]$, which measures the expected difference in outcomes if the entire population were treated versus the scenario where none were treated. In many observational and policy evaluation contexts, a more relevant estimand is the ATT, $\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid D = 1]$ which measures the average causal impact of the treatment on those units that actually received it.

3.1 Cross-sectional data: regression-based approaches

When data consist of independent observations taken at a single time point (cross-sectional data), causal effects are often estimated via outcome regression models or propensity score methods. A common approach is to specify a linear regression model:

$$Y_i = \alpha + \tau D_i + X_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

where D_i is a binary treatment indicator, X_i are observed covariates, and τ captures the average effect of treatment. Under the *conditional unconfoundedness* assumption

$$\{Y_i(1), Y_i(0)\} \perp D_i \mid X_i,$$

the ordinary least squares estimator of τ identifies the ATE. Bayesian approaches proceed by specifying priors over $(\alpha, \beta, \tau, \sigma^2)$ and computing the posterior distribution [Saarela et al., 2016]. Extensions include flexible outcome regression, propensity score weighting, and doubly robust estimators. The strength of these approaches lies in their simplicity and interpretability. However, they rely critically on observing all relevant confounders and correctly specifying the outcome model. When these conditions fail — for instance, if unobserved heterogeneity influences both treatment and outcomes — the estimated causal effects will likely be biased.

3.2 Panel data: latent factor and synthetic control models

In many applied settings, interventions are introduced at different times across units, a situation known as *staggered adoption*. Extending the potential outcomes framework [Rubin, 1974] to longitudinal data, we define for each unit i and time t the potential outcomes $Y_{it}(1)$ and $Y_{it}(0)$ under treatment and control, respectively. The causal effect of interest at the unit–time level is

$$\tau_{it} = Y_{it}(1) - Y_{it}(0),$$

but as in the cross-sectional setting, we only observe one of the two potential outcomes at any given time, making causal inference a missing data problem. Specifically, let T_i be the time treatment begins for unit i; for $t \geq T_i$ we observe $Y_{it}(1)$ and must infer the missing counterfactual $Y_{it}(0)$. The ATT and its dynamic form can then be written as

$$\tau_{\text{ATT}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \tau_{\text{ATT}}(t) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \left[Y_{it}(1) - Y_{it}(0) \right] \right),$$
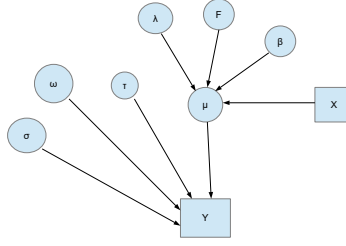
Figure 1: Directed acyclic graph representing the structure of the generalized Bayesian causal model. The treatment effect τ and noise scale σ govern the outcome Y, while latent factors F with loadings λ generate the baseline structure. Covariates X influence both the latent structure and outcome through regression coefficients β. The learning rate ω modulates the likelihood contribution in the generalized posterior, and μ denotes prior location hyperparameters.

where $\mathcal{I}_t = \{i : T_i \leq t\}$ is the set of units treated by time t and $\mathcal{T}$ denotes the set of post-treatment periods. These dynamic effects capture how the impact of treatment evolves with exposure length. A widely used modelling strategy for staggered adoption settings, following Xu [2017], specifies the observed outcome as

$$Y_{it} = \tau_{it} D_{it} + \lambda_i^\top f_t + X_{it}^\top \beta + \varepsilon_{it}, \tag{5}$$

where D_{it} is a binary treatment indicator, λ_i are unit-specific factor loadings, f_t represent time-varying latent factors common across units, X_{it} are observed covariates with coefficient vector β, and ε_{it} is an idiosyncratic error. This specification allows for both unobserved heterogeneity and common shocks affecting units through the interactive fixed effects $\lambda_i^\top f_t$. A graphical representation of the models defined in 5 (and 4) is depicted in Figure 1 where we group the mean behaviour in μ and split the uncertainty/structural parameters affecting the response data Y.

Under the assumption that treatment affects outcomes only through $\tau_{it} D_{it}$, the untreated potential outcomes follow

$$Y_{it}(0) = \lambda_i^\top f_t + X_{it}^\top \beta + \varepsilon_{it}, \tag{6}$$

which corresponds to the data-generating process for control units ($D_{it} = 0$). Estimation of λ_i and f_t from pre-treatment data enables counterfactual prediction of $Y_{it}(0)$ in post-treatment periods, forming the basis of generalized synthetic control and related matrix completion approaches [Xu, 2017, Athey et al., 2021].

Conditional on estimates $\widehat{Y}_{it}(0)$ of untreated potential outcomes, we can write the overall and time-specific ATT in a single numbered display

$$\widehat{\tau}_{\text{ATT}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \widehat{\tau}_{\text{ATT}}(t) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \left[Y_{it} - \widehat{Y}_{it}(0) \right] \right). \tag{7}$$

Despite their flexibility and empirical success, latent factor and matrix completion approaches rely on structural assumptions that may be violated in practice, leading to biased $\widehat{Y}_{it}(0)$ and therefore distorted τ_{ATT} and $\widehat{\tau}_{\text{ATT}}(t)$ estimates.

3.3 Baseline Likelihood and Estimation Methodology

Cross-sectional and panel data are widely analysed using frequentist, ML and Bayesian approaches. In cross-sectional settings, causal effects are often estimated under conditional unconfoundedness via regression or propensity score methods [Hogan and Lancaster, 2004, Seaman and Vansteelandt, 2018, Abadie et al., 2024]. Flexible ML approaches [Athey and Wager, 2019, Van der Laan et al., 2011] relax functional assumptions but typically offer weaker uncertainty quantification.

Bayesian methods provide coherent uncertainty assessment and are well-suited to complex data structures, including latent confounding. However, robustness to *model mis-specification* has received less attention in this literature, with some notable exceptions such as BART [Hill et al., 2020] and Bayesian doubly robust estimators [Saarela et al., 2016]. This motivates a generalised Bayesian approach designed to remain reliable even when the outcome model is imperfect.

3.3.1 Bayesian estimation in cross-sectional studies

In the Gaussian linear model (4), the generalized Bayesian posterior from (1) is defined in terms of the loss function $L(\theta, Y)$, where $\theta = (\alpha, \beta, \tau, \sigma^2)$. We take $L(\theta, Y)$ to be the negative log-likelihood, which (up to an

additive constant) is proportional to the usual squared-error loss:

$$L(\theta, Y) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - X_i^\top \beta\right)^2.$$

Substituting this choice of L into (1) gives

$$\pi_\omega(\theta \mid Y) \propto \pi(\theta) \exp\left\{-\frac{\omega}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - X_i^\top \beta\right)^2\right\}.$$

Equivalently (up to a factor independent of θ when σ^2 is fixed),

$$\pi_\omega(\theta, \sigma^2 \mid Y) \propto \pi(\theta) \left[\mathcal{N}\left(Y \mid \alpha\mathbf{1} + \tau D + X\beta, \ \sigma^2 I_n\right)\right]^\omega$$

$$\propto \ \pi(\theta) \mathcal{N}\left(Y \mid \alpha\mathbf{1} + \tau D + X\beta, \ (\sigma^2/\omega) I_n\right).$$

Clearly, the learning rate ω controls the relative weight of the data and the prior in the posterior distribution and directly influences the effective noise level in the model. Hence, in this simple case it holds that raising the Gaussian likelihood to the power ω is equivalent (up to a proportionality constant) to using a Gaussian likelihood with variance σ^2/ω. This means that ω rescales the variance term and thus the precision with which parameters are estimated. When $\omega = 1$, we recover standard Bayesian updating under the true model. Values $\omega < 1$ downweight the likelihood relative to the prior, effectively inflating the variance to $\sigma^2/\omega > \sigma^2$ leading to broader credible intervals, similarly to the sandwich estimators. This type of conservative inference is desirable when the model is mis-specified, including when the likelihood is overly informative relative to the true data-generating process. Conversely, $\omega > 1$ shrinks the effective variance producing sharper posteriors. Overall, ω acts as a tuning parameter that balances robustness and efficiency: smaller values favour robustness to mis-specification by tempering the likelihood's influence, while larger values favour efficiency by emphasising the information in the data.

3.3.2 Panel data estimation

Estimation of causal effects in panel data settings with staggered treatment adoption typically relies on predicting the unobserved counterfactual outcomes $Y_{it}(0)$ using latent factor models such as (6). The most common frequentist approach estimates the latent factors and loadings by least squares or principal components methods [Bai, 2009], after which counterfactuals $\widehat{Y}_{it}(0)$ are constructed and plugged into (3.2) to obtain estimates of τ_{ATT} and $\tau_{\text{ATT}}(t)$. Several approaches build on this framework. Specifically, synthetic control methods [Abadie et al., 2010] construct a weighted combination of untreated units to approximate each treated unit's counterfactual trajectory, while generalized synthetic controls [Xu, 2017] combine factor structure with observed covariates. Matrix completion approaches [Athey et al., 2021] use low-rank structure to impute missing potential outcomes directly. Machine learning techniques such as nuclear norm minimization, regularized factor models, and neural-network-based imputations have also been proposed to improve predictive accuracy. These approaches are widely used and often perform well empirically but remain subject to model mis-specification which can bias $\widehat{Y}_{it}(0)$ and lead to misleading treatment effect estimates. Also, frequentist and ML approaches can provide limited uncertainty quantification thus complicating inference on τ_{ATT} and $\tau_{\text{ATT}}(t)$.

Similarly to the case of cross-sectional studies we adopt the generalised Bayesian approach to estimate the latent factor model in (6). Let Y the vectorised form of the $\mathbb{R}^{N \times T}$ matrix that collects outcomes with entries Y_{it} and $\theta = (f_{1:T}, \lambda_{1:n}, \beta, \sigma^2)$ collect all unknown parameters. The Gaussian likelihood implied by this model reads

$$p(Y \mid \theta) = \prod_{i,t} \mathcal{N}\left(Y_{it} \mid \lambda_i^\top f_t + X_{it}^\top \beta, \ \sigma^2\right).$$

In the generalized Bayesian framework of (1), we replace the likelihood with the loss

$$L(\theta, Y) = \frac{1}{2\sigma^2} \sum_{i,t} \left(Y_{it} - \lambda_i^\top f_t - X_{it}^\top \beta\right)^2,$$

leading to the generalized posterior

$$\pi_\omega(\theta \mid Y) \propto \pi(\theta) \left[\mathcal{N}\left(Y \mid \Lambda F + X\beta, \ \sigma^2 I\right)\right]^\omega,$$

where ΛF denotes the matrix of factor components with $(\Lambda F)_{it} = \lambda_i^\top f_t$. This formulation highlights the role of the scaling parameter ω and has the same interpretation as in the cross-sectional setting for $\omega < 1, = 1, > 1$ respectively.

Our approach relates to Athey et al. [2021] who select the number of latent factors by minimizing a penalised loss function; a procedure that balances model fit and complexity. Our framework can be viewed as a Bayesian generalisation of the same principle: rather than fixing the factor structure ex ante and penalising the loss function once, we adaptively calibrate ω to control the effective influence of the likelihood throughout the posterior learning process. In doing so, we address structural uncertainty (such as factor dimensionality) and provide a coherent mechanism for uncertainty quantification, also when the factor model is an approximation of the true data-generating process. Given posterior draws of θ, posterior predictive samples $\widehat{Y}_{it}(0)$ are generated and used to compute $\widehat{\tau}_{\mathrm{ATT}}$ and $\widehat{\tau}_{\mathrm{ATT}}(t)$ along with their posterior distributions. This approach therefore combines coherent uncertainty quantification with robustness to model mis-specification through the calibrated choice of ω.

4 Simulation study

We now illustrate the performance of the proposed methodology through simulation experiments designed to mimic typical causal inference settings. The goal is to investigate how the proposed approach behaves under different levels of model mis-specification and to assess the effect of the learning rate ω in the two scenarios of cross-sectional regression and panel-data factor models. The details on prior specification and Bayesian computation are provided in the supplementary material.

4.1 Simulation I: Cross-sectional regression

Our first simulation investigates the behaviour of the generalized posterior in a regression model under varying degrees of unobserved confounding. Following Luo et al. [2023], we generate three covariates $X_1, X_2, X_3 \overset{iid}{\sim} \mathcal{N}(0,1)$ and construct a latent confounder $U = |X_1|/\sqrt{1 - 2/\pi}$. Treatment assignment is then generated according to a logistic model

$$D \mid U, X_2, X_3 \sim \mathrm{Bernoulli}(\mathrm{expit}(0.4U + 0.4X_2 + 0.8X_3)),$$

so that U and the observed covariates jointly influence the probability of receiving treatment. Finally, outcomes are drawn from the model

$$Y \mid D, U, X_1 \sim \mathcal{N}(\tau D + X_1 + \gamma U, 1),$$

where the parameter γ controls the degree of model mis-specification with larger values of γ inducing stronger unobserved confounding and therefore greater mis-specification of the working model. We simulate $N = 500$ observations from the model and assess inference regarding the causal estimand of interest, the treatment effect τ. We estimate τ by targeting the generalized posterior in (1) with the loss function set to the baseline log-likelihood of the working model

$$y_i = \beta_D d_i + \beta x_{i1} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), i = 1, \ldots, N,$$

where $\tau = \beta_D$ and the latent confounder U is omitted on purpose to introduce mis-specification. We generate data under several γ values to vary the severity of this mis-specification.

We compute the posterior density of τ under different values of ω and evaluate the score function in (2). Figure 2 illustrates two key findings. First, in the left panel it is clear that the minimiser of $S(F_{\Pi_\omega}, \tau)$ shifts systematically as the degree of model mis-specification increases. When the omitted confounding is weak (small γ), the score is minimised near $\omega = 1$, which corresponds to standard Bayesian updating and indicates that the working model is approximately correct. As the confounding strength grows, however, the optimal ω becomes smaller, downweighting the likelihood and effectively inflating posterior uncertainty. This behaviour illustrates how our framework adapts to model mis-specification: rather than over-confidently relying on a mis-specified likelihood, it automatically reduces the influence of the data and produces more conservative inference. Second, the right panel shows that this adaptive scaling has tangible benefits for inference on the treatment effect. The posterior distribution of β_D becomes wider under mis-specification and also shifts towards the true value, correcting for the bias that would otherwise arise from ignoring the latent confounder, as appropriate. This combination of variance inflation and location adjustment leads to credible intervals with better coverage properties and more reliable causal conclusions. Together, these results highlight the key advantage of the proposed approach: by calibrating ω to the causal effect, the generalized posterior shifts towards the data-generating process, thus achieving robustness to model mis-specification without requiring explicit model corrections or complex structural assumptions.

4.2 Simulation II: Panel data and latent factor models

Our second simulation examines a staggered–adoption panel-data setting with latent confounding. We generate a $T \times N$ panel with $N = 30$ units observed over $T = 100$ time periods. For each unit i we draw a treatment start time T_i from a discrete uniform distribution on $\{40, \ldots, 95\}$, yielding heterogeneous post–intervention windows. The observed outcome is generated according to

$$y_{it} = \tau_{it} D_{it} + \lambda_i^\top f_t + \beta_U U_{it}(1 - D_{it}) + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0,1),$$
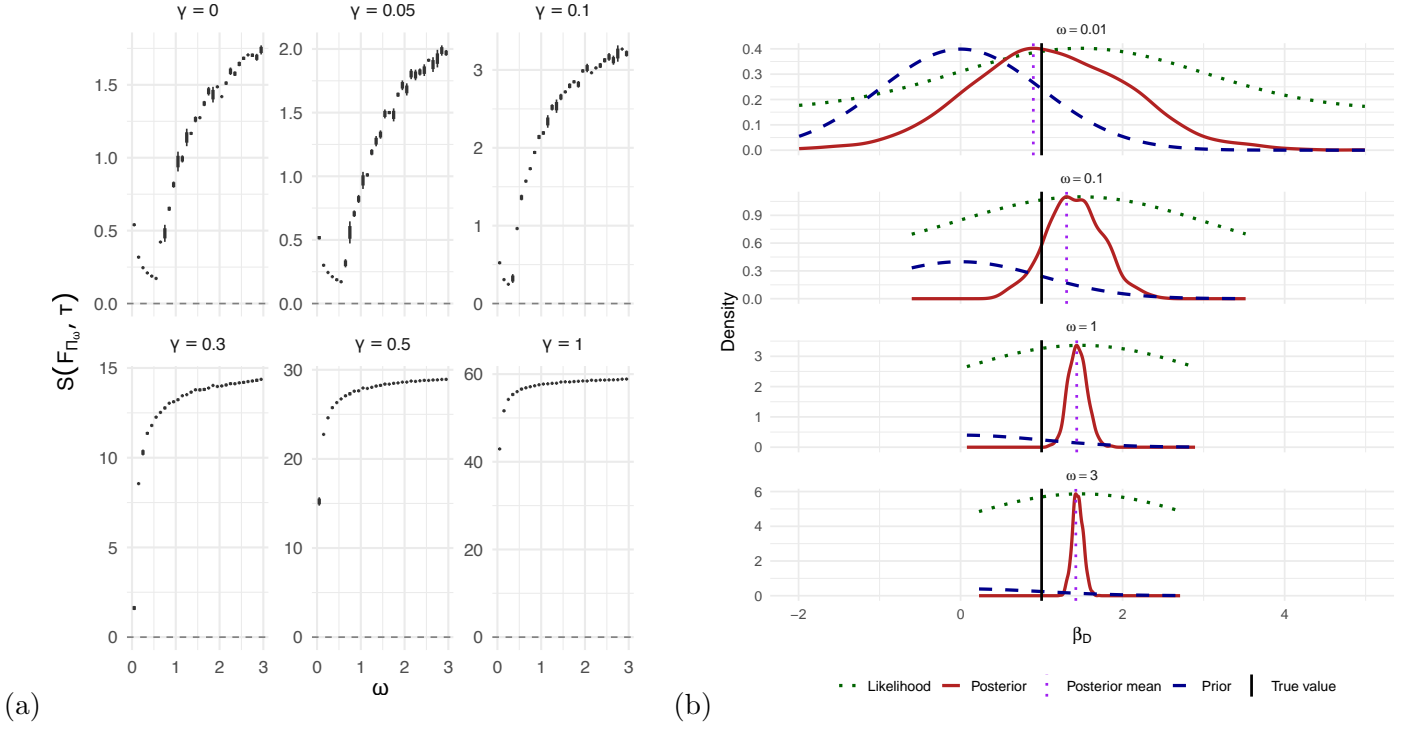
Figure 2: Cross-sectional regression under model mis-specification. Panel (a): score $S(F_{\Pi_\omega}, \tau)$ for different values of ω and γ; the boxplots correspond to 10 runs of the MCMC algorithm in Step 3 of Algorithm 1 to account for Monte Carlo variability. Panel (b): the prior/likelihood/posterior triplot of β_D for $\gamma = 0.3$.

where τ_{it} denotes the heterogeneous treatment effect, $D_{it} = \mathbb{I}\{t \geq T_i\}$ is the staggered treatment indicator, and $U_{it} = |X_{it}|/\sqrt{1 - 2/\pi}$ represents a latent source of time-varying outcome heterogeneity with $X_{it} \sim \mathcal{N}(0, 1)$. The parameter β_U controls the degree of latent confounding and induces increasing model mis-specification as it grows.

We simulate a low-rank latent structure using $K = 2$ latent factors with $\lambda_i \sim \mathcal{N}_K(0, I_K)$ and $f_t \sim \mathcal{N}_K(0, I_K)$ corresponding to the Gaussian priors specified in the supplementary material. These factors drive strong cross-sectional dependence and capture shared temporal dynamics unrelated to treatment. For treated units the counterfactual trajectories $y_{it}(0)$ are unobserved whenever $t \geq T_i$. Following the latent factor literature [Athey et al., 2021, Samartsidis et al., 2024], the working model for the untreated potential outcomes is specified as

$$y_{it}(0) = \lambda_i^\top f_t + \epsilon_{it}.$$

Figure 3 illustrates how the average posterior score $S(\omega, K)$ varies over a grid of learning rate and factor dimension choices, under two different levels of model misspecification. The left panel corresponds to the correctly specified scenario ($\beta_U = 0$), while the right panel examines strong unobserved confounding ($\beta_U = 1$). Two notable patterns emerge: when the model is correctly specified, the score is minimised near $\omega = 1$ and at the true number of latent factors, confirming that standard Bayesian updating remains optimal in well-specified settings. By contrast, as misspecification increases, the minimiser shifts favouring $\omega < 1$ and alternate values of K indicating that the generalized posterior adapts by downweighting the likelihood and adjusting model complexity. Overall, this joint surface analysis provides guidance on how the pair (ω, K) should be tuned to balance variance, bias, and complexity in different misspecification regimes.

Having established the behaviour of the score across tuning parameters, we now compare the resulting generalized posterior treatment effect estimates with those from the state-of-the-art matrix completion estimator of Athey et al. [2021]. We select (K, ω) using a predictive cross-validation score that targets unbiased counterfactual recovery by randomly masking 20% of pre-treatment observations for never-treated units for which τ_{it} should be identically equal to zero. We vary both the degree of model mis-specification (via β_U) and the number of latent factors K, and, for each scenario, evaluate the minimiser of $S(F_{\Pi_\omega}, \tau)$ in the joint two-dimensional space. This setup allows us to assess how the proposed approach compensates for both (i) structural model errors and (ii) factor number mis-specification when estimating causal effects; see in the supplementary material for the detailed description of the adopted cross-validation procedure.

Figure 4 summarizes predictive causal performance under increasing levels of latent factor mis-specification. We report both the unit–time–specific squared bias of the estimated average treatment effect and the 95% interval score, comparing the proposed RBCI method against matrix completion. When the factor structure is properly specified ("zero" mis-specification), both approaches perform similarly. As mis-specification increases, however, RBCI systematically achieves lower squared bias together with substantially smaller interval scores, yielding more
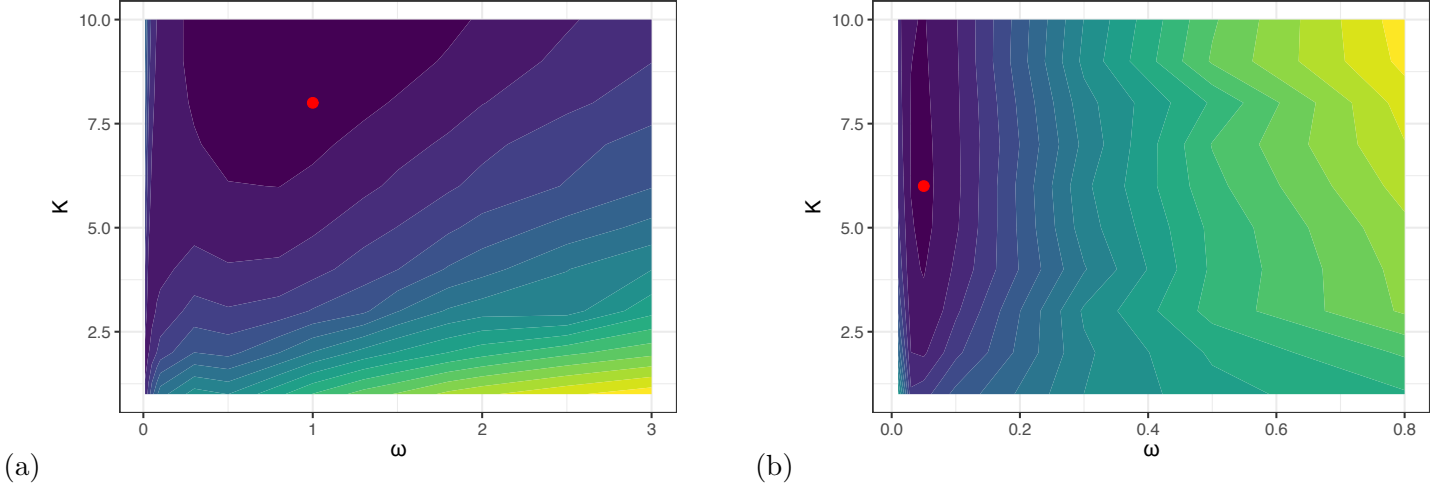
9

Figure 3: Average posterior score surface $S(F_{\Pi_\omega}, \tau)$ across a grid of values for the learning rate ω and the number of latent factors K. Panel (a) displays the score surface for different ω and γ under correct model specification ($\beta_U = 0$), panel (b) corresponds to strong mis-specification ($\beta_U = 1$). The location of the minimum indicates the optimal balance between likelihood weighting and model complexity.
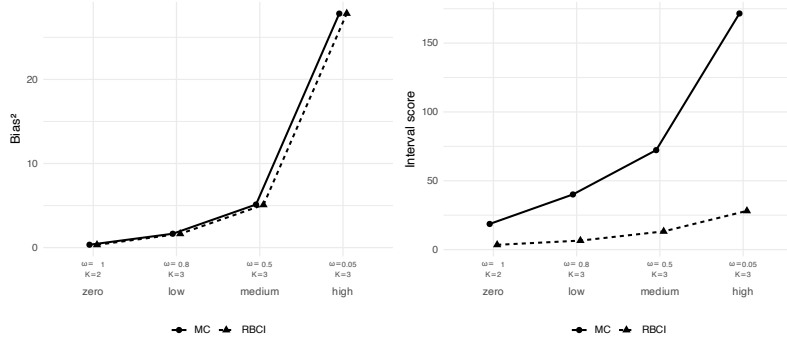

Figure 4: Comparison of squared bias (left) and interval score (right) for the proposed RBCI method and matrix completion (MC) across increasing latent factor mis-specification. Under each scenario we report the selected learning rate $\widehat{\omega}$ and number of latent factors $K^\star$.

accurate point estimates and sharper yet well-calibrated uncertainty. The figure also displays the selected learning rate $\widehat{\omega}$ and number of factors $K^\star$ chosen under each condition. Specifically, $\widehat{\omega} < 1$ is chosen once unobserved confounding becomes more severe, tempering overconfident likelihood contributions and improving predictive robustness. These results demonstrate that adaptively tuning $(\widehat{\omega}, K^\star)$ leads to more reliable estimation of τ_{ATT} without sacrificing performance when the model is correctly specified.

5   Real data analysis

We evaluate the performance of the proposed RBCI framework on three real-world panel datasets. In all applications we fit the latent factor model introduced in (5), with the unit–time–specific treatment effect τ_{it} as the estimand of interest, and we estimate the generalized posterior using the MCMC methodology described in Section 3.3.2. The learning rate ω is selected using the same cross-validation design as in Section 4: we randomly mask never-treated outcomes and choose ω by evaluating how accurately the model recovers the known values of τ_{it}. This is possible because τ_{it} is identically zero for all control units and for treated units prior to their intervention time, providing observed benchmarks against which the score $S(F_{\Pi_\omega}, \tau_{it})$ can be computed.

The first application revisits the well-known California smoking intervention, a benchmark setting for synthetic control and matrix completion methods [Abadie et al., 2010, Athey et al., 2021]. The second examines regional employment outcomes in France following a spatially targeted industrial policy [Gobillon and Magnac, 2016]. The third evaluates a digital tax enforcement intervention introduced by the Greek IAPR in the energy market. Across all three studies, we compare RBCI to state-of-the-art matrix completion approaches [Xu, 2017, Athey et al., 2021], focusing on treatment-effect estimation and uncertainty quantification under latent factor mis-specification.
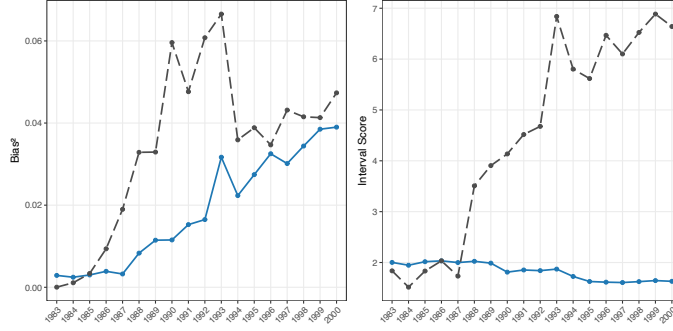
Figure 5: Per-Year predictive accuracy on masked pseudo-treated states in the California smoking study. Squared bias (left) and interval score (right), reported relative to the Proposition 99 intervention year (1989), for the proposed robust Bayesian causal inference method (solid) and matrix completion (dashed).

5.1 The California smoking intervention

We begin with the California tobacco control program analyzed in Abadie et al. [2010]. The dataset contains annual cigarette consumption for 39 U.S. states observed over 31 years (1970–2000), with California serving as the single treated state from 1989. Following Athey et al. [2021], we treat the remaining $N = 38$ states as potential controls. We assess both causal point estimation and uncertainty quantification relative to matrix completion as implemented in `gsynth` [Xu et al., 2021].

Hyperparameters in our RBCI framework are selected by minimizing a localized predictive scoring rule that combines unit–time–specific squared bias and interval score. To do so, we randomly mask 20% of the outcomes belonging to never-treated states and evaluate predictions on these held-out cells, targeting unbiased out-of-sample validation. The selected configuration ($K^\star = 2$, $\widehat{\omega} = 0.5$) reflects both the extent of latent factor structure and the degree of structural mis-specification: $\widehat{\omega} < 1$ indicates that a standard Bayesian update would have been overly confident and likelihood tempering yields better-calibrated uncertainty. To evaluate causal prediction performance, we follow a placebo strategy by masking the post–1989 outcomes of a random subset (15%) of states and treating them as pseudo-treated. Both our method and matrix completion are fit to the unmasked panel and then evaluated strictly on these masked placebo outcomes, treating them as ground truth. Full details are provided in the supplementary material.

Figure 5 summarizes post–intervention predictive accuracy. The left panel shows *time–unit–specific* squared bias computed averaged across pseudo-treated units at each post–treatment year. Moreover, the right panel reports the interval score for 95% predictive intervals, jointly quantifying sharpness and calibration. Our approach achieves systematically lower squared bias in most post–intervention years, together with smaller interval scores, implying more accurate and more efficient uncertainty quantification than matrix completion. These advantages reflect upon a tempered likelihood which adapts to latent factor mis-specification and produces improved time–unit–level causal estimates.

5.2 The French regional industrial policy

We next examine the regional employment intervention studied in Gobillon and Magnac [2016], where industrial subsidies were allocated to selected French labour market areas. The dataset contains annual employment measures for $N = 148$ areas observed over $T = 20$ years (1997-2016), among which 13 received treatment while the remaining 135 serve as never-treated controls.

We evaluate predictive performance using the same experimental pipeline as in Section 5.1. Table 1 reports results for three employment measures: firm entries, firm exits, and total employment. Our approach produces substantially improved uncertainty quantification, interval scores are reduced by 35–65% in all cases, while achieving similar point accuracy. The benefits are most pronounced for the "entries" outcome, where matrix completion yields overly dispersed uncertainty intervals despite lower squared bias. Notably, for all outcomes the selected learning rate satisfies $\widehat{\omega} \geq 1$, in contrast to the California study, suggesting that the latent factor structure in this French employment panel is comparatively well specified and requires little likelihood tempering for accurate predictive uncertainty.

11

Table 1: Predictive performance on masked placebo-treated regions in the French employment study. Average unit–time–specific squared bias (Bias^2) and 95% interval score (IS) over post-intervention times for the matrix completion (MC) and the proposed robust Bayesian causal inference (RBCI) framework.

Outcome	Method	Bias^2	IS	$K^\star$	$\widehat{\omega}$
Entries	RBCI	0.0204	0.7251	4	1.5
	MC	0.0063	2.0753	4	1.5
Exits	RBCI	0.00078	0.4672	1	2.0
	MC	0.00055	0.5259	1	2.0
Employment	RBCI	0.00319	0.4932	1	2.0
	MC	0.00457	1.3319	1	2.0

5.3 Evaluation of digital intervention in the Greek energy market

Our third dataset is concerned with a digital tax enforcement programme by the Greek IAPR aimed at improving compliance in the energy sector. The intervention rolled out across firms and regions in phases, yielding a panel with staggered treatment adoption for $T = 53$ weekly observations from $1,399$ gas stations in Greece during 2020. The outcome of interest is the log-transformed weekly fuel sales of each station. The policy is operationalised through a QR-code authentication system on fuel receipts, enabling consumers to verify in real time whether transactions have been transmitted to the tax authority's central database, thereby increasing the perceived probability of detection for non-compliant stations and altering incentives to under-report sales. Because we observe the district of each gas station, we analyse treatment effects both at the national level and separately within 13 administrative districts. Additional data details and descriptive statistics are provided in the supplementary material.
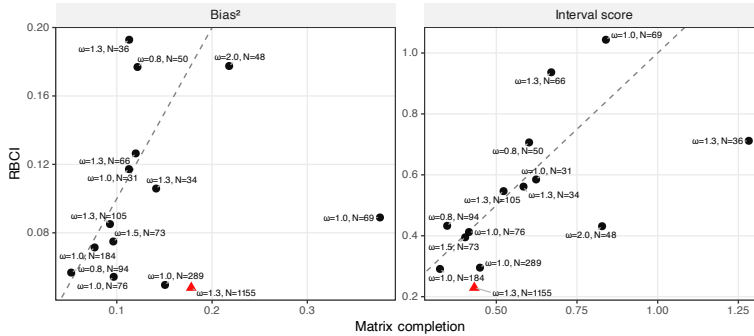

Figure 6: Each point summarises average time–unit performance for a district (Bias^2 on the left; interval score on the right), with labels indicating the selected learning rate $\widehat{\omega}$ and the number of observations N contributing to that point. The 45° reference line indicates equal performance in point accuracy and uncertainty quantification between the matrix completion method (x-axis) and the proposed robust Bayesian causal inference (RBCI) framework. The analysis of the complete dataset is depicted using a red triangle.

Figure 6 compares district–level and overall performance of matrix completion and RBCI. The analysis of all (N=1155) the data (depicted using a red triangle) used $\widehat{\omega} = 1.3$ and resulted in improved bias and IS performance of RBCI indicating desirable point accuracy as well as sharper *and* well–calibrated predictive intervals. The district-level results suggest that RBCI achieves smaller or comparable bias (with two exceptions) and typically lower interval scores. Selected learning rates vary across districts, with $\widehat{\omega}$ often in the 0.7–2.0 range, reflecting heterogeneous degrees of latent–factor mis–specification: values $\widehat{\omega} < 1$ temper the likelihood when the factor structure is too rigid (yielding better calibration), whereas $\widehat{\omega} \geq 1$ rates indicate that down–weighting is unnecessary and the data can be exploited more aggressively. Overall, the Greek energy application mirrors our earlier findings: likelihood tempering adapts to structural uncertainty and delivers robust uncertainty quantification across heterogeneous panels, while remaining competitive in counterfactual point prediction.

6 Discussion

This paper postulates that causal inference based on observational data should assume the possibility for residual model error that one ought to quantify and adjust for. The robust Bayesian method outlined is broadly applicable to a range of causal models. We suggest that one should initiate this process using a model that may reasonably

be considered the state of the art for the problem at hand and then allow for model mis-specification. We then adopt a non-asymptotic Bayesian approach that facilitates model correction and access to the uncertainty of the decision space. We are mostly concerned with causal inference where the issues are often structural due to missing confounders but the method is applicable to other such problems, like the presence of outliers. In addition to causal inference, the proposed framework may be used in other imperfect likelihood settings, like variational Bayes or composite likelihood cases.

We target causal effects at the unit–time level, estimating counterfactual outcomes per treated unit and post–treatment period. We adopt a tempered posterior and select the learning rate ω to balance calibration and sharpness. When the working model is adequate ($\omega \approx 1$) RBCI coincides with standard Bayesian inference while under mis-specification it adjusts the location and scale of the posterior for the unit–time–specific effects as appropriate. Our simulations and applications show that selecting ω with a proper scoring rule—combining point accuracy and interval calibration—provides improved coverage and reduced or comparable bias compared to state-of-the-art estimators.

Default choices and future work A number of choices made in the present work require further discussion and exploration. The main ones are concerned with (i) the choice of loss function. This was dictated by the presence of suitable baseline models and the requirement to adjust for covariates but alternative, perhaps causality or estimand-focused, loss functions are likely suitable in different problems, including doubly-robust-like or orthogonal estimating-equation-type losses. (ii) The calibration and sharpness measures; the assessment of estimation is non-trivial in the generalised Bayes case beyond Gaussian cases and alternative accuracy measures may be operationalised by the problem, such as the nature of the estimand of interest. (iii) We found that both $\omega < 1$ and $\omega > 1$ may offer optimal solutions. The former offers an alternative to the tradition Bayesian prior-likelihood-posterior triplot with the posterior shifting towards the prior while the interpretation of the latter suggests 'super-learning' with the posterior concentrating closer to the data and further theoretical investigation is warranted. (iv) Depending on the data, the joint posterior score space of the number of factors and ω can appear relatively flat. This space deserves further investigation and the necessary tools may be based on the Geometry of the Fisher Information and the sandwich covariance matrix. (v) We focused on tempered likelihoods using a single ω but additional flexibility may be gained using multiple, likely related, ω and insights from the weighted/composite likelihood literature may offer fruitful suggestions. (vi) Our inference procedure largely ignored the design of the intervention process. Its efficiency could further improve by adopting appropriate such designs, see for example Imai et al. [2013] and Chattopadhyay et al. [2026]. (vii) Our proposed framework is general and can be extended to broader classes of models, including generalised linear mixed models and survival models; this is the subject of current work. Finally (viii), recent developments (e.g. Fong et al. [2024]) cast Bayesian inference as a predictive task and drop the prior via a stronger focus on the estimand of interest. However, it is not immediately apparent how one may include in this framework factor-like and related general models and this remains an open problem.

7 Acknowledgments and code

References

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490): 493–505, 2010.

A. Abadie, A. Agarwal, R. Dwivedi, and A. Shah. Doubly robust inference in causal latent factor models. *arXiv preprint arXiv:2402.11652*, 2024.

J. Antonelli, G. Papadogeorgou, and F. Dominici. Causal inference in high dimensions: a marriage between bayesian modeling and good frequentist properties. *Biometrics*, 78(1):100–114, 2022.

S. Athey and S. Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

D. Bruns-Smith, O. Dukes, A. Feller, and E. L. Ogburn. Augmented balancing weights as linear regression (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear, 2025. URL https://doi.org/10.1093/jrsssb/qkaf019.

A. Chattopadhyay, C. N. Morris, , and J. R. Zubizarreta. Balanced and robust randomized treatment assignments: The finite selection model for the health insurance experiment and beyond (with discussion). *Journal of the Royal Statistical Society Series A: Statistics in Society*, to appear, 2026.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):1–68, 2018.

G. Claeskens and N. Hjort. The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916, 2003.

K. Claxton, M. Sculpher, S. Palmer, and A. J. Culyer. Causes for concern: Is nice failing to uphold its responsibilities to all nhs patients? *Health Economics*, 24(1):1–7, 2015. URL https://doi.org/10.1002/hec.3130.

J. Dorn, K. Guo, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120(549):331–342, 2025.

E. Fong, C. C. Holmes, and S. G. Walker. Martingale posterior distributions (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(5):1357–1391, 2024.

A. Ghosh and D. Rothenhäusler. Assumption-robust causal inference. *Arxiv*, 2025. URL https://arxiv.org/abs/2505.08729.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

L. Gobillon and T. Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551, 2016.

M. A. Hernan and J. M. Robins. Causal inference: What if. 2020.

J. Hill, A. Linero, and J. Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, 2020.

J. W. Hogan and T. Lancaster. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical methods in medical research*, 13(1):17–48, 2004.

C. C. Holmes and S. G. Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

P. Huber. The behavior of the maximum likelihood estimates under nonstandard conditions. *in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.*, 5th:221—-233, 1967.

K. Imai, D. Tingley, and T. Yamamoto. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176(1):5—51, 2013. URL https://doi.org/10.1111/j.1467-985X.2012.01032.x.

K. Imai, Z. Jiang, D. J. Greiner, R. Halen, and S. Shin. Experimental evaluation of algorithm-assisted human decision-making: application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167—189, 2023. URL https://doi.org/10.1093/jrsssa/qnad010.

Z. Jia, E. Ben-Michael, and K. Imai. Bayesian safe policy learning with chance constrained optimization: application to military security assessment during the vietnam war. *Journal of the Royal Statistical Society Series A: Statistics in Society*, to appear, 2025. URL https://doi.org/10.1093/jrsssa/qnaf122.

J. Jin and V. Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *COLT*, 2025. URL https://arxiv.org/abs/2402.14264.

G. Lewis and V. Syrgkanis. Double/debiased machine learning for dynamic treatment effects. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22695–22707. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bf65417dcecc7f2b0006e1f5793b7143-Paper.pdf.

F. Li, P. Ding, and F. Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153, 2023.

Y. Luo, D. J. Graham, and E. J. McCoy. Semiparametric bayesian doubly robust causal estimation. *Journal of Statistical Planning and Inference*, 225:171–187, 2023.

S. P. Lyddon, C. Holmes, and S. Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

Y. McLatchie, E. Fong, D. T. Frazier, and J. Knoblauch. Predictive performance of power posteriors. *Biometrika*, 112(3), 2025. URL `https://doi.org/10.1093/biomet/asaf034`.

U. Muller. Risk of bayesian inference in misspecified models and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

K. Ray and A. Van de Vaart. Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

M. d. l. A. Resa and J. R. Zubizarreta. Direct and stable weight adjustment in non-experimental studies with multivalued treatments: Analysis of the effect of an earthquake on post-traumatic stress. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(4):1387—1410, 2020. URL `https://doi.org/10.1111/rssa.12561`.

K. Rice. Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies. *Journal of the American Statistical Association*, 103(481):385–396, 2008. URL `https://doi.org/10.1198/016214507000001463`.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

O. Saarela, L. R. Belzile, and D. A. Stephens. A Bayesian view of doubly robust causal inference. *Biometrika*, 103(3):667–681, 2016.

P. Samartsidis, S. R. Seaman, A. Harrison, A. Alexopoulos, G. J. Hughes, C. Rawlinson, C. Anderson, A. Charlett, I. Oliver, and D. De Angelis. A Bayesian multivariate factor analysis model for causal inference using time-series observational data on mixed outcomes. *Biostatistics*, 25(3):867–884, 2024.

S. R. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies. *Biometrika*, 91(1):15–25, 2004. URL `https://doi.org/10.1093/biomet/91.1.15`.

S. R. Seaman and S. Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

D. A. Stephens, W. S. Nobre, E. E. M. Moodie, and A. M. Schmidt. Causal inference under mis-specification: Adjustment based on the propensity score (with discussion). *Bayesian Analysis*, 18(2):639–694, 2023.

N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

M. J. van der Laan and S. Rose. Targeted learning in data science: Causal inference for complex longitudinal studies. 2018.

M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

N. Whiteley, A. Gray, and P. Rubin-Delanchy. Statistical exploration of the manifold hypothesis (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear, 2025.

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

Y. Xu, L. Liu, and M. Y. Xu. *gsynth: Generalized Synthetic Control Method for Causal Inference with Panel Data*, 2021. URL `https://cran.r-project.org/web/packages/gsynth/gsynth.pdf`. R package version 1.2.4.

Appendix/Supplementary

Methodology

Properties of the Scoring rule $S(\tau, \omega)$

Proof of Proposition 1. Let G_0 denote the true distribution of τ. Write $m(G) = \mathbb{E}_G[\tau]$ and let $\ell(G) = G^{-1}(\alpha/2)$ and $u(G) = G^{-1}(1 - \alpha/2)$ denote the central quantiles.

The squared-error term $(m(G) - \tau)^2$ is a strictly proper scoring rule for the mean: its expected value under G_0 is minimized if and only if $m(G) = m(G_0)$. The interval score $S_{\mathrm{int}}(\ell(G), u(G); \tau)$ is a strictly proper scoring rule for the pair of quantiles $(\ell(G), u(G))$ [Gneiting and Raftery, 2007, Theorem 3], so its expected value under G_0 is minimized if and only if $(\ell(G), u(G)) = (\ell(G_0), u(G_0))$.

For any G,
$$\mathbb{E}_{\tau \sim G_0}[S(G, \tau)] = \mathbb{E}_{\tau \sim G_0}\big[(m(G) - \tau)^2\big] + \mathbb{E}_{\tau \sim G_0}\big[S_{\mathrm{int}}(\ell(G), u(G); \tau)\big].$$

Each term is minimized at its true value, so
$$\mathbb{E}_{\tau \sim G_0}[S(G, \tau)] \geq \mathbb{E}_{\tau \sim G_0}[S(G_0, \tau)],$$

with equality if and only if $m(G) = m(G_0)$ and $(\ell(G), u(G)) = (\ell(G_0), u(G_0))$. Thus $S(G, \tau)$ is proper. It is not strictly proper for the full distribution, since distinct G can share the same mean and central quantiles. □

The proposition implies that, in expectation, the score $S(F, \tau)$ is minimized when the reported distribution has the correct posterior mean and central $(1 - \alpha)$ credible interval of τ. Consequently, tuning ω by minimizing $S(F_{\Pi_\omega}, \tau) = S(\Pi_\omega, \tau)$ constitutes a proper decision-theoretic rule targeted at the inferential features of interest.

To enforce strict propriety for the full predictive distribution, a natural alternative is the continuous ranked probability score (CRPS),
$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \big(F(z) - \mathbf{1}\{y \leq z\}\big)^2 \, dz = \mathbb{E}_F |X - y| - \tfrac{1}{2}\mathbb{E}_F |X - X'|,$$

which is strictly proper for all univariate predictive distributions [Gneiting and Raftery, 2007]. The CRPS can be viewed as the integral of interval scores across all nominal levels α and hence enforces strict propriety.

Proof of Proposition 2. Let $S(F_{\Pi_\omega}, \tau) = (\widehat{\tau}_\omega - \tau)^2 + S_{\mathrm{int}}(\ell_\omega, u_\omega; \tau)$ and define the expected score
$$R_n(\omega) := \mathbb{E}\big[S(F_{\Pi_\omega}, \tau)\big],$$

where the expectation is with respect to the sampling distribution under the correctly specified model.

Let $\pi(\tau) \propto 1$ and Gaussian likelihood, the generalized posterior for τ is $\mathcal{N}(\bar{Y}, \sigma^2/(\omega n))$, so $\widehat{\tau}_\omega = \bar{Y}$ and the equal-tailed $(1 - \alpha)$ interval is $[\ell_\omega, u_\omega] = [\bar{Y} \pm a_\omega]$ with $a_\omega = z_{1-\alpha/2}\, \sigma/\sqrt{\omega n}$. Hence,
$$\mathbb{E}\big[(\widehat{\tau}_\omega - \tau)^2\big] = \mathbb{E}[(\bar{Y} - \tau)^2] = \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

Therefore,
$$R_n(\omega) = \frac{\sigma^2}{n} + f_n(\omega),$$

where
$$f_n(\omega) := \mathbb{E}[S_{\mathrm{int}}(\ell_\omega, u_\omega; \tau)].$$

Let $Z = \sqrt{n}(\bar{Y} - \tau)/\sigma \sim \mathcal{N}(0, 1)$ and $t(\omega) = a_\omega/(\sigma/\sqrt{n}) = z_{1-\alpha/2}/\sqrt{\omega}$. A standard computation yields
$$f_n(\omega) = \frac{2\sigma}{\sqrt{n}}\left[t(\omega) + \frac{2}{\alpha}\Big(\phi(t(\omega)) - t(\omega)\,\bar{\Phi}(t(\omega))\Big)\right],$$

where ϕ and $\bar{\Phi}$ denote the standard normal pdf and upper tail. Differentiating w.r.t. ω (using $t'(\omega) = -\tfrac{1}{2}t(\omega)\,\omega^{-1}$, $\phi'(t) = -t\phi(t)$, $\bar{\Phi}'(t) = -\phi(t)$) gives
$$f_n'(\omega) = -\frac{\sigma}{\sqrt{n}}\frac{z_{1-\alpha/2}}{\omega^{3/2}}\left[1 - \frac{2}{\alpha}\bar{\Phi}\Big(\frac{z_{1-\alpha/2}}{\sqrt{\omega}}\Big)\right].$$

Since $\bar{\Phi}(z_{1-\alpha/2}) = \alpha/2$, we have $f_n'(1) = 0$. A second-derivative check shows $f_n''(1) > 0$, hence $R_n(\omega)$ is uniquely minimized at $\omega = 1$.

Conjugate priors and unknown variance. The same conclusion holds for any regular conjugate prior on τ or when σ^2 is unknown and assigned a normal–inverse-gamma prior. In both settings the generalized posterior for τ is approximately normal with mean $m_\omega = w_\omega \bar{Y} + (1 - w_\omega)m_0$ and variance proportional to $\sigma^2/(\omega n)$ (conditionally on σ^2 in the latter case). As $n \to \infty$, standard conjugate-updating results imply $w_\omega \to 1$, $m_\omega \xrightarrow{P} \bar{Y}$, and the posterior scale parameter converges in probability to σ^2. Hence the posterior mean and variance coincide asymptotically with those under the flat-prior, known-variance case, and $R_n(\omega)$ converges to the same limit $\sigma^2/n + f(\omega)$ whose unique minimizer is $\omega = 1$. $\qquad\square$

Choosing K and ω in panel data

To tune the hyperparameters (K, ω) of the Bayesian factor model and assess robustness to model mis-specification, we employ a placebo-based validation design that operates at the time–unit level as illustrated by Figure 7. We randomly draw 15% of units from this pool and assign them artificial treatment at randomly sampled post intervention times. All post-placebo outcomes for these units are removed from the data matrix and treated as missing during model fitting. For each hyperparameter combination, the masked pseudo-treated outcomes are reconstructed solely from pre-treatment information via posterior prediction by minimising the score $(\hat{\mu}_{it} - Y_{it})^2 + \mathrm{IS}_{it}$. Here, $\hat{\mu}_{it}$ denotes the posterior predictive mean and thus $(\hat{\mu}_{it} - Y_{it})$ is its unit- time-specific bias, while IS_{it} is the interval score based on full predictive intervals. The hyperparameters $(K^\star, \omega^\star)$ are selected by minimizing the average of the score across (which is an estimation of $S(F_{\Pi_\omega}, \tau)$) all masked placebo cells.

After selection, a final model is refit using the full observed pre-placebo data, and predictive performance is calculated exclusively on the held-out placebo outcomes. This procedure avoids any information leakage from masked outcomes into factor learning, jointly accounts for pointwise causal bias and uncertainty calibration, and preserves the heterogeneous, time–unit–specific structure of causal effects. Comparisons against matrix completion methods (e.g., matrix completion and generalized synthetic controls [Xu, 2017, Athey et al., 2021]) are conducted using the exact same pseudo-treated time–unit pairs to guarantee a fair benchmark.
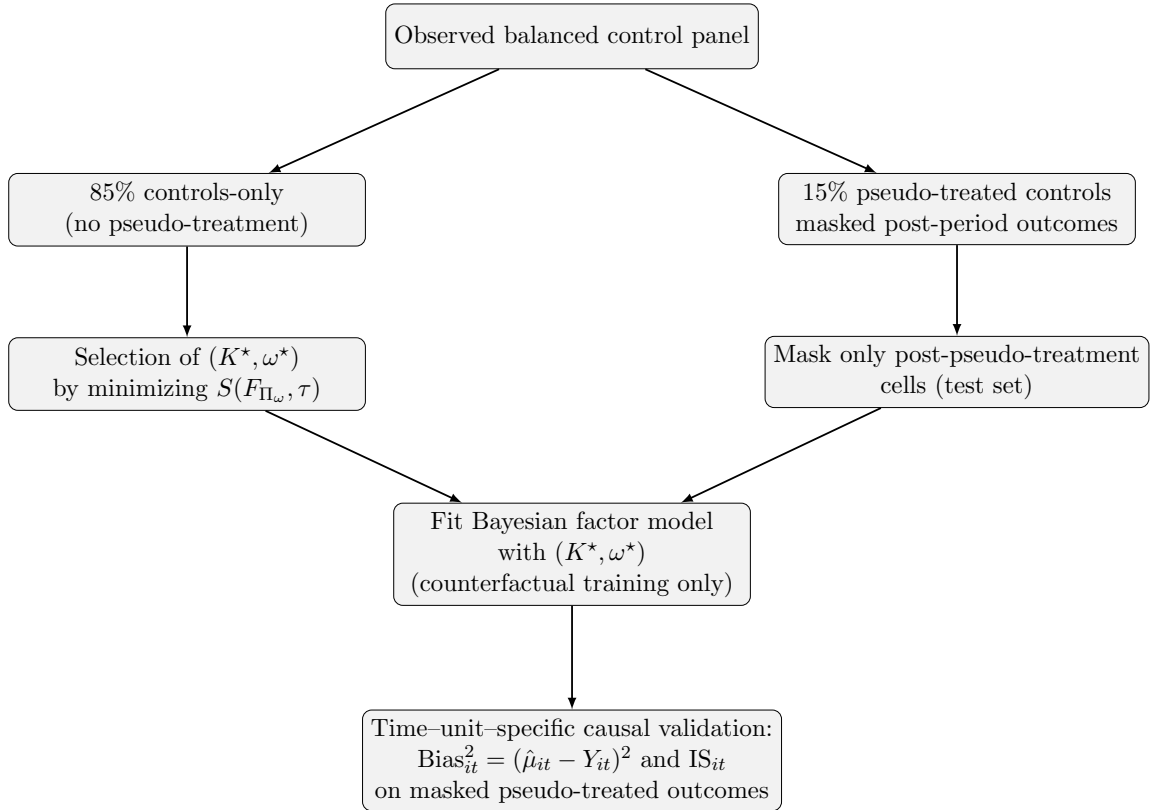

Figure 7: Time–unit–specific placebo-based validation and causal evaluation. We designate 15% of controls as pseudo-treated and mask their post-period outcomes. Hyperparameters (K, ω) are selected to minimize local prediction error on masked cells. The final model is evaluated by time–unit–specific squared bias and interval score.

MCMC and modelling details

Cross-sectional regression

Following the main text, inference targets the generalized posterior for the treatment effect τ under the *working* Gaussian regression that omits the latent confounder:

$$Y_i = \beta_D D_i + \beta X_{i1} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, N,$$

with $\tau \equiv \beta_D$. (In the experiments we let σ^2 be unknown for generality; if desired it may be fixed to 1 to match the data generating variance.)

Priors. We place conjugate priors
$$(\beta_D, \beta)^\top \sim \mathcal{N}(m_0, V_0),$$
$$\sigma^{-2} \sim \mathrm{Gamma}(a_0, b_0),$$

with hyperparameters (m_0, V_0, a_0, b_0) chosen to be weakly informative.

Generalized (tempered) posterior. For a learning rate $\omega > 0$, the generalized posterior is

$$\pi_\omega(\theta \mid \mathcal{D}) \ \propto \ \left\{ \prod_{i=1}^N \mathcal{N}\big(Y_i; \, \beta_D D_i + \beta X_{i1}, \ \sigma^2\big) \right\}^\omega \times \pi(\theta), \qquad \theta = (\beta_D, \beta, \sigma^2).$$

Raising the likelihood to the power ω is equivalent to inflating the effective noise from σ^2 to σ^2/ω in the Gaussian kernel, thereby down-weighting the data when $\omega < 1$.

Gibbs sampler. Let $X = [D \ \ X_1]$ be the $N \times 2$ design and $\beta = (\beta_D, \beta)^\top$. Conditional posteriors are standard:

1. $\beta \mid \sigma^2, Y \sim \mathcal{N}(m_\omega, V_\omega)$ with

$$V_\omega \ = \ \left(V_0^{-1} + \omega \, \sigma^{-2} X^\top X \right)^{-1}, \qquad m_\omega \ = \ V_\omega \left(V_0^{-1} m_0 + \omega \, \sigma^{-2} X^\top Y \right).$$

2. $\sigma^{-2} \mid \beta, Y \sim \mathrm{Gamma}(a_\omega, b_\omega)$ with

$$a_\omega \ = \ a_0 + \tfrac{\omega N}{2}, \qquad b_\omega \ = \ b_0 + \tfrac{\omega}{2} \, \|Y - X\beta\|_2^2.$$

The sampler iterates these two updates for M iterations, discarding burn-in and keeping M_p posterior draws $\{\beta^{(m)}, \sigma^{2(m)}\}_{m=1}^{M_p}$. The causal parameter of interest is $\tau \equiv \beta_D$, for which we retain $\{\tau^{(m)}\}_{m=1}^{M_p}$.

From the retained draws we report:

$$\widehat{\tau}_\omega = \mathsf{E}_m[\tau^{(m)}], \qquad \big(\tau_\omega^{\mathrm{lo}}, \ \tau_\omega^{\mathrm{hi}}\big) = \big(\mathsf{q}_{0.025}(\tau^{(m)}), \ \mathsf{q}_{0.975}(\tau^{(m)})\big).$$

For each candidate ω and each simulated dataset (indexed by γ), we compute the score

$$\mathrm{Bias}_\tau^2(\omega) \ = \ \big(\widehat{\tau}_\omega - \tau_{\mathrm{true}}\big)^2, \qquad \mathrm{IS}_\tau(\omega) \ = \ \big(\tau_\omega^{\mathrm{hi}} - \tau_\omega^{\mathrm{lo}}\big) + \tfrac{2}{\alpha}\big(\tau_\omega^{\mathrm{lo}} - \tau_{\mathrm{true}}\big) \mathbf{1}\{\tau_{\mathrm{true}} < \tau_\omega^{\mathrm{lo}}\} + \tfrac{2}{\alpha}\big(\tau_{\mathrm{true}} - \tau_\omega^{\mathrm{hi}}\big) \mathbf{1}\{\tau_{\mathrm{true}} > \tau_\omega^{\mathrm{hi}}\},$$

with $\alpha = 0.05$. Our overall scalar score is

$$S(\omega; \gamma) \ = \ \mathrm{Bias}_\tau^2(\omega) \ + \ \mathrm{IS}_\tau(\omega).$$

We select $\omega^\star(\gamma)$ that minimizes $S(\omega; \gamma)$ on a grid of values. To account for Monte Carlo variability in Step 3 of Algorithm 1, we repeat the sampler R times for each (γ, ω) and report boxplots of $S(\omega; \gamma)$ across replicates (the main text uses $R = 10$).

When the working model is approximately correct ($\gamma \approx 0$), the minimizer $\omega^\star$ is typically near 1, recovering standard Bayesian updating. As mis-specification increases ($\gamma > 0$), the optimal $\omega^\star$ drifts below 1, down-weighting the likelihood and widening the posterior for τ. Empirically this reduces bias and improves the interval score, yielding more reliable uncertainty quantification for the causal effect.

Factor model for panel data

For evaluation we split never-treated donor units into a *train* subset (for hyperparameter calibration) and a *test* subset (held out for pseudo treatment). For each test unit i, we assign a staggered placebo start T_i and mask its post-placebo outcomes

$$\mathcal{M} \;=\; \{(i,t): \; t \geq T_i, \; i \in \text{test}\}, \qquad Y_{it} = \text{NA}, \; (i,t) \in \mathcal{M}.$$

We also define a binary placebo indicator $Z_{it} = \mathbb{I}\{t \geq T_i, \; i \in \text{test}\}$. When fitting, unit i contributes only its pre-placebo observations $\{t < T_i\}$ if it is in test. Following the main text, the untreated potential outcome admits the interactive fixed-effects representation

$$Y_{it}(0) \;=\; \lambda_i^\top f_t \;+\; X_{it}^\top \beta \;+\; \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_i^2),$$

where $f_t \in \mathbb{R}^K$ are common across units factors and $\lambda_i \in \mathbb{R}^K$ are loadings. We place conjugate priors

$$f_t \sim \mathcal{N}(0, I_K), \quad \lambda_i \sim \mathcal{N}(0, \lambda I_K), \quad \sigma_i^{-2} \sim \text{Gamma}(a, b), \quad \beta \sim \mathcal{N}(0, \Sigma_\beta),$$

and target the generalized (tempered) posterior with learning rate $\omega > 0$

$$\pi_\omega(\theta \mid Y_{\text{obs}}) \;\propto\; \left\{ \prod_{(i,t) \in \mathcal{O}} \mathcal{N}\big(Y_{it} \,; \lambda_i^\top f_t + X_{it}^\top \beta, \; \sigma_i^2\big) \right\}^\omega \;\times\; \pi(\theta),$$

where $\mathcal{O}$ are observed (unmasked) cells and $\theta = (\{f_i\}, \{\lambda_t\}, \beta, \{\sigma_i^2\})$. Raising the likelihood to the power ω is equivalent to inflating the effective noise variance from σ_i^2 to σ_i^2/ω in the Gaussian terms, thus down-weighting the likelihood when $\omega < 1$.

At iteration $m = 1, \ldots, M$

1. Update $\{\lambda_i\}$ conditionally on $\{f_t\}, \beta, \{\sigma_i^2\}$ using only $(i,t) \in \mathcal{O}$.
2. Update $\{f_t\}$ conditionally on $\{\lambda_i\}, \beta, \{\sigma_i^2\}$ using only times $t < T_i$ if $i \in \text{test}$.
3. Update β from its Gaussian conditional.
4. Update each σ_i^2 from its inverse-Gamma/Gamma conditional (with residual sums scaled by ω).

After discarding burn-in, we retain M_p draws for prediction. For each retained draw m,

$$\mu_{it}^{(m)} = \lambda_t^{(m)\top} f_i^{(m)} + X_{it}^\top \beta^{(m)}, \qquad \tilde{Y}_{it}^{(m)} = \mu_{it}^{(m)} + \tilde{\varepsilon}_{it}^{(m)}, \;\; \tilde{\varepsilon}_{it}^{(m)} \sim \mathcal{N}\Big(0, \frac{\sigma_i^{2(m)}}{\omega}\Big).$$

We report

$$\widehat{\mu}_{it} = \mathsf{E}_m[\mu_{it}^{(m)}], \quad \widehat{Y}_{it}^{\text{lo}} = \mathsf{q}_{0.025}(\tilde{Y}_{it}^{(m)}), \quad \widehat{Y}_{it}^{\text{hi}} = \mathsf{q}_{0.975}(\tilde{Y}_{it}^{(m)}).$$

On the panel used to select K and ω we randomly mask 20% of cells and, for each (K, ω), fit the model to the partially observed matrix. Validation uses only masked cells (i,t):

$$\text{Bias}^2 = \mathsf{mean}\,(\widehat{\mu}_{it} - Y_{it})^2, \qquad \text{IS} = \mathsf{mean}\left[(\widehat{Y}_{it}^{\text{hi}} - \widehat{Y}_{it}^{\text{lo}}) + \tfrac{2}{\alpha}(\widehat{Y}_{it}^{\text{lo}} - Y_{it})\mathbf{1}\{Y_{it} < \widehat{Y}_{it}^{\text{lo}}\} + \tfrac{2}{\alpha}(Y_{it} - \widehat{Y}_{it}^{\text{hi}})\mathbf{1}\{Y_{it} > \widehat{Y}_{it}^{\text{hi}}\}\right],$$

with $\alpha = 0.05$. We pick $(K^\star, \omega^\star)$ minimizing $\text{Bias}^2 + \text{IS}$.

Wild bootstrap intervals for matrix completion

The `gsynth` r-package provides counterfactual paths $\widehat{Y}_{it}^{\text{mc}}$ for treated columns by using the matrix completion framework [Athey et al., 2021, Xu, 2017]. We construct predictive intervals on the masked test cells via a wild bootstrap procedure that preserves the panel structure and the placebo masking. Let $\mathcal{T}$ denote the set of test units retained by `gsynth` after name-safe matching. Define residuals on the masked test cells

$$E_{it} \;=\; Y_{it} - \widehat{Y}_{it}^{\text{mc}}, \qquad (i,t) \in \mathcal{M} \cap \{i \in \mathcal{T}\}.$$

For $b = 1, \ldots, B$ (e.g. $B = 200$):

1. Draw independent Rademacher multipliers $\xi_i^{(b)} \in \{-1, +1\}$ for each $i \in \mathcal{T}$, and set $S_{it}^{(b)} = \xi_i^{(b)}$ for all t (unit-wise sign flips).
2. Form starred treated trajectories on masked test cells:

$$Y_{it}^{\star(b)} \;=\; \widehat{Y}_{it}^{\text{mc}} + S_{it}^{(b)} E_{it}, \quad (i,t) \in \mathcal{M} \cap \{i \in \mathcal{T}\},$$

and replace the corresponding columns in the full controls panel Y by $Y_{\cdot i}^{\star(b)}$ at the *same unit names*.

3. Refit `gsynth` on the starred panel (same placebo design Z_{it}, same options), retrieve $\widehat{Y}_{it}^{\mathrm{mc},(b)}$ for $i \in \mathcal{T}$, and extract predictions *only* at the original masked cells $(i,t) \in \mathcal{M} \cap \{i \in \mathcal{T}\}$.

For each masked cell, take empirical quantiles over $b = 1, \ldots, B$:

$$\widehat{Y}_{it}^{\mathrm{lo}} = \mathsf{q}_{0.025}\big(\widehat{Y}_{it}^{\mathrm{mc},(b)}\big), \qquad \widehat{Y}_{it}^{\mathrm{hi}} = \mathsf{q}_{0.975}\big(\widehat{Y}_{it}^{\mathrm{mc},(b)}\big).$$

On the masked test set we compute

$$\mathrm{Bias}^2 = \mathsf{mean}\big(\widehat{Y}_{it}^{\mathrm{mc}} - Y_{it}\big)^2, \qquad \mathrm{IS}^{\mathrm{mc}} = \mathrm{IS}\big(Y_{it}; \widehat{Y}_{it}^{\mathrm{lo}}, \widehat{Y}_{it}^{\mathrm{hi}}\big),$$

and their per-time averages across units with $t \geq T_i$. Name-safe alignment (by unit labels) is enforced in every bootstrap refit and extraction step to prevent column-order drift.

Additional simulation results

This appendix complements the main simulation results in Section 4.2 by providing a more detailed view of time–unit–specific causal estimation performance across varying degrees of latent factor mis-specification.

For each value of $\beta_U \in \{0, 1, 2, 5\}$, we consider the optimally selected configuration $(K^\star, \widehat{\omega})$ and evaluate predictive accuracy for $\tau_{\mathrm{ATT}}(t)$ exclusively on masked post-treatment pseudo-treated observations (see Section 4.2 for design details). We report squared bias $\big(\widehat{\tau}_{\mathrm{ATT}}(t) - \tau_{\mathrm{ATT}}(t)\big)^2$, and 95% interval score, both averaged across units that are masked at each time point t. These figures provide a dynamic view of causal prediction quality beyond the global averages shown in Figure 4 in the main text.

Figures 8–10 show that RBCI is able to track the post-treatment dynamics with improved accuracy relative to matrix completion (MC) when latent confounding is present. In the low and medium mis-specification settings, RBCI achieves a uniformly lower interval score and reduced squared bias across most post-treatment years. Under high mis-specification these gains become substantial, reflecting the ability of the selected learning rate $\widehat{\omega} < 1$ to temper overconfident likelihood contributions and improve predictive robustness. Importantly, RBCI does not degrade coverage or point accuracy at earlier post-treatment times where latent confounding exerts weaker influence, further supporting the benefits of adaptively tuning (K, ω).
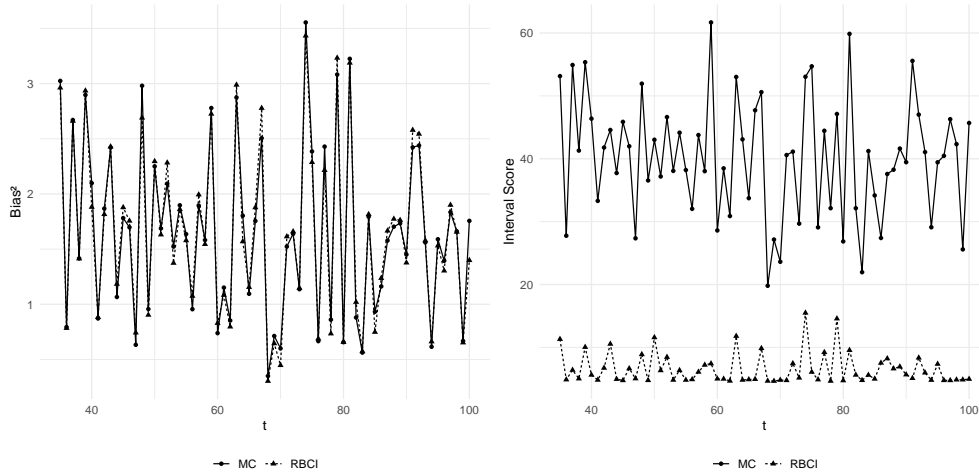
Figure 8: Time-varying squared bias (left) and interval score (right) under low latent factor mis-specification ($\beta_U = 1$). Robust Bayesian causal inference (RBCI) yields consistently lower interval scores and improved point estimation relative to matrix completion (MC).
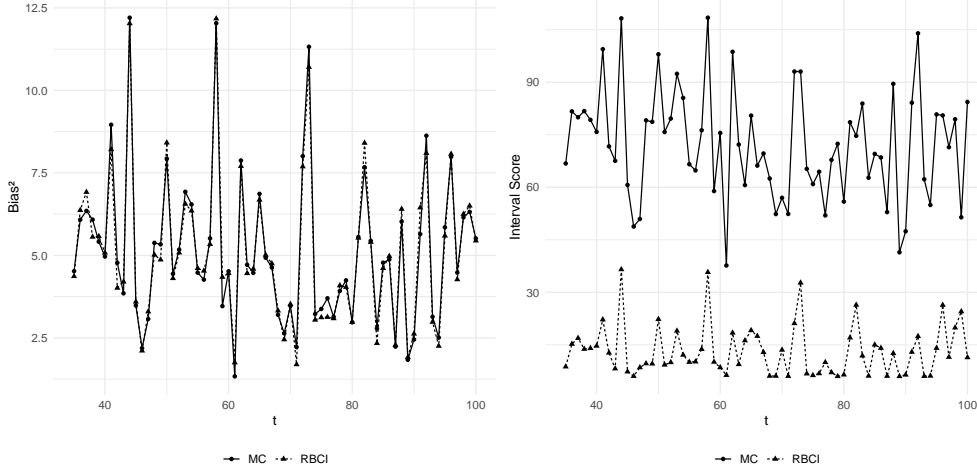
Figure 9: Medium latent factor mis-specification ($\beta_U = 2$). Performance gains of robust Bayesian causal inference (RBCI) increase relative to the low mis-specification scenario against matrix completion (MC).
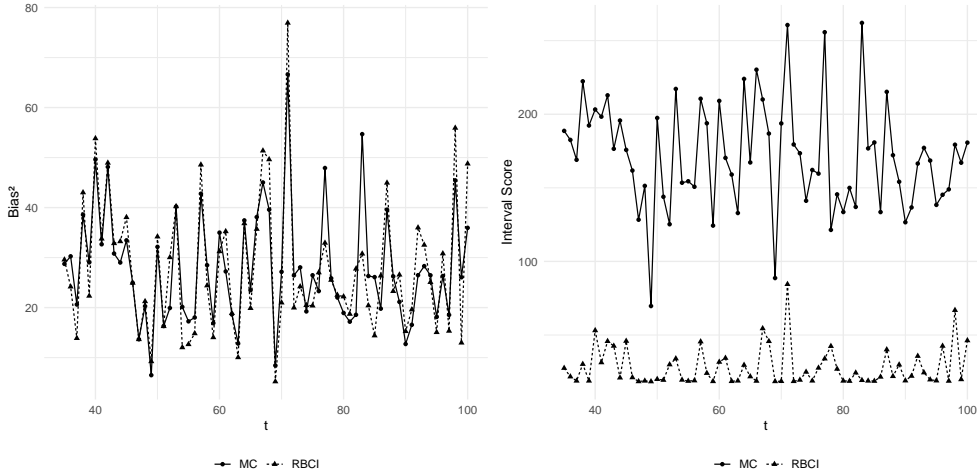

Figure 10: High latent factor mis-specification ($\beta_U = 5$). Robust Bayesian causal inference (RBCI) shows reductions in squared bias and interval score compared with matrix completion (MC), highlighting robustness to severe unobserved confounding.

8 Real datasets details

8.1 California Smoking Data

We provide further details on the California tobacco intervention dataset analysed in Section 5.1. The balanced panel includes annual per-capita cigarette consumption for 39 U.S. states from 1970 to 2000, with all outcomes log-transformed prior to analysis. California is treated as the intervention state following Proposition 99 in 1989, while the remaining 38 states serve as never-treated controls. To enable unbiased hyperparameter selection, we randomly mask 20% of pre-intervention observations belonging to control states and evaluate predictive performance on these held-out entries. To assess causal prediction accuracy, we additionally construct a placebo evaluation set by masking post–1989 outcomes for a random 15% subset of control states and treating those cells as pseudo-treated. Both masking steps strictly respect time ordering, ensuring valid causal evaluation. Figure S1 displays the panel layout, showing that all states are observed in every year of the study period and highlighting the absence of missing observations outside the deliberately masked entries. Figure S2 plots the average (log) cigarette sales in California and the corresponding mean across control states. The parallel pre-intervention trajectories and the clear divergence after 1989 emphasise the need for factor-based counterfactual estimation.
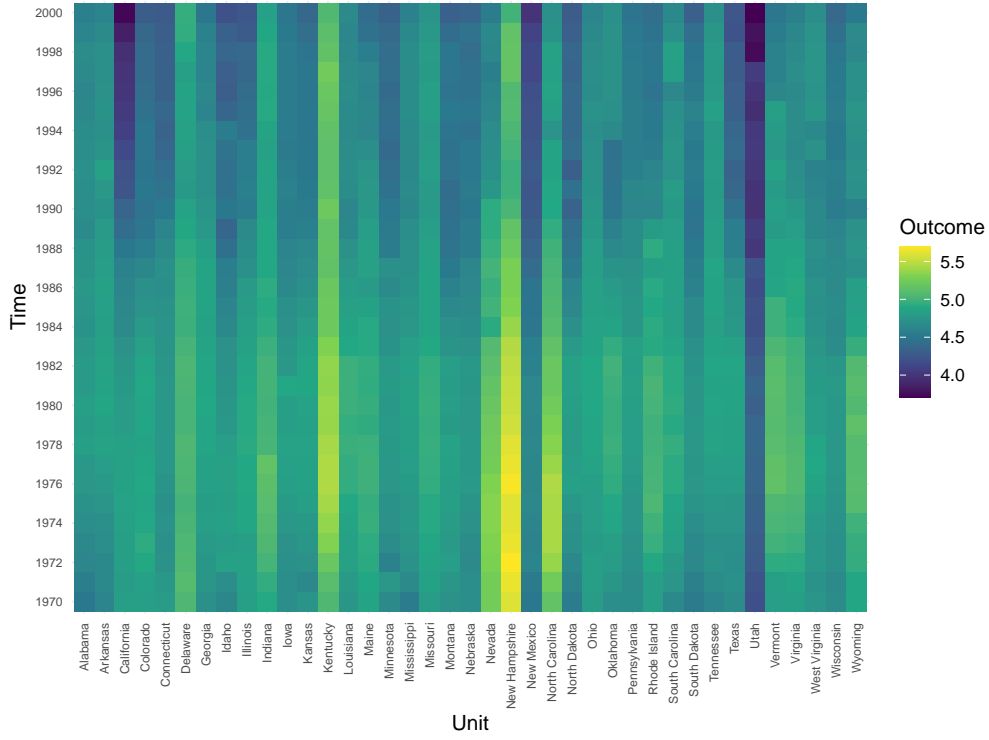
Figure 11: Panel structure for the California smoking dataset: log per-capita cigarette sales observed for 39 states over 1970–2000. California is the treated unit (red stripe at bottom); controls shown above.
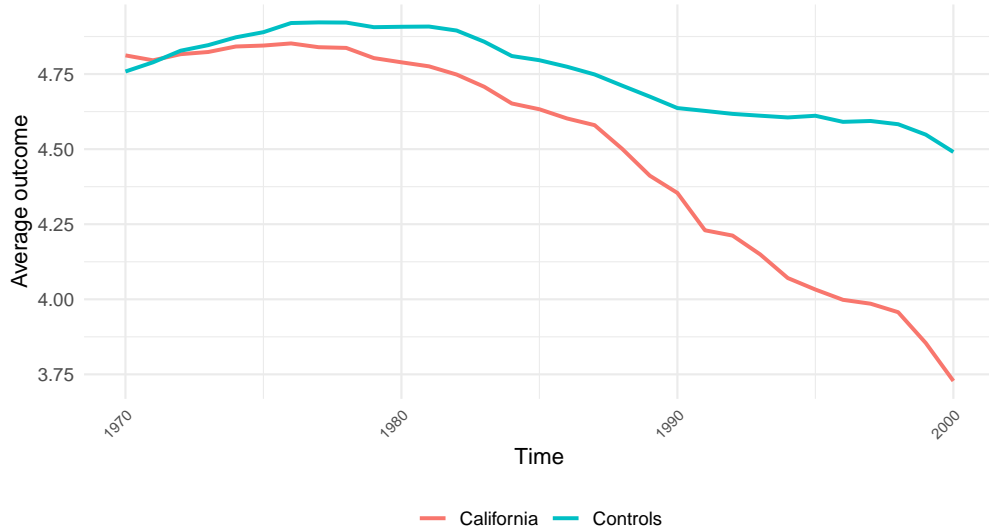
Figure 12: Average log per-capita cigarette sales for California and the mean across never-treated states, 1970–2000. Proposition 99 is enacted in 1989.

8.2 French regional industrial policy dataset

Here, we provide exploratory figures for the French regional industrial policy dataset used in Section 5.2. The panel comprises annual observations for $N = 148$ French employment areas over $T = 20$ years (1993–2012), with 13 treated regions and 135 never-treated controls. We analyse three outcomes: firm entries, firm exits, and total employment.
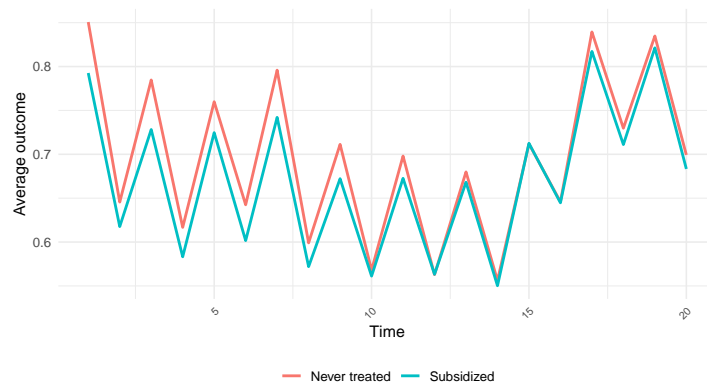
Figure 13: Mean firm entry rates over time for subsidized against never-treated regions.
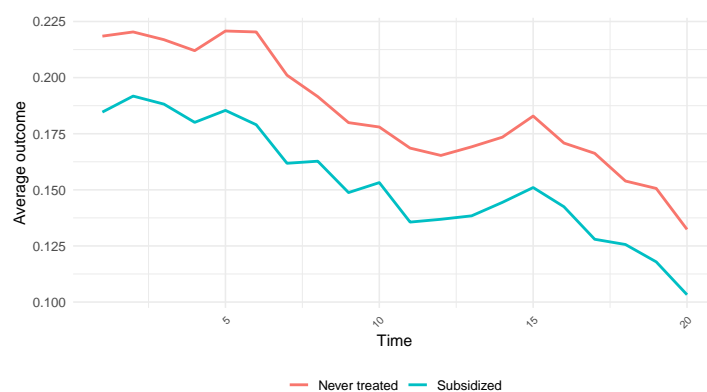
Figure 14: Mean firm exit rates over time for subsidized vs. never-treated regions.
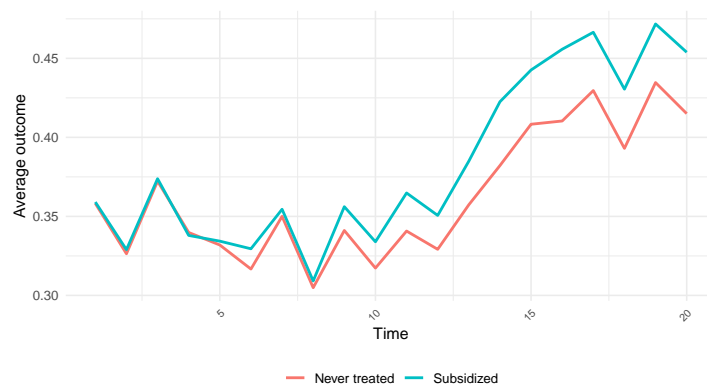
Figure 15: Mean total employment over time for subsidized vs. never-treated regions.

Department of Economics
Athens University of Economics and Business

List of Recent Working Papers

2023

01-23   Real interest rate and monetary policy in the post Bretton Woods United States, George C. Bitros and Mara Vidali

02-23   Debt targets and fiscal consolidation in a two-country HANK model: the case of Euro Area, Xiaoshan Chen, Spyridon Lazarakis and Petros Varthalitis

03-23   Central bank digital currencies: Foundational issues and prospects looking forward, George C. Bitros and Anastasios G. Malliaris

04-23   The State and the Economy of Modern Greece. Key Drivers from 1821 to the Present, George Alogoskoufis

05-23   Sparse spanning portfolios and under-diversification with second-order stochastic dominance, Stelios Arvanitis, Olivier Scaillet, Nikolas Topaloglou

06-23   What makes for survival? Key characteristics of Greek incubated early-stage startup(per)s during the Crisis: a multivariate and machine learning approach, Ioannis Besis, Ioanna Sapfo Pepelasis and Spiros Paraskevas

07-23   The Twin Deficits, Monetary Instability and Debt Crises in the History of Modern Greece, George Alogoskoufis

08-23   Dealing with endogenous regressors using copulas; on the problem of near multicollinearity, Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis

09-23   A machine learning approach to construct quarterly data on intangible investment for Eurozone, Angelos Alexopoulos and Petros Varthalitis

10-23   Asymmetries in Post-War Monetary Arrangements in Europe: From Bretton Woods to the Euro Area, George Alogoskoufis, Konstantinos Gravas and Laurent Jacque

11-23   Unanticipated Inflation, Unemployment Persistence and the New Keynesian Phillips Curve, George Alogoskoufis and Stelios Giannoulakis

12-23   Threshold Endogeneity in Threshold VARs: An Application to Monetary State Dependence, Dimitris Christopoulos, Peter McAdam and Elias Tzavalis

13-23   A DSGE Model for the European Unemployment Persistence, Konstantinos Giakas

14-23   Binary public decisions with a status quo: undominated mechanisms without coercion, Efthymios Athanasiou and Giacomo Valletta

15-23   Does Agents' learning explain deviations in the Euro Area between the Core and the Periphery? George Economides, Konstantinos Mavrigiannakis and Vanghelis Vassilatos

16-23   Mild Explocivity, Persistent Homology and Cryptocurrencies' Bubbles: An Empirical Exercise, Stelios Arvanitis and Michalis Detsis

17-23   A network and machine learning approach to detect Value Added Tax fraud, Angelos Alexopoulos, Petros Dellaportas, Stanley Gyoshev, Christos Kotsogiannis, Sofia C. Olhede, Trifon Pavkov

18-23   Time Varying Three Pass Regression Filter, Yiannis Dendramis, George Kapetanios, Massimiliano Marcellino

19-23   From debt arithmetic to fiscal sustainability and fiscal rules: Taking stock, George Economides, Natasha Miouli and Apostolis Philippopoulos

20-23   Stochastic Arbitrage Opportunities: Set Estimation and Statistical Testing, Stelios Arvanitis and Thierry Post

21-23   **Behavioral Personae, Stochastic Dominance, and the Cryptocurrency Market, Stelios Arvanitis, Nikolas Topaloglou, and Georgios Tsomidis**

22-23   **Block Empirical Likelihood Inference for Stochastic Bounding: Large Deviations Asymptotics Under m-Dependence, Stelios Arvanitis and Nikolas Topaloglou**

23-23   **A Consolidation of the Neoclassical Macroeconomic Competitive General Equilibrium Theory via Keynesianism (Part 1 and Part 2), Angelos Angelopoulos**

24-23   **Limit Theory for Martingale Transforms with Heavy-Tailed Noise, Stelios Arvanitis and Alexandros Louka**

2024

01-24   **Market Timing & Predictive Complexity, Stelios Arvanitis, Foteini Kyriazi, Dimitrios Thomakos**

02-24   **Multi-Objective Frequentistic Model Averaging with an Application to Economic Growth, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, Nikolas Topaloglou**

03-24   **State dependent fiscal multipliers in a Small Open Economy, Xiaoshan Chen, Jilei Huang, Petros Varthalitis**

04-24   **Public debt consolidation: Aggregate and distributional implications in a small open economy of the Euro Area, Eleftherios-Theodoros Roumpanis**

05-24   **Intangible investment during the Global Financial Crisis in the EU, Vassiliki Dimakopoulou, Stelios Sakkas and Petros Varthalitis**

06-24   **Time will tell! Towards the construction of instantaneous indicators of different agent-types, Iordanis Kalaitzoglou, Stelios Arvanitis**

07-24   **Norm Constrained Empirical Portfolio Optimization with Stochastic Dominance: Robust Optimization Non-Asymptotics, Stelios Arvanitis**

08-24   **Asymptotics of a QLR-type test for optimal predictive ability, Stelios Arvanitis**

09-24   **Strongly Equitable General Equilibrium Allocations, Angelos Angelopoulos**

10-24   **The Greek macroeconomy: A note on the current situation and future outlook, Apostolis Philippopoulos**

11-24   **Evolution of Greek Tax System, A Survey of Legislated Tax Changes from 1974 to 2018, Panagiotis Asimakopoulos**

12-24   **Macroeconomic Impact of Tax Changes, The case of Greece from 1974 to 2018, Panagiotis Asimakopoulos**

13-24   **`Pareto, Edgeworth, Walras, Shapley' Equivalence in a Small Economy, Angelos Angelopoulos**

14-24   **Stimulating long-term growth and welfare in the U.S, James Malley and Apostolis Philippopoulos**

15-24   **Distributionally Conservative Stochastic Dominance via Subsampling, Stelios Arvanitis**

16-24   **Before and After the Political Transition of 1974. Institutions, Politics, and the Economy of Post-War Greece, George Alogoskoufis**

17-24   **Gaussian Stochastic Volatility, Misspecified Volatility Filters and Indirect Inference Estimation, Stelios Arvanitis, Antonis Demos**

18-24   **Endogenous Realtor Intermediation and Housing Market Liquidity, Miroslav Gabrovski, Ioannis Kospentaris, Victor Ortego-Marti**

19-24   **Universal Choice Spaces and Expected Utility: A Banach-type Functorial Fixed Point, Stelios Arvanitis, Pantelis Argyropoulos, Spyros Vassilakis**

2025

01-25   **Democracy, redistribution, and economic growth: Some evidence from post-1974 Greece, George C. Bitros**

02-25   **Assessing Downside Public Debt Risks in an Environment of Negative Interest Rates Growth Differentials, Yiannis Dendramis, Georgios Dimitrakopoulos and Elias Tzavalis**

03-25   **Wicksellian General Equilibrium, Angelos Angelopoulos**

04-25   **Institutions and Ethics, Angelos Angelopoulos**

Department of Economics
Athens University of Economics and Business

The Department is the oldest Department of Economics in Greece with a pioneering role in organising postgraduate studies in Economics since 1978. Its priority has always been to bring together highly qualified academics and top quality students. Faculty members specialize in a wide range of topics in economics, with teaching and research experience in world-class universities and publications in top academic journals.

The Department constantly strives to maintain its high level of research and teaching standards. It covers a wide range of economic studies in micro-and macroeconomic analysis, banking and finance, public and monetary economics, international and rural economics, labour economics, industrial organization and strategy, economics of the environment and natural resources, economic history and relevant quantitative tools of mathematics, statistics and econometrics.

Its undergraduate program attracts high quality students who, after successful completion of their studies, have excellent prospects for employment in the private and public sector, including areas such as business, banking, finance and advisory services. Also, graduates of the program have solid foundations in economics and related tools and are regularly admitted to top graduate programs internationally. Three specializations are offered:1. Economic Theory and Policy, 2. Business Economics and Finance and 3. International and European Economics. The postgraduate programs of the Department (M.Sc and Ph.D) are highly regarded and attract a large number of quality candidates every year.

For more information:

https://www.dept.aueb.gr/en/econ/