Empirical Bayes Estimation in Heterogeneous Coefficient Panel Models

Myunghyun Song* Sokbae Lee† Serena Ng‡

February 4, 2026

Abstract

We develop an empirical Bayes (EB) G-modeling framework for short-panel linear models with multidimensional heterogeneity and nonparametric prior. Specifically, we allow heterogeneous intercepts, slopes, dynamics, and a non-spherical error covariance structure. We establish identification and consistency of the nonparametric maximum likelihood estimator (NPMLE) under general conditions, and provide low-level sufficient conditions for several models of empirical interest. Conditions for regret consistency of the EB estimators are also established. The NPMLE is computed using a Wasserstein-Fisher-Rao gradient flow algorithm adapted to panel regressions. Using data from the Panel Study of Income Dynamics, we find that the slope coefficient for potential experience is substantially heterogeneous and negatively correlated with the random intercept, and that error variances and autoregressive coefficients vary significantly across individuals. The EB estimates reduce mean squared prediction errors relative to individual maximum likelihood estimates.

Keywords: G-modeling, nonparametric maximum likelihood, shrinkage estimation, Wasserstein-Fisher-Rao gradient flow, income dynamics

*Department of Economics, Columbia University

†Department of Economics, Columbia University and Institute for Fiscal Studies

‡Department of Economics, Columbia University and NBER

We thank Roger Koenker, Soonwoo Kwon, Bodhisattva Sen, Kaizheng Wang, and seminar participants at Indiana University, the University of Pennsylvania, the University of Toronto, and Yale University for helpful and encouraging comments. The third author would like to thank the National Science Foundation for financial support (SES: 2018369).

1 Introduction

Understanding differences in individual behavior is the primary goal of many economic analyses. Why do workers with similar experience exhibit different earnings profiles? Why do economic agents with similar observed characteristics respond differently to economic shocks and policy interventions? A challenge for empirical researchers in addressing these questions is the presence of individual-level differences that affect behavior but are not captured by observed variables, commonly referred to as unobserved heterogeneity. While it is not hard to allow for unit-specific intercepts in linear regressions, entertaining heterogeneous slope parameters and innovation variances is more challenging, and in a frequentist setting, this usually requires parametric assumptions to gain tractability at the cost of robustness.

This paper considers a setting in which we observe $\{(Y_{it}, X_{2,it}) : t = 1, \ldots, T\}$ for each unit $i = 1, \ldots, N$, assuming they are independent and identically distributed (i.i.d.) across i. We propose a framework for short-T linear panel models with heterogeneous intercepts, slope coefficients, and error distributions. To fix ideas, consider the dynamic linear panel model with an exogenous covariate $X_{2,it}$ and AR(1) errors:

$$Y_{it} = a_i + b_i X_{2,it} + u_{it}, \tag{1a}$$

$$u_{it} = \rho_i u_{it-1} + \sigma_i e_{it}, \quad e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad t \geq 1, \tag{1b}$$

where $a_i \in \mathbb{R}$ is a random intercept, $b_i \in \mathbb{R}$ a random slope, $\rho_i \in (-1, 1)$ a random autoregressive coefficient, $\sigma_i^2 > 0$ a random variance, and u_{i0} follows the stationary distribution $\mathcal{N}(0, \sigma_i^2/(1-\rho_i^2))$ conditional on (σ_i^2, ρ_i). We refer to the model in (1) as the HIVDX model since it admits heterogeneity in four dimensions: intercepts (I), error variances (V), dynamics (D) through the autoregressive coefficient, and the coefficient on the exogenous covariate X_2. The widely studied Gaussian location-scale model $Y_{it} = a_i + \sigma_i e_{it}$, as well as models that impose $\rho_i = \rho$ and $b_i = b$, are special cases of HIVDX.

We analyze the HIVDX model and the special cases in Section 2 using a general framework that partitions the random parameters into two sets: regression coefficients $\beta_i = (a_i, b_i)$ of dimension $d_\beta = 2$, and covariance parameters $\delta_i = (\sigma_i^2, \rho_i)$ of dimension $d_\delta = 2$. The parameters can be estimated separately for each i by max-

imum likelihood. Instead, we assume that the random parameters $\theta_i \equiv (\beta_i, \delta_i)$ are drawn from a true but unknown common distribution G_*. Viewing G_* as a prior, the optimal decision rule for each unit under compound squared-error loss is the posterior mean $\theta_i^* := \mathbb{E}_{G_*}[\theta_i \mid Y_i, X_{2,i}]$, where $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ is a $T \times 1$ vector of outcomes, $X_{2,i} = (X_{2,i1}, \ldots, X_{2,iT})'$ is a vector of observed covariates, and $\mathbb{E}_{G_*}[\cdot]$ denotes expectation with respect to $\theta_i \sim G_*$.

Specifying that $\theta_i \overset{\text{i.i.d.}}{\sim} G_*$ is grounded in the principle of empirical Bayes (EB) introduced by Robbins (1956). This approach is intrinsically frequentist because it estimates the prior distribution from the data. EB estimation improves on individual-level maximum likelihood estimation through shrinkage, borrowing strength across units to learn both the individual parameters θ_i and their population distribution G_*. Following the terminology of Efron (2014), the procedure is referred to as G-modeling when G_* is estimated directly. The resulting EB estimator is given by $\hat{\theta}_i^{\text{EB}} := \mathbb{E}_{\hat{G}}[\theta_i \,|\, Y_i, X_{2,i}]$, where $\hat{G}$ denotes an estimator of G_*. An EB estimator $\hat{\tau}_i^{\text{EB}}(\theta_i) = \mathbb{E}_{\hat{G}}[\tau_i \,|\, Y_i, X_{2,i}]$ for general unit-specific parameters τ_i defined as functions of θ_i can also be defined. While parametric G-modeling is convenient, nonparametric modeling guards against misspecification. Recent reviews of EB and its applications in economics include Koenker and Gu (2024), Koenker and Gu (2025), and Walters (2024). To our knowledge, no nonparametric EB analysis has considered all four sources of heterogeneity present in the HIVDX model.

We study identification and estimation of the HIVDX model within an EB framework. The generality of our model with $d_\beta > 1$ and $d_\delta > 0$ introduces several challenges. First, while assuming the identification of G_* in the Gaussian location model with i.i.d. errors without formal proof may be defensible, this assumption cannot be taken for granted in the more complex settings. Theorem 1 establishes that identification requires: (i) the design matrix to have full column rank almost surely, and (ii) the covariance structure to remain identifiable after removing the mean component by appropriately differencing the data. These conditions imply that $d_\beta < T$, leaving $T - d_\beta$ degrees of freedom to identify the covariance structure; this imposes a constraint on the dimension of heterogeneous variance parameters (d_δ).

We apply Theorem 1 to the location model ($d_\beta = 1$) and illustrate how differencing facilitates identification. Specifically, we show that $T \geq 3$ suffices for AR(1) errors (Proposition 1), whereas ARMA(1,1) errors require $T \geq 4$ (Proposition 2). The analysis is, however, more involved when $d_\beta > 1$, as the mean and variance components

are convolved within the outcome distribution. Now whether and how G_* can be identified depends critically on the specific time-series variation of the covariate. We examine two canonical examples. The first example includes models with a common time trend and heterogeneous initial conditions (e.g., potential experience in income dynamics). For this example, Proposition 3 establishes that G_* is nonparametrically identified with $T \geq 4$. The second example includes a common mean shift as a special case. For this case, Proposition 4 shows that identification of G_* is possible with $T \geq 5$. Identification must be worked out on a case by case basis.

The second challenge lies in establishing the consistent estimation of the HIVDX model. The classic proof of Kiefer and Wolfowitz (1956) assumes compactness, but when β_i is allowed to be unbounded, the space of mixing distributions is non-compact. While the case of $d_\beta = 1$ can be resolved by embedding $\mathbb{R}$ into its compact completion $[-\infty, \infty]$, the multidimensional setting ($d_\beta \geq 2$) presents unique difficulties because mass can diverge along infinitely many distinct directions. To address this, Section 4 adopts a metric induced by the vague topology to effectively compactify the space of prior distributions. Theorem 2 establishes the almost sure consistency of the nonparametric maximum likelihood estimator (NPMLE) with respect to this metric. Building on this result, Theorem 3 in Section 5 provides conditions for regret consistency (equivalently, asymptotic optimality) of the EB estimator for τ_i, taking into account that τ_i may involve functions of the random coefficients and the data. For example, the optimal (but infeasible) one-step-ahead prediction under quadratic loss is $\tau_i^* = \mathbb{E}_{G_*}[a_i + b_i X_{2,iT+1} + \rho_i Y_{iT} \,|\, Y_i, X_i, X_{2,iT+1}] - \mathbb{E}_{G_*}[\rho_i a_i + \rho_i b_i X_{2,iT} \,|\, Y_i, X_i]$. The prediction error $\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*$ thus involves interaction of the errors in estimating the random coefficients and (Y_i, X_i). Proposition 5 provides low-level sufficient conditions for the EB prediction to be regret-consistent.

The third challenge arises from the computational complexity associated with multiple sources of heterogeneity. In the analysis of location models, the NPMLE of G is typically computed by optimizing over a fixed grid of m points. The location of the grid points and the size of the grid are usually fixed during the iterations. However, this fixed-grid approach can be intractable as d_θ increases because the grid size required to maintain accuracy needs to grow exponentially. It would be desirable to have an algorithm that allows the support points to move adaptively. This motivates a shift from Euclidean optimization to optimization on the space of probability measures, where the updates are governed by gradient flows. In this framework, the

choice of geometry is critical. For example, the Fisher-Rao gradient flow corresponds to reweighting fixed support points, whereas the Wasserstein gradient flow transports mass across the parameter space, effectively moving the support. Recently, Yan, Wang, and Rigollet (2024) demonstrate the theoretical and computational advantages of combining these geometries for the Gaussian location mixture model. Motivated by this encouraging result, Section 6 develops a Wasserstein-Fisher-Rao (WFR) gradient flow algorithm for panel regressions with multiple sources of heterogeneity.

We use the EB approach to analyze the income data in the Panel Study of Income Dynamics (PSID) in Section 7. The model allows earnings Y_{it} to have unit-specific mean a_i and response b_i to potential experience $X_{2,it}$, with a heteroskedastic error covariance structure that depends on internal dynamics ρ_i and exposure to shocks through σ_i^2. We find that (i) the slope coefficient b_i for potential experience exhibits substantial heterogeneity and is negatively correlated with the random intercept a_i, and (ii) there is pronounced cross-sectional heterogeneity in both the variance (σ_i^2) and the dynamics (ρ_i). The EB estimates reduce mean squared prediction errors relative to individual maximum likelihood estimates. Section 8 shows that features of the EB estimates and forecast error reduction are replicated in Monte Carlo experiments calibrated to the application.

Before turning to the main part of the paper, we briefly discuss recent work that, while not directly related to our approach, are conceptually connected to the broader EB literature. Chen (2024) analyzes conditionally Gaussian settings with known variances σ_i^2, constructing studentized data by estimating the conditional mean and variance of outcomes given σ_i^2, and then estimating the prior for use in EB decision rules. The paper emphasizes the importance of nonparametric estimation in both stages. In contrast, we allow for unobserved σ_i^2 and heterogeneous dynamics across multiple time periods.

Kwon (2025) considers a multivariate Gaussian model with a hierarchical normal prior, while Cheng, Ho, and Schorfheide (2025) propose a parametric EB estimator for two-way effects with assortative matching. Both methods rely on linear shrinkage parameters selected by minimizing unbiased risk estimates. In contrast, we adopt a fully nonparametric EB approach via NPMLE. This avoids distributional assumptions and the restriction to linear shrinkage, allowing us to recover the optimal (potentially non-linear) decision rule and accommodate arbitrary dependence among the latent parameters.

Gaillac (2024) develops posterior mean estimators for cross-sectional random coefficient models in which the coefficients are independent of the regressors. In a panel extension, the paper constructs time averages for each unit and allows heterogeneity only in time-invariant regressors, while assuming homogeneous coefficients for time-varying regressors. In contrast, our framework focuses on random coefficients associated with regressors that vary over time.

Adusumilli, Gu, and Tao (2025) investigate EB methods in compound adaptive experiments, where the outcome of each arm follows a normal distribution with an unknown mean. They demonstrate that the risk guarantees for G-modeling, originally derived under i.i.d. sampling, continue to hold for adaptively collected data without knowledge of the sampling algorithm. In contrast, F-modeling is shown to yield biased estimates in this setting.

Recent work has extended nonparametric EB methods beyond the classical Gaussian location model to a variety of settings, including heteroskedastic models (e.g., Soloff, Guntuboyina, and Sen, 2025), high-dimensional or multivariate regression (e.g., Mukherjee, Sen, and Sen, 2023; Fan, Guan, Shen, and Wu, 2023; Jiang and Guntuboyina, 2025), and variance estimation (e.g., Ignatiadis and Sen, 2025). These works only consider a static model while our focus is the heterogeneous dynamic panel model presented above as HIVDX.

2 The Econometric Setup and EB Modeling

We consider a panel of observations on N cross-sectional units over T time periods, focusing on a short-panel framework in which N grows while T remains fixed. Our main setup is a general heterogeneous coefficient (HC) model that nests the HIVDX model:

$$Y_i = X_i\beta_i + P_i e_i, \tag{2}$$

where $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ is a $T \times 1$ vector of outcomes, $X_i = (X_{i1}, \ldots, X_{iT})'$ is a $T \times d_\beta$ matrix of observed covariates, and $e_i = (e_{i1}, \ldots, e_{iT})'$ is a vector of independent Gaussian errors. We assume that the first column of X_i is ι_T, the T-dimensional vector of ones. For each i, P_i is lower triangular, and the corresponding conditional covariance matrix $\Sigma_i \equiv P_i P_i'$ is positive definite. To capture its dependence on heterogeneous variance parameters $\delta_i \in \mathbb{R}^{d_\delta}$, we write $P_i = P(\delta_i)$, where $\delta \mapsto P(\delta)$ is a

continuously differentiable function of δ.

To see that the HIVDX model is a special case of the HC model for particular (β_i, δ_i), note that since e_{it} are i.i.d. $\mathcal{N}(0,1)$, $u_i = (u_{i1}, \ldots, u_{iT})'$ is normally distributed with conditional covariance matrix Σ_i, whose (t,s) entry is given by

$$\Sigma_{i,ts} = \sigma_i^2 \frac{\rho_i^{|t-s|}}{1-\rho_i^2}, \quad t,s = 1,\ldots,T. \tag{3}$$

Hence, P_i is defined via the Cholesky decomposition of $\Sigma_i = P_i P_i'$ with $\delta_i = (\sigma_i^2, \rho_i)$. We develop our theoretical framework through the lens of the HC model since it provides a unified way to analyze a broader class of dynamics such as ARMA errors. Let I_T denote the $T \times T$ identity matrix. The following assumptions will be used throughout.

Assumption 1 (DGP)**.** Let $(Y_i, X_i)_{i=1}^N$ be identically and independently distributed (i.i.d.) data, satisfying (2). In addition, the following conditions hold.

(i) The pairs $(\beta_i, \delta_i)_{i=1}^N$ are i.i.d. draws from an unknown distribution G_* and are independent of $(X_i)_{i=1}^N$.

(ii) The parameter space for δ_i is a compact subset $\mathcal{K}_\delta \subseteq \mathbb{R}^{d_\delta}$. For each $\delta_i \in \mathcal{K}_\delta$, the covariance matrix $\Sigma_i \equiv P(\delta_i)P(\delta_i)'$ satisfies $\underline{c}\, I_T \leq \Sigma_i \leq \bar{c}\, I_T$ for some $0 < \underline{c} < \bar{c} < \infty$.

(iii) $e_i \sim \mathcal{N}(0, I_T)$.

Conditional on (β_i, δ_i) and X_i, the outcome vector Y_i follows the distribution $\mathcal{N}(X_i\beta_i, \Sigma_i)$. Individuals in this model are distinguished by types characterized by their heterogeneous features. Given the information structure, individual types cannot be directly uncovered from the data, but their distributions can be identified.

Let $\theta_i := (\beta_i, \delta_i)$ collect all heterogeneous parameters, lying in the space $\Theta := \mathbb{R}^{d_\beta} \times \mathcal{K}_\delta \subseteq \mathbb{R}^{d_\theta}$ with $d_\theta = d_\beta + d_\delta$. If θ_i were observed, the conditional density of Y_i given X_i is

$$\ell(Y_i \,|\, X_i, \theta_i) := \frac{1}{(2\pi)^{T/2}|P(\delta_i)|} \exp\left(-\frac{1}{2}\|P(\delta_i)^{-1}(Y_i - X_i\beta_i)\|^2\right).$$

When θ_i is unobserved but its distribution is given as G, the marginal density of Y_i given X_i, obtained by integrating out θ_i from the likelihood, is

$$f_G(Y_i, X_i) := \int_\Theta \ell(Y_i \mid X_i, \theta) dG(\theta),$$

where the independence between θ_i and X_i is used.[1]

For the estimation of $(\theta_i)_{i=1}^N$, two conceptual frameworks are available. The first is the *separate* decision framework that treats each θ_i as fixed and minimizes the individual risk $\mathbb{E}[\|\hat{\theta}_i - \theta_i\|^2 \mid \theta_i]$. This yields the individual MLE,

$$\hat{\theta}_i^{\text{MLE}} := \arg\max_{\theta \in \Theta} \ell(Y_i \mid X_i, \theta),$$

which provides an asymptotically efficient estimator for θ_i as $T \to \infty$. However, Stein (1956) and James and Stein (1961) demonstrate for independent Gaussian observations that the MLE is inadmissible under squared-error loss when the dimension of the parameter vector is at least three (here, $N \geq 3$). In this setting, the MLE is dominated by a shrinkage estimator that achieves lower total risk by pooling information across observations.

The second approach is the *compound* decision framework. Unlike the separate framework, which minimizes risk for each unit individually, this approach seeks to minimize the aggregate risk over the ensemble of N units. While the strict compound decision formulation treats $(\theta_i)_{i=1}^N$ as fixed, Robbins (1956) demonstrates that this problem can be effectively solved by adopting an EB perspective. By modeling the parameters as i.i.d. draws from a common latent distribution G_*, we can derive an optimal (oracle) decision rule, which takes the form of the posterior mean:

$$\theta_i^* := \mathbb{E}_{G_*}[\theta_i \mid Y_i, X_i] = \frac{\int_\Theta \theta \ell(Y_i \mid X_i, \theta) dG_*(\theta)}{\int_\Theta \ell(Y_i \mid X_i, \theta) dG_*(\theta)}. \tag{4}$$

This estimator improves efficiency by pooling information across units, thereby reducing the variance relative to the individual maximum likelihood estimators $\hat{\theta}_i^{\text{MLE}}$. These efficiency gains come at the cost of introducing individual bias (i.e., $\mathbb{E}[\theta_i^* \mid \theta_i] \neq \theta_i$), but the reduction in the overall compound mean squared error dominates this cost.

[1] We use G_* to denote the unknown true distribution and G to denote a generic distribution in the class of possible distributions that includes G_*. The notation $f_G(Y_i, X_i)$ emphasizes that the marginal density of Y_i given X_i depends on G.

In essence, θ_i^* shrinks $\hat{\theta}_i^{\text{MLE}}$ toward the prior.

In practice, the compound decision rule is infeasible due to the unknown G_*. In Bayesian analysis, the prior G_* is typically specified as a parametric distribution, either by fixing its hyperparameters or by assigning them their own distributions. Efron and Morris (1973) show that the James-Stein shrinkage estimator can be interpreted as a parametric EB estimator, with G_* serving as a parametric prior whose hyperparameters are estimated from the data.

An Empirical Bayes analysis can approximate the infeasible oracle decision rule in one of two ways: F-modeling or G-modeling. F-modeling directly computes the shrinkage term from the data without explicit knowledge of G_*. For example, assuming $\Sigma_i \equiv \Sigma$ for some known homogeneous covariance matrix Σ (that is, $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$ for some known ρ and σ^2), the (infeasible) F-modeling EB estimator is

$$\hat{\beta}_i^{\text{F-EB}} := \hat{\beta}_i^{\text{MLE}} + (X_i'\Sigma^{-1}X_i)^{-1}\frac{\partial}{\partial\hat{\beta}}\log\hat{f}_{\hat{\beta}|X}(\hat{\beta}_i^{\text{MLE}}, X_i), \tag{5}$$

where $\hat{\beta}_i^{\text{MLE}} = (X_i'\Sigma^{-1}X_i)^{-1}X_i'\Sigma^{-1}Y_i$ and $\hat{f}_{\hat{\beta}|X}(\hat{\beta}, x)$ is an estimator for $f_{\hat{\beta}|X}(\hat{\beta}, x)$ – the conditional density of $\hat{\beta}_i^{\text{MLE}}$ given $X_i = x$ under G_*. In practice, Σ is typically unknown and must also be estimated. Equation (5), commonly known as *Tweedie's formula* in the literature, shows that the posterior mean of β_i adjusts the corresponding MLE by a shrinkage term derived from the cross-sectional distribution of the MLE.[2] The magnitude of the shrinkage term grows with the variance of $\hat{\beta}_i^{\text{MLE}}$, as measured by $(X_i'\Sigma^{-1}X_i)^{-1}$, and with the informativeness of the prior, as reflected in $f_{\hat{\beta}|X}$.[3]

F-modeling typically relies on the assumption that the likelihood belongs to the linear exponential family, offering computational convenience when this condition holds. Walters (2024) provides a survey of F-modeling applications in labor economics. Liu, Moon, and Schorfheide (2020) estimate the posterior mean of a_i by F-modeling using Tweedie's formula. However, this method does not readily extend to the HIVDX model. We adopt the alternative of G-modeling and explicitly estimates the unknown prior G_*. Shen and Wu (2025) demonstrate that G-modeling

[2] In the simplest case when $Y \mid \theta \overset{\text{ind.}}{\sim} \mathcal{N}(\theta, 1)$ with $T = 1$, the Tweedie formula is $\mathbb{E}_{G_*}[\theta|y] = y + \frac{d}{dy}\log f_{G_*}(y)$, where $f_{G_*}(\cdot)$ denotes the marginal density of Y under G_*.

[3] It can be shown that the oracle-decision errors and the shrinkage term are uncorrelated, that is, $\text{Cov}_{G_*}(\beta_i^* - \beta_i, \hat{\beta}_i^{\text{MLE}} - \beta_i^*) = 0$, where $\beta_i^* := \mathbb{E}_{G_*}[\beta_i \mid Y_i, X_i]$, implying that the variance reduction is attributable to the shrinkage term in (5).

has superior statistical properties in Poisson mixture models than F-modeling.

While parametric G-modeling imposes a specific functional form on the prior, nonparametric G-modeling characterizes the true distribution G_* as the maximizer of the population log-likelihood:

$$G_* \in \arg\max_{G \in \mathcal{G}} F(G), \tag{6}$$

where $\mathcal{G}$ is the class of all probability distributions on Θ and $F(G) := \mathbb{E}[\log f_G(Y_i, X_i)]$ denotes the expected marginal log-likelihood (i.e., the negative risk). This formulation motivates estimating G_* via its sample analog:

$$\hat{G} \in \arg\max_{G \in \mathcal{G}} F_N(G), \tag{7}$$

where the empirical criterion is given by $F_N(G) := N^{-1} \sum_{i=1}^{N} \log f_G(Y_i, X_i)$. We refer to $\hat{G}$ as the NPMLE of G_*. Substituting $\hat{G}$ for G_* in (4) yields the G-modeling EB estimator for θ_i:

$$\hat{\theta}_i^{\text{EB}} := \frac{\int_\Theta \theta \ell(Y_i \,|\, X_i, \theta) d\hat{G}(\theta)}{\int_\Theta \ell(Y_i \,|\, X_i, \theta) d\hat{G}(\theta)}. \tag{8}$$

The NPMLE is built on the same principle as parametric MLE, namely that G_* is the maximizer of the population log-likelihood defined in (6). If G_* is the unique maximizer of F over $\mathcal{G}$, it is said to be (point-)identified. However, it is known that the NPMLE in (7) may not be unique in multidimensional settings; in such cases, $\hat{G}$ refers to any maximizer. Thus, before presenting an algorithm to compute $\hat{G}$ in Section 6, we first establish conditions for the identification of G_* in Section 3, and then show in Section 4 that consistent estimation is possible despite multiple sources of heterogeneity.

3 Identification

This section establishes identification of the HC model defined in (2). Let $\mathcal{G}$ denote the class of all distributions on $\Theta \equiv \mathbb{R}^{d_\beta} \times \mathcal{K}_\delta$, defined as

$$\mathcal{G} := \mathcal{P}(\mathbb{R}^{d_\beta} \times \mathcal{K}_\delta) = \{G \in \mathcal{P}(\mathbb{R}^{d_\theta}) : \text{supp}_G(\delta_i) \subseteq \mathcal{K}_\delta\}, \tag{9}$$

where $\mathcal{K}_\delta \subseteq \mathbb{R}^{d_\delta}$ is a compact subset in which δ_i is supported, as in Assumption 1. Since any solution to the MLE in (6) must lie in the set

$$\{G \in \mathcal{G} : f_G(Y_i, X_i) = f_{G_*}(Y_i, X_i) \ \text{ a.s.}\}, \tag{10}$$

the identified set of distributions consists of distributions consistent with the observed marginal likelihood. We now state the conditions on the covariates and the parametric form of $P(\cdot)$ required for the identification of G_*.

Assumption 2. There exists a pair $(\boldsymbol{X}, \boldsymbol{M})$ of fixed matrices with $\boldsymbol{X} \in \text{supp}(X_i)$ and $\boldsymbol{M} \in \mathbb{R}^{T \times (T - d_\beta)}$ such that the following conditions hold:

(i) $\text{rank}(\boldsymbol{X}) = d_\beta \leq T - 1$, $\text{rank}(\boldsymbol{M}) = T - d_\beta \geq 1$, and $\boldsymbol{M}'\boldsymbol{X} = 0$.

(ii) For any $\delta, \tilde{\delta} \in \mathcal{K}_\delta$,

$$\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M} = \boldsymbol{M}'P(\tilde{\delta})P(\tilde{\delta})'\boldsymbol{M} \quad \text{if and only if} \quad \delta = \tilde{\delta}.$$

Assumption 2(i) requires $\boldsymbol{X} \in \text{supp}(X_i)$ to have full column rank, paralleling the familiar rank condition in regression. This ensures the existence of an annihilator matrix $\boldsymbol{M}$ with rank $T - d_\beta$.[4] We further require that $d_\beta \leq T - 1$ so that $\boldsymbol{M}$ has non-null rank and can be used in part (ii). This restriction is needed because, in our setting, the covariances Σ_i depend on unknown δ_i.

Assumption 2(ii) imposes that $\delta \mapsto \boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}$ is a one-to-one function of $\delta \in \mathcal{K}_\delta$, meaning that the variance parameters remain identifiable after removing the effect of $\boldsymbol{X}\beta_i$. The intuition is as follows. Since the concatenated matrix $(\boldsymbol{X}, \boldsymbol{M})$ is of full rank, the information in Y_i, conditional on $X_i = \boldsymbol{X}$, can be expressed as

$$\begin{bmatrix} \boldsymbol{M}'Y_i \\ \boldsymbol{X}'Y_i \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{X}'\boldsymbol{X}\beta_i \end{bmatrix} + \begin{bmatrix} \boldsymbol{M}'P_i e_i \\ \boldsymbol{X}'P_i e_i \end{bmatrix},$$

where $\boldsymbol{M}'Y_i$ is a Gaussian mixture with mean zero and variance $\boldsymbol{M}'P_iP_i'\boldsymbol{M} = \boldsymbol{M}'P(\delta_i)P(\delta_i)'\boldsymbol{M}$. The distribution of $\boldsymbol{M}'P(\delta_i)P(\delta_i)'\boldsymbol{M}$ can therefore be identified from the distribution of $\boldsymbol{M}'Y_i$. Then, the distribution of δ_i is identified in view of Assumption 2(ii).

[4]The choice of $\boldsymbol{M}$ is unique (up to a nonsingular linear transformation) given $\boldsymbol{X}$, since a different choice of $\boldsymbol{M}$ can be written as $\tilde{\boldsymbol{M}} = \boldsymbol{M}A$ for some nonsingular A.

To identify G_*, the remaining step is to recover the conditional distribution of β_i given δ_i. Let $\mathcal{E} \in \mathbb{R}^{T\times(T-d_\beta)}$. The key idea is to express

$$\boldsymbol{M}'Y_i + \mathcal{E}'\boldsymbol{X}'Y_i = (\boldsymbol{M}' + \mathcal{E}'\boldsymbol{X}')P(\delta_i)e_i + \mathcal{E}'(\boldsymbol{X}'\boldsymbol{X})\beta_i$$

as a perturbation of $\boldsymbol{M}'Y_i$ with noise $\mathcal{E}'\boldsymbol{X}'Y_i$ injected. We measure the shift in the mixing distribution of $\boldsymbol{M}'Y_i + \mathcal{E}'\boldsymbol{X}'Y_i$ relative to the mixing distribution of $\boldsymbol{M}'Y_i$ identified earlier. This comparison reveals the conditional distribution of $\mathcal{E}'(\boldsymbol{X}'\boldsymbol{X})\beta_i$ given δ_i as $\mathcal{E}$ traverses $\mathbb{R}^{T\times(T-d_\beta)}$, which then identifies β_i given δ_i via the Cramer-Wold device.

The following theorem formalizes this result. The proof of Theorem 1 as well as the proofs of all other theorems and propositions are given in Supplemental Appendix S-1.

Theorem 1 *Under Assumptions 1 and 2, G_* is identified in the class $\mathcal{G} = \mathcal{P}(\mathbb{R}^{d_\beta} \times \mathcal{K}_\delta)$.*

The primary challenge in identifying G_* is to distinguish the mean and variance components that are convolved into the outcome distribution. For instance, assume that $Y_i \sim \mathcal{N}(0,1)$ is a scalar observation. The observed distribution of Y_i can be rationalized by a continuum of models: $Y_i = a_i + \sigma_i e_i$, where $a_i \sim \mathcal{N}(0, v_a)$ and $\sigma_i \equiv \sqrt{1-v_a}$ for $v_a \in [0,1]$. Thus, one cannot tell which model generates the observed distribution without imposing further restrictions on a_i or σ_i. Bruni and Koch (1985) establish general identification results for Gaussian mixtures under a compact support assumption for both the mean and variance components.[5] Using a type of identification-at-infinity argument, they show that both components can be separated by leveraging variation in the tails of Y_i. However, this argument breaks down as soon as the mean component places a slight nonzero mass in the tails, producing observationally indistinguishable yet fundamentally different models, as illustrated above. In contrast, our identification strategy rests on the rank and one-to-one restrictions that preclude this type of identification failure.

Theorem 1 extends to models that include homogeneous slope parameters:

$$Y_i = X_i\beta_i + W_i\gamma_* + P_i e_i, \tag{11}$$

[5]The results of Bruni and Koch (1985) are also used in identification of latent distributions in labor economic analyses (see, e.g., Pastorino (2024); Bunting, Diegert, and Maurel (2024); de Paula, Gualdani, Pastorino, and Salgado (2025)), but the context is different from identification of mixture models in econometrics such as discussed in Compiani and Kitamura (2016).

where $W_i = (W_{i1}, \ldots, W_{iT})'$ is a $T \times d_\gamma$ matrix of additional controls associated with common slopes $\gamma_* \in \mathbb{R}^{d_\gamma}$. The common parameters γ_* can often be identified without knowledge of G_* by pooling cross-sectional information.

To see how γ_* can be identified, rewrite (11) as a pooled-effects model:

$$Y_i = X_i\beta_* + W_i\gamma_* + u_i,$$

where $\beta_* := \mathbb{E}_{G_*}[\beta_i]$ and $u_i = X_i(\beta_i - \beta_*) + P_i e_i$ denotes the composite error. The parameter γ_* is identified as the familiar regression coefficient

$$\gamma_* = \mathbb{E}[\tilde{W}_i' W_i]^{-1}\mathbb{E}[\tilde{W}_i' Y_i],$$

provided that $\mathbb{E}[\tilde{W}_i'\tilde{W}_i] > 0$, where $\tilde{W}_i := W_i - X_i\mathbb{E}[X_i'X_i]^{-1}\mathbb{E}[X_i'W_i]$. Once γ_* is identified, the model can be analyzed using the residuals $Y_i - W_i\gamma_*$. Theorem 1 serves as a general tool for analyzing identification in HC models. Before applying Theorem 1 to the HIVDX model, we first consider a HIVD model without the covariate X_{2i}. We verify Assumption 2 by demonstrating how to choose a pair $(\boldsymbol{X}, \boldsymbol{M})$.

3.1 Identification of the HIVD Model

We consider two variants of the HIVD model: one with AR(1) errors and the other with ARMA(1,1) errors.

3.1.1 AR(1) Errors

Consider the following AR(1) panel model without covariates:

$$\begin{aligned} Y_{it} &= a_i + u_{it}, \\ u_{it} &= \rho_i u_{it-1} + \sigma_i e_{it}, \quad t = 1, \ldots, T, \end{aligned} \tag{12}$$

where the initial condition u_{i0} is assumed to be drawn from the stationary distribution $\mathcal{N}(0, \sigma_i^2/(1-\rho_i^2))$. The parameters are collected in $\theta_i = (a_i, \sigma_i^2, \rho_i) \overset{\text{i.i.d.}}{\sim} G_*$.

Recall the form of the covariance matrix Σ_i given in (3). This yields the likelihood

function evaluated at a generic parameter $\theta = (a, \sigma^2, \rho)$:

$$\ell(Y_i \mid \theta) = \frac{(1-\rho^2)^{1/2}}{(2\pi\sigma^2)^{T/2}} \exp\left(-\frac{(1-\rho^2)(Y_{i1}-a)^2}{2\sigma^2} - \sum_{t=2}^{T} \frac{(Y_{it} - \rho Y_{it-1} - (1-\rho)a)^2}{2\sigma^2} \right). \tag{13}$$

Proposition 1 *Assume that $T \geq 3$. Then, G_* is identified in $\mathcal{G} = \mathcal{P}(\Theta)$, where $\Theta = \mathbb{R} \times \mathcal{K}_{(\sigma^2,\rho)}$ and $\mathcal{K}_{(\sigma^2,\rho)}$ is a compact subset of $(0,\infty) \times (-1,1)$.*

The restriction on $\mathcal{K}_{(\sigma^2,\rho)}$ implies that ρ_i is bounded away from $\{-1, 1\}$ and that σ_i^2 is bounded away from 0 and infinity. The condition $T \geq 3$ is not only sufficient but also necessary for the identification of the three-dimensional parameter θ_i. The variance parameters are identified from the covariance structure after purging the fixed effect a_i via first-differencing Y_i. Finally, the compact support for δ_i ensures that Y_i has non-degenerate and bounded covariance matrices across all δ_i; in particular, this condition rules out unit root processes.

We conclude this subsection by illustrating the choice of $(\boldsymbol{X}, \boldsymbol{M})$ for Assumption 2 in this example. For simplicity, consider $T = 3$. Let $\boldsymbol{X}$ be the column vector of ones whose rank is 1. Set $\boldsymbol{M}$ equal to the first-difference matrix:

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

This choice of $\boldsymbol{M}$ yields

$$\mathcal{V} := \boldsymbol{M}' P(\delta) P(\delta)' \boldsymbol{M} = \frac{\sigma^2}{1+\rho} \begin{pmatrix} 2 & -1+\rho \\ -1+\rho & 2 \end{pmatrix}.$$

Since $\rho = 1 + 2\mathcal{V}_{21}/\mathcal{V}_{11}$ and $\sigma^2 = \mathcal{V}_{11} + \mathcal{V}_{21}$, there exists a one-to-one relationship between $\mathcal{V}$ and δ, which ensures that δ is identified from $\boldsymbol{M}' P(\delta) P(\delta)' \boldsymbol{M}$.

3.1.2 ARMA(1,1) Errors

Consider the ARMA(1,1) panel model:

$$\begin{aligned} Y_{it} &= a_i + u_{it}, \\ u_{it} &= \rho_i u_{it-1} + \sigma_i(e_{it} + \varphi_i e_{it-1}), \quad t = 1, \ldots, T, \end{aligned}$$

where $e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and the initial condition u_{i0} is drawn from the stationary distribution. The additional random coefficient $\varphi_i \in [-1,1]$ captures heterogeneity in the MA component. Under the given assumptions, u_i is normally distributed with a conditional covariance matrix Σ_i, whose (t,s) entry is given by

$$\Sigma_{i,ts} = \frac{\sigma_i^2}{1-\rho_i^2} \times \begin{cases} 1+\varphi_i^2+2\varphi_i\rho_i & \text{if } t=s, \\ (\rho_i+\varphi_i)(1+\rho_i\varphi_i) & \text{if } |t-s|=1, \\ \rho_i^{|t-s|-1}(\rho_i+\varphi_i)(1+\rho_i\varphi_i) & \text{if } |t-s|\geq 2. \end{cases}$$

Let $G_*(\rho_i+\varphi_i=0)=0$ denote that the probability that $\rho_i+\varphi_i=0$ equals zero under G_*.

Proposition 2 *Let $\theta_i=(a_i,\sigma_i^2,\rho_i,\varphi_i)$ collect all heterogeneous coefficients with $\theta_i \overset{\text{i.i.d.}}{\sim} G_*$. Assume that $T\geq 4$ and $G_*(\rho_i+\varphi_i=0)=0$. Then G_* is identified in $\mathcal{G}=\mathcal{P}(\Theta)$, where $\Theta=\mathbb{R}\times\mathcal{K}_{(\sigma^2,\rho,\varphi)}$ and $\mathcal{K}_{(\sigma^2,\rho,\varphi)}$ is a compact subset of $(0,\infty)\times(-1,1)\times[-1,1]$.*

The condition $T\geq 4$ is necessary to observe the process over a horizon sufficient to distinguish the two sources of persistence, namely ρ_i and φ_i. The restriction $G_*(\rho_i+\varphi_i=0)=0$ ensures that the two dynamic components do not perfectly offset one another. We provide the specific choice of $(\boldsymbol{X},\boldsymbol{M})$ for Assumption 2 in the arXiv version of this paper (Song, Lee, and Ng, 2026).

3.2 Identification of the HIVDX Model

We now return to the HIVDX model given in (1). In this model, the parameters are partitioned as $\beta_i=(a_i,b_i)$ and $\delta_i=(\sigma_i^2,\rho_i)$, yielding the full parameter vector $\theta_i=(a_i,b_i,\sigma_i^2,\rho_i)$ with dimension $d_\theta=4$. Define $\Delta x_t := x_t - x_{t-1}$ as the first difference operator applied to the sequence $(x_t)_{t=1}^T$. Whether and how G_* can be identified depends critically on the specific time-series variation of the covariates. We illustrate this using two processes for X_{2it}. In the first case, the parameters of the HIVDX model are identified when $T=4$, whereas identification requires $T=5$ in the second case.

3.2.1 Case I

We start with the following proposition.

Proposition 3 *Assume that $T \geq 4$ and either $\Delta X_{2,i2} = \pm \Delta X_{2,i3} \neq 0$ or $\Delta X_{2,i2} \neq \Delta X_{2,i4}$ holds with positive probability. Then, G_* is identified in $\mathcal{G} = \mathcal{P}(\Theta)$, where $\Theta = \mathbb{R} \times \mathcal{K}_{(\sigma^2,\rho)}$ and $\mathcal{K}_{(\sigma^2,\rho)}$ is a compact subset of $(0,\infty) \times (-1,1)$.*

To illustrate, suppose that $X_{2,it} = X_{2,i1} + (t-1)$ is a linear trend with a random initial value. Then, we have $\Delta X_{2,it} \equiv 1$ for all t. We need to find an appropriate $(\boldsymbol{X}, \boldsymbol{M})$. Pick $X_{2,1}$ in $\text{supp}(X_{2,i1})$ and define

$$\boldsymbol{X} := \begin{pmatrix} 1 & X_{2,1} \\ 1 & X_{2,1}+1 \\ 1 & X_{2,1}+2 \\ 1 & X_{2,1}+3 \end{pmatrix}, \quad \boldsymbol{M} := \begin{pmatrix} 1 & 0 \\ -2 & 1 \\ 1 & -2 \\ 0 & 1 \end{pmatrix},$$

where $\boldsymbol{M}$ represents the second-difference operator in matrix form. It is clear that $\boldsymbol{X}$ has full column rank. Furthermore, because $\boldsymbol{M}$ annihilates both the intercept (column of ones) and $X_{2,it}$, the transformation $\boldsymbol{M}'Y_i$ eliminates the fixed effects β_i entirely. Let

$$\mathcal{V} := \boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M} \equiv \mathbb{V}(\boldsymbol{M}'Y_i \mid \delta_i = \delta, X_i = \boldsymbol{X}),$$

where $\boldsymbol{M}'Y_i \in \mathbb{R}^2$ collects the second differences of the errors, $(\Delta^2 u_{i3}, \Delta^2 u_{i4})$, conditional on $X_i = \boldsymbol{X}$. We obtain the following equations:

$$\begin{aligned}
\mathcal{V}_{11} &= \mathbb{V}(\Delta^2 u_{i3} \mid \delta_i = \delta) = \mathbb{V}(u_{i3} - 2u_{i2} + u_{i1} \mid \delta_i = \delta) = \frac{2(3-\rho)}{1+\rho}\sigma^2, \\
\mathcal{V}_{12} &= \text{Cov}(\Delta^2 u_{i3}, \Delta^2 u_{i4} \mid \delta_i = \delta) = \text{Cov}(u_{i3} - 2u_{i2} + u_{i1}, u_{i4} - 2u_{i3} + u_{i2} \mid \delta_i = \delta) \\
&= \frac{(3-\rho)(\rho-1)}{1+\rho}\sigma^2.
\end{aligned}$$

Hence, the system can be inverted:

$$(\sigma^2, \rho) = \left(\frac{\mathcal{V}_{11}}{2} \frac{1 + \mathcal{V}_{12}/\mathcal{V}_{11}}{1 - \mathcal{V}_{12}/\mathcal{V}_{11}}, \frac{2\mathcal{V}_{12}}{\mathcal{V}_{11}} + 1 \right).$$

This shows that the mapping $\delta \mapsto \mathcal{V}$ is one-to-one, thereby establishing Assumption 2(ii). However, as shown in the arXiv version of this paper (Song, Lee, and Ng, 2026), Proposition 3 does not identify Case II, which is discussed below.

3.2.2 Case II

Proposition 4 *Assume that $T \geq 5$ and that $\Delta X_{2,it}$ is not identically zero for some $t \in \{2, \ldots, T\}$. Then, G_* is identified in $\mathcal{G} = \mathcal{P}(\Theta)$, where $\Theta = \mathbb{R} \times \mathcal{K}_{(\sigma^2,\rho)}$ and $\mathcal{K}_{(\sigma^2,\rho)}$ is a compact subset of $(0, \infty) \times (-1, 1)$.*

An example of $X_{2,it}$ that satisfies the condition in the proposition is a one-time universal level shift: $X_{2,it} = \mathbf{1}\{t \geq 3\}$, as in a nationwide policy change. Then, the sequence of first differences is $(\Delta X_{2,i2}, \Delta X_{2,i3}, \Delta X_{2,i4}, \Delta X_{2,i5}) = (0, 1, 0, 0)$. Then, $\boldsymbol{M}$ can now be constructed as

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

which yields $\boldsymbol{M}'Y_i = (\Delta u_{i2}, \Delta u_{i4}, \Delta u_{i5})'$. Since the $(1,1)$ and $(2,3)$ elements of the covariance matrix $\mathcal{V} = \mathbb{V}(\boldsymbol{M}'Y_i \mid \delta_i = \delta)$ are given by

$$\begin{aligned} \mathcal{V}_{11} &= \mathbb{V}(\Delta u_{i2} \mid \delta_i = \delta) = \frac{2\sigma^2}{1+\rho}, \\ \mathcal{V}_{23} &= \mathrm{Cov}(\Delta u_{i4}, \Delta u_{i5} \mid \delta_i = \delta) = \frac{(\rho-1)\sigma^2}{1+\rho}, \end{aligned}$$

we can identify ρ via the ratio $2\mathcal{V}_{23}/\mathcal{V}_{11} + 1$, and subsequently recover σ^2. This establishes that the mapping $\delta \mapsto \mathcal{V}$ is one-to-one, as required by Assumption 2(ii). By Proposition 4, this result extends to the analysis of heterogeneous treatment effects implemented at a known time $s \geq 2$, giving

$$Y_{it} = a_i + b_i \, \mathbf{1}\{t \geq s\} + u_{it}, \quad t = 1, \ldots, 5.$$

Then the term $\mathbf{1}\{t \geq s\}$ represents a common treatment or policy, and b_i captures the individual-specific treatment effect. Identification holds provided the data cover both pre- and post-treatment periods.

As these examples illustrate, incorporating covariates necessitates a significantly more involved analysis. This is because the transformation matrix $\boldsymbol{M}$ forces reliance on differenced data, which often obscures the specific autocovariance structure re-

quired to identify δ_i. It is worth emphasizing that while the sufficient conditions in Proposition 4 are simple to verify, the underlying proof is combinatorially complex. Establishing this general result requires an exhaustive analysis of all distinct patterns of $\Delta X_{2,it}$ to verify the one-to-one mapping between δ and $\mathcal{V}$ in every possible scenario, as shown in the Appendix.

4 Consistency of the NPMLE

This section establishes that G_* can be consistently estimated by the NPMLE under Assumption 2 and the additional regularity assumptions stated below.

Assumption 3. (i) $\mathbb{E}[\|Y_i\|^2] < \infty$ and $\mathbb{E}[\|X_i\|^2] < \infty$.

(ii) $\mathbb{P}(\text{rank}(X_i) = d_\beta) = 1$.

Assumption 3(i) imposes standard moment conditions. Assumption 3(ii) requires full column rank of each X_i. This ensures that each individual likelihood distinguishes between different values of β_i, so that the individual MLEs are well-defined.

Our proof of the consistency of the NPMLE builds upon Kiefer and Wolfowitz (1956), who consider the case where $d_\beta = 1$. Since we are interested in cases where $d_\beta > 1$, we verify a multivariate version of Assumptions 1–5 in Kiefer and Wolfowitz (1956). Distinct from the issue of identification, the multivariate setting presents a topological complexity not found in the scalar case: probability mass can diverge along multiple directions. A primary technical challenge therefore lies in finding a metric that renders the space $\mathcal{G}$ totally bounded (thereby ensuring the compactness of its completion) while enabling the comparison of distributions with such diverging mass.

To this end, the vague topology is the natural choice. We adopt the canonical metrization of this topology, defined by

$$d(G_0, G_1) = \sum_{r=1}^{\infty} \frac{1}{2^r} \left| \int_\Theta h_r(\theta) dG_0(\theta) - \int_\Theta h_r(\theta) dG_1(\theta) \right|,$$

where $\{h_r\}_{r=1}^{\infty}$ is a dense sequence in the unit ball of $C_c(\Theta)$ with respect to the supremum norm $\|h\|_\infty = \sup_{\theta \in \Theta} |h(\theta)|$.[6] Since the test functions h_r vanish at infinity,

[6]Here, $C_c(\Theta)$ denotes the space of continuous real-valued functions on Θ with compact support.

this metric treats any mass that drifts to infinity as if it were collapsed to a single "point at infinity," irrespective of the direction of divergence. This choice of metric is compatible with the model because the likelihood cannot distinguish between different parameter values at infinity (i.e., the likelihood vanishes as parameters diverge). In effect, the metric compactifies the domain by collapsing all unbounded directions into a single point, mirroring the likelihood's asymptotic behavior. Finally, we note that convergence in the vague topology coincides with standard weak convergence whenever the limiting distribution is itself a probability measure.

The next theorem presents the almost sure consistency of the NPMLE with respect to the weak metric d.

Theorem 2 *Let Assumptions 1, 2, and 3 hold. Then,* $d(\hat{G}, G_*) \xrightarrow{\text{a.s.}} 0$ *as* $N \to \infty$.

Alternative metrics are discussed in the NPMLE literature, though they serve theoretical roles distinct from our metric d. The first is the Hellinger distance between the marginal densities, $\frac{1}{2}\int(\sqrt{f_{\hat{G}}(y)} - \sqrt{f_{G_*}(y)})^2 dy$. While Hellinger distance is frequently considered in mixture models (Jiang and Zhang, 2009; Soloff, Guntuboyina, and Sen, 2025), it measures proximity in the data space rather than the parameter space. In our setting, the convergence of the fitted marginal density $f_{\hat{G}}$ to the truth f_{G_*} is of secondary importance. Another alternative is the Wasserstein distance (e.g., W_2 as in Soloff, Guntuboyina, and Sen, 2025). However, W_2 requires the convergence of second moments and consequently becomes unbounded if probability mass escapes to infinity. This renders it unsuitable for our goal of establishing consistency through compactification. By contrast, our metric d allows us to successfully apply the Kiefer-Wolfowitz machinery. It strikes a necessary balance: it induces a topology sufficiently weak to ensure compactness (accommodating mass at infinity), yet strong enough to imply the consistency of downstream EB estimation.

The consistency of the NPMLE extends to the constrained NPMLE, defined as

$$\tilde{G} \in \operatorname*{arg\,max}_{G \in \tilde{\mathcal{G}}} F_N(G),$$

where $\tilde{\mathcal{G}} \subseteq \mathcal{G}$ is a subset that contains G_*. For instance, Gu and Koenker (2017) employ a profile likelihood approach to estimate a dynamic panel model similar to (12) under the restriction $\rho_i = \rho_*$. In our framework, imposing homogeneity ($\rho_i \equiv \rho$) corresponds to optimizing over a restricted subspace $\tilde{\mathcal{G}}$.

The estimator $\hat{G}$ allows for the recovery of various features of the data-generating process that can be expressed as continuous functionals. In particular, consistent estimators of prior and posterior moments can be obtained from the NPMLE under suitable regularity conditions. For example, the covariance matrix $\mathbb{V}_{G_*}(\theta_i)$ of the true prior can be estimated by $\mathbb{V}_{\hat{G}}(\theta_i)$, provided that $\|\theta\|^2$ is uniformly integrable with respect to the sequence of estimated priors. See the discussion following Assumption 4 for sufficient conditions ensuring such uniform integrability.

5 Regret Consistency and ℓ_p Convergence

Let $(\tau_i)_{i=1}^N = (\tau(\theta_i))_{i=1}^N$ denote the unit-level latent parameters of interest, each defined as a function of θ_i.[7] The estimation problem consists of constructing a compound decision rule $(Y_i, X_i)_{i=1}^N \mapsto (\hat{\tau}_i)_{i=1}^N$. G-modeling yields the estimator:

$$\hat{\tau}_i^{\mathrm{EB}} := \mathbb{E}_{\hat{G}}[\tau_i \mid Y_i, X_i] \equiv \frac{\int_\Theta \tau(\theta)\ell(Y_i \mid X_i, \theta)d\hat{G}(\theta)}{f_{\hat{G}}(Y_i, X_i)}.$$

This section establishes asymptotic guarantees for the EB method in terms of regret $R(\hat{\tau}^{\mathrm{EB}}, \tau) - R(\tau^*, \tau)$, where $R(\hat{\tau}, \tau) := \mathbb{E}_{G_*}[L(\hat{\tau}, \tau)]$ is compound risk. Specifically, we consider quadratic loss $L(\hat{\tau}, \tau) := \frac{1}{N}\sum_{i=1}^N (\hat{\tau}_i - \tau_i)^2$, and hence the *oracle* posterior mean is defined by:

$$\tau_i^* := \mathbb{E}_{G_*}[\tau_i \mid (Y_i, X_i)_{i=1}^N] = \mathbb{E}_{G_*}[\tau_i \mid Y_i, X_i], \quad i = 1, \ldots, N, \tag{14}$$

where the second equality follows from the independence across units. Write

$$R(\hat{\tau}^{\mathrm{EB}}, \tau) - R(\tau^*, \tau) = \mathbb{E}_{G_*}\left[\frac{1}{N}\sum_{i=1}^N (\hat{\tau}_i^{\mathrm{EB}} - \tau_i)^2 - (\tau_i^* - \tau_i)^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N (\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*)^2\right] \geq 0,$$

since $\mathbb{E}_{G_*}[(\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*)(\tau_i^* - \tau_i)] = 0$ by the properties of conditional expectation. The condition $\lim_{N\to\infty}[R(\hat{\tau}^{\mathrm{EB}}, \tau) - R(\tau^*, \tau)] = 0$ is referred to as *regret consistency* (see, e.g., Adusumilli, Gu, and Tao (2025)) or *asymptotic optimality* (see, e.g., Zhang (1997)).

Consider the case when τ_i is the prediction of Y_{iT+1}. The oracle predictor τ_i^* in

[7] We focus on the case of a scalar parameter here. However, the discussion and results in this section readily extend to the vector-valued case.

the HIVDX model is given by

$$\tau_i^* = \mathbb{E}_{G_*}[a_i + b_i X_{2,iT+1} + \rho_i Y_{iT} \mid Y_i, X_i, X_{2,iT+1}] - \mathbb{E}_{G_*}[\rho_i a_i + \rho_i b_i X_{2,iT} \mid Y_i, X_i]. \tag{15}$$

The EB one-period-ahead prediction $\hat{\tau}_i^{\mathrm{EB}}$ is then obtained by substituting the NPMLE for G_* in (15). Evaluating the EB method in this setting requires controlling the compound error arising from the interaction between the estimation discrepancy $\hat{\theta}_i^{\mathrm{EB}} - \theta_i^*$ and these observed variables. Consequently, the analysis necessitates controlling the ℓ_p distance (where $p \geq 2$, with $p = 4$ being a particularly important case) between the EB and oracle estimates.

Assumption 4. Assume that $\mathbb{E}_{G_*}[|\tau_i|^p] < \infty$ for some $p \in [2, \infty)$. Moreover, the following conditions hold.

(i) The function $\tau(\theta)$ is continuous in $\theta \in \Theta$ and, for all $c > 0$, satisfies

$$\lim_{\|\beta\| \to \infty} \exp(-c\|\beta\|^2)|\tau(\theta)| = 0.$$

(ii) For all $\epsilon > 0$, there exists $M < \infty$ such that

$$\limsup_{N \to \infty} \mathbb{E}\left[\int_\Theta |\tau(\theta)|^p \mathbf{1}\{|\tau(\theta)| \geq M\} d\hat{G}_N(\theta)\right] \leq \epsilon,$$

where $\hat{G}_N$ is the NPMLE defined in (7), with N denoting the sample size.

Assumption 4(i) requires that $\tau(\theta)$ depends continuously on θ and grows slower than any quadratic exponential function of $\|\beta\|$. This ensures that $|\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*| \to 0$ for each unit i as $d(\hat{G}_N, G_*) \to 0$. The continuity requirement can be relaxed to allow $\tau(\theta)$ to be discontinuous on a set where G_* places zero mass.

Assumption 4(ii) imposes a high-level condition on the NPMLE sequence, requiring $|\tau_i|^p$ to be uniformly integrable with respect to the sequence of estimated priors. This condition is automatically satisfied when τ_i has bounded support (e.g., $\tau_i = \delta_i$) but can be nontrivial for unbounded τ_i such as slope parameters. For such cases, Assumption 4(ii) requires specific moment conditions on (Y_i, X_i); for example, the p-th moment of β_i in the HIVDX model is bounded by the p-th moment of Y_i, provided X_i satisfies certain regularity conditions.

Theorem 3 *Let Assumptions 1, 2, 3, and 4 hold for some $p \in [2, \infty)$. Then $(\hat{\tau}_i^{\mathrm{EB}})_{i=1}^N$ is regret-consistent, i.e.,*

$$R(\hat{\tau}^{\mathrm{EB}}, \tau) - R(\tau^*, \tau) \to 0 \quad \text{as } N \to \infty.$$

Moreover, the EB estimators converge in ℓ_p-norm:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*|^p\right] \to 0 \quad \text{as } N \to \infty.$$

Theorem 3 provides the theoretical justification for using $(\hat{\tau}_i^{\mathrm{EB}})_{i=1}^N$ in large samples. The primary challenge in establishing this result is that the weak convergence of the priors, $d(\hat{G}, G_*) \xrightarrow{\text{a.s.}} 0$, implies the convergence of integrals only for *bounded* continuous functions. However, $\tau(\theta)$ and the squared-error loss are generally unbounded functions of the parameters. Assumption 4(ii) is critical because uniform integrability of the estimated prior moments ensures uniform integrability of the posterior moments across units. This allows us to upgrade the weak convergence of $\hat{G}$ to the convergence of the posterior expectations (the EB estimators) in ℓ_p-norm. Notably, the second result of the theorem, that is, convergence in ℓ_p-norm, is stronger than simple regret consistency ($p = 2$) and is particularly useful when the latent parameters enter the decision problem nonlinearly or interact with unbounded covariates.

5.1 Regret Consistency in the HIVDX Model

This subsection applies Theorem 3 to the HIVDX model, focusing on several quantities of interest: the individual-level parameters θ_i and the one-period-ahead prediction. These examples provide the theoretical underpinnings for the empirical analysis presented in Section 7.

When $\tau_i = \theta_i$ and $\theta_i = (\beta_i', \delta_i')'$ in the HIVDX model, the EB estimator of θ_i can be partitioned as $\hat{\theta}_i^{\mathrm{EB}} = (\hat{\beta}_i^{\mathrm{EB}}, \hat{\delta}_i^{\mathrm{EB}})$. Regret consistency of $\hat{\delta}_i^{\mathrm{EB}}$ follows immediately from Theorem 3 and Assumption 4, as δ_i lies in a compact set. For $\hat{\beta}_i^{\mathrm{EB}}$, however, Assumption 4(ii) requires $\|\beta_i\|^p$ to be uniformly integrable with respect to the sequence of estimated priors. This condition is guaranteed by a moment condition on (Y_i, X_i).

Once $(\theta_i)_{i=1}^N$ have been estimated, we turn to predicting future outcomes. We focus on the one-period-ahead EB prediction $\hat{Y}_{iT+1}^{\mathrm{EB}}$ of Y_{iT+1} (recall the oracle estimator given

in (15)). Let $\bar{X}_{2,i} = T^{-1}\sum_{t=1}^{T} X_{2,it}$ denote the time average of $X_{2,it}$ for unit i.

Proposition 5 *Let Assumption 1 hold. Assume that* $|\bar{X}_{2,i}| \leq M$ *and* $\sum_{t=1}^{T}(X_{2,it} - \bar{X}_{2,i})^2 \geq c$ *for some constants* $M, c > 0$ *and for all* $i \geq 1$.

(i) $(\hat{\sigma}_i^{2,\mathrm{EB}})_{i=1}^N$ *and* $(\hat{\rho}_i^{\mathrm{EB}})_{i=1}^N$ *are regret-consistent.*

(ii) In addition, assume that $\mathbb{E}[\|Y_i\|^{2+\varepsilon}] < \infty$ *for some* $\varepsilon > 0$. *Then* $(\hat{a}_i^{\mathrm{EB}})_{i=1}^N$ *and* $(\hat{b}_i^{\mathrm{EB}})_{i=1}^N$ *are regret-consistent.*

(iii) Assume further that $\mathbb{E}[\|Y_i\|^{4+\varepsilon}] < \infty$ *for some* $\varepsilon > 0$, $\mathbb{E}[|X_{2,iT}|^4] < \infty$, *and* $\mathbb{E}[|X_{2,iT+1}|^4] < \infty$. *Then the predictions* $(\hat{Y}_{iT+1}^{\mathrm{EB}})_{i=1}^N$ *are regret-consistent.*

Proposition 5(iii) requires fourth (and slightly higher) moments of $(X_{2,iT}, X_{2,iT+1})$ and Y_i to control the components of the prediction risk $\mathbb{E}[N^{-1}\sum_{i=1}^{N}(\hat{Y}_{iT+1}^{\mathrm{EB}} - \hat{Y}_{iT+1}^{*})^2]$. For instance, applying the Cauchy-Schwarz inequality to the component involving the slope coefficient yields

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}} - \hat{b}_i^{*})^2 X_{2,iT+1}^2\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}} - \hat{b}_i^{*})^4\right]}\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_{2,iT+1}^4\right]}.$$

The moment assumption on Y_i ensures that the first term on the right-hand side vanishes asymptotically (by Theorem 3 with $p = 4$), while the fourth moment of $X_{2,iT+1}$ ensures that the second term remains bounded.

6 Computational Algorithms

A variety of computational methods have been proposed for implementing the NPMLE. Although the optimization problem is convex, computation is challenging because the optimization variable G is an infinite-dimensional object, rendering direct implementation intractable. A common remedy is to approximate G by a discrete distribution supported on m atoms (with $m \in \mathbb{N}$), thereby reducing the problem to fitting an m-component mixture model. This approach is justified by the well-known result that a discrete NPMLE with at most N_0 support points exists (possibly among multiple maximizers), where N_0 is the number of distinct data points (Lindsay, 1983).[8]

[8] Shen and Wu (2025) show that the required number of components m grows at the rate $O(\log N)$ in the one-dimensional case under subgaussian priors, and conjecture that in higher dimensions m

Specifically, let $\boldsymbol{\theta} = (\theta^j)_{j=1}^m$ denote a grid of support points for G, and let $\boldsymbol{\omega} = (\omega^j)_{j=1}^m$ be the associated weights satisfying $\omega^j \geq 0$ for all j and $\sum_{j=1}^m \omega^j = 1$. Let $G^{\boldsymbol{\theta},\boldsymbol{\omega}}$ denote the discrete distribution induced by $(\boldsymbol{\theta},\boldsymbol{\omega})$, i.e.,

$$G^{\boldsymbol{\theta},\boldsymbol{\omega}}(\Theta') = \sum_{j=1}^m \omega^j \, \mathbf{1}[\theta^j \in \Theta'] \qquad \text{for all subsets } \Theta' \subseteq \Theta,$$

where we use the superscript j to denote the grid $(\theta^j)_{j=1}^m \in \Theta^m$, as distinct from the individual-specific parameters $(\theta_i)_{i=1}^N \in \Theta^N$. The NPMLE of the HC model is given by

$$\underset{\boldsymbol{\theta}\in\Theta^m,\boldsymbol{\omega}\geq 0}{\arg\max}\; F_N(G^{\boldsymbol{\theta},\boldsymbol{\omega}}) = \underset{\boldsymbol{\theta}\in\Theta^m,\boldsymbol{\omega}\geq 0}{\arg\max}\; \frac{1}{N}\sum_{i=1}^N \log f_{G^{\boldsymbol{\theta},\boldsymbol{\omega}}}(Y_i, X_i) \quad \text{s.t.} \;\; \sum_{j=1}^m \omega^j = 1,$$

where

$$f_{G^{\boldsymbol{\theta},\boldsymbol{\omega}}}(Y_i, X_i) = \int_\Theta \ell(Y_i \,|\, X_i, \theta)\, dG^{\boldsymbol{\theta},\boldsymbol{\omega}}(\theta) = \sum_{j=1}^m \ell(Y_i \,|\, X_i, \theta^j)\, \omega^j.$$

By the Karush-Kuhn-Tucker theorem, this problem is equivalent to maximizing the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\omega}) = \frac{1}{N}\sum_{i=1}^N \log\left(\sum_{j=1}^m \ell(Y_i \,|\, X_i, \theta^j)\omega^j\right) - \lambda\left(\sum_{j=1}^m \omega^j - 1\right). \tag{16}$$

The first-order conditions with respect to $\boldsymbol{\omega}$ imply that the optimal Lagrange multiplier is fixed at $\lambda = 1$.

Most EB estimators are solutions to the discretized problem described above. For example, Jiang and Zhang (2009) introduce a fixed-point EM algorithm that mirrors the standard EM routine and establish that it attains a global optimum up to an explicit error bound. Gu and Koenker (2017) estimate a dynamic panel model with two-dimensional heterogeneity using the interior-point algorithm of Koenker and Mizera (2014) to fit weights on a fixed grid. More recently, Kim, Carbonetto, Stephens, and Anitescu (2020), Zhang, Cui, Sen, and Toh (2024), and Wang, Ibrahim, and Mazumder (2025) propose alternative methods based on general-purpose convex programming to improve scalability.

increases on the order of $O((\log N)^C)$, where C depends on the dimension of θ_i. For one-dimensional Gaussian location mixtures, Polyanskiy and Sellke (2025) establish computational guarantees, including an algorithm that computes the exact support size.

A feature shared by all these methods is that the location of the support points is fixed, so optimization is performed solely with respect to the weights $\boldsymbol{\omega}$. Unlike in simple mixture models, determining where to position the grid points is not obvious when the dimension of θ exceeds two. For instance, Zhang, Cui, Sen, and Toh (2024) suggest using observed data points as the grid when the dimension is three or higher. The choice of grid size is also delicate: a grid dense enough to minimize approximation error becomes computationally intractable due to the curse of dimensionality, while a sparse grid risks significant discretization bias.

To overcome the limitations of fixed grids, we consider an algorithm that allows the support points to move adaptively. This necessitates a shift from optimizing weights on a static grid to optimizing the probability distribution G itself within the space of probability measures $\mathcal{P}(\Theta)$. In this infinite-dimensional setting, the standard gradient descent used in Euclidean spaces is replaced by *gradient flow*, where the specific trajectory depends critically on the geometric structure imposed on $\mathcal{P}(\Theta)$. Because $\mathcal{P}(\Theta)$ can have a complex geometry, different metrics can generate markedly different gradient-flow trajectories.

Consider the gradient flow induced by two commonly used metrics: the Fisher-Rao (FR) and the Wasserstein (W). The FR gradient flow updates G_t^{FR} according to

$$G_{t+\Delta t}^{\mathrm{FR}} - G_t^{\mathrm{FR}} \approx G_t^{\mathrm{FR}}\big(\alpha_t - \mathbb{E}_{G_t^{\mathrm{FR}}}[\alpha_t]\big),$$

where the reweighting strategy $\alpha_t : \Theta \to \mathbb{R}$ is chosen to maximize the likelihood gain $F_N(G_{t+\Delta t}^{\mathrm{FR}}) - F_N(G_t^{\mathrm{FR}})$ subject to the normalization $\mathbb{E}_{G_t^{\mathrm{FR}}}[(\alpha_t - \mathbb{E}_{G_t^{\mathrm{FR}}}[\alpha_t])^2] = 1$. Intuitively, the FR flow increases probability mass at locations with high fitness but cannot create mass at new locations or move existing support points. In contrast, the W gradient flow evolves according to the continuity (transport) equation

$$G_{t+\Delta t}^{\mathrm{W}} - G_t^{\mathrm{W}} \approx -\mathrm{div}(G_t^{\mathrm{W}} V_t),$$

where the velocity field $V_t : \Theta \to \mathbb{R}^{d_\theta}$ describes the movement of mass and is chosen to maximize the gain in F_N under the normalization $\mathbb{E}_{G_t^{\mathrm{W}}}[\|V_t\|^2] = 1$. This trajectory effectively moves the support points across the parameter space Θ.

Yan, Wang, and Rigollet (2024) show for the Gaussian location mixture case that the FR flow is guaranteed to reach a global maximum when initialized with full support, while W may be stuck at a local maximum. To also exploit the gains

from transporting mass across space, Yan, Wang, and Rigollet (2024) consider a discrete-time Wasserstein-Fisher-Rao (WFR) gradient flow algorithm that combines both reweighting of FR and transport dynamics of W when updating G_t^{WFR}:

$$G_{t+\Delta t}^{\mathrm{WFR}} - G_t^{\mathrm{WFR}} \approx -\mathrm{div}(G_t^{\mathrm{WFR}} V_t) + G_t^{\mathrm{WFR}}\big(\alpha_t - \mathbb{E}_{G_t^{\mathrm{WFR}}}[\alpha_t]\big).$$

The algorithm adaptively relocates support points so that m remains moderate and need not exceed the sample size N. It thus inherits the global convergence guarantees of the FR flow, and adaptivity of W through mass transport. Again in a Gaussian location mixture setting, The WFR is shown to be superior to the EM algorithm, and gradient-descent methods based solely on the FR or W flow. Motivated by these encouraging results, we extend the WFR from a Gaussian location mixture to the HC model.

6.1 WFR Algorithm for the HC Model

The WFR gradient-descent algorithm updates $(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n)_{n\geq 0}$ by alternately reweighting and moving the mass along the steepest descent path. Let $\eta \in (0, 1]$ denote the step size (learning rate), and define the posterior weights

$$\pi_{ij}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{\ell(Y_i \,|\, X_i, \theta^j)\, \omega^j}{\sum_{k=1}^m \ell(Y_i \,|\, X_i, \theta^k)\, \omega^k}, \qquad i = 1, \ldots, N,\ j = 1, \ldots, m,$$

which correspond to the E-step in the EM algorithm. Holding $\boldsymbol{\theta}_n$ fixed, the algorithm first updates the weights $\boldsymbol{\omega}_n$ via

$$\begin{aligned} \omega_{n+1}^j &= \omega_n^j \left(1 + \eta \left[\frac{1}{N}\sum_{i=1}^N \frac{\ell(Y_i \,|\, X_i, \theta_n^j)}{\sum_{k=1}^m \ell(Y_i \,|\, X_i, \theta_n^k)\, \omega_n^k} - 1\right]\right) \\ &= (1-\eta)\omega_n^j + \frac{\eta}{N}\sum_{i=1}^N \pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n), \qquad j = 1, \ldots, m, \end{aligned} \tag{FR-step}$$

which we refer to as the FR-step. The reweighting function is chosen as the gradient of the objective function in (16) with respect to ω^j, thereby ensuring descent. Next,

with $\boldsymbol{\omega}_{n+1}$ fixed, the grid $\boldsymbol{\theta}_n$ is updated according to

$$\begin{aligned}\tilde{\theta}^j_{n+1} &= \theta^j_n + \frac{\eta}{N}\sum_{i=1}^{N}\pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_{n+1})\frac{\partial}{\partial\theta^j_n}\log\ell(Y_i \mid X_i, \theta^j_n),\\ \theta^j_{n+1} &= P_\Theta(\tilde{\theta}^j_{n+1}), \qquad j = 1, \ldots, m,\end{aligned} \tag{W-step}$$

where $P_\Theta(\cdot)$ denotes the metric projection onto Θ, ensuring that $\theta^j_{n+1} \in \Theta$.[9] This W-step moves each grid point in the direction of the posterior-weighted average score, corresponding to a discrete version of the W gradient flow. Starting from $n = 0$, the FR- and W-steps are applied alternately until $n = \overline{n}$, the maximum number of iterations, or until the following convergence criterion is met: namely,

$$\max_j \left\{\frac{1}{N}\sum_{i=1}^{N}\pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n) - 1\right\} \leq \texttt{tol}$$

for some tolerance parameter $\texttt{tol} > 0$.[10] Let $\hat{n}$ denote the stopping index (either the first n satisfying the convergence criterion or $\overline{n}$ if it is not met). The resulting estimate is the NPMLE $\hat{G} = G^{\boldsymbol{\theta}_{\hat{n}}, \boldsymbol{\omega}_{\hat{n}}}$. Implementation of the HC model is summarized in Algorithm 1.

Starting values Choosing an initial grid that is well spread over Θ is essential for the WFR gradient-descent algorithm to converge to a global optimum (Yan, Wang, and Rigollet, 2024, Theorem 4). If, instead, all atoms are initialized at the same location, i.e., $\theta^j_0 = \tilde{\theta}$ for all $j = 1, \ldots, m$, then the subsequent updates satisfy $\theta^1_n = \cdots = \theta^m_n$ for every $n \geq 0$. In this case, the distribution $G^{\boldsymbol{\theta}_n, \boldsymbol{\omega}_n}$ degenerates to a single point mass at $\theta^j_n = \tilde{\theta}$, failing to capture any heterogeneity in the θ_i's and hence unable to recover the NPMLE.

In practice, we generate a dispersed initialization by computing MLEs on random subsamples of the data. The procedure is as follows. Fix the subsample size B. If the time dimension T is sufficient to ensure that individual MLEs are well-defined, one may choose $B = 1$ to maximize the diversity of the grid points. For $j = 1, \ldots, m$:

[9] Formally, for any $\tilde{\theta} \in \mathbb{R}^{d_\theta}$, the metric projection is $P_\Theta(\tilde{\theta}) := \arg\min_{\theta\in\Theta} \|\tilde{\theta} - \theta\|$. It yields the closest point in Θ to $\tilde{\theta}$ under the Euclidean norm. For example, if $\Theta = [a, b]^{d_\theta}$ is a box constraint, then $P_\Theta(\tilde{\theta})$ simply clips each coordinate of $\tilde{\theta}$ to lie within $[a, b]$. This ensures that each updated grid point θ^j_{n+1} remains feasible, even if the gradient step temporarily leaves the parameter space Θ.

[10] A necessary and sufficient condition for $\hat{G}$ to be NPMLE is $\frac{1}{N}\sum_{i=1}^{N}\frac{\ell(Y_i \mid X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \leq 1$ for all $\theta \in \Theta$.

Algorithm 1: Wasserstein-Fisher-Rao algorithm for the HC model

Input: Dataset $(Y_i, X_i)_{i=1}^N$; initial grid and weights $(\theta_0^j, \omega_0^j)_{j=1}^m$; step size $\eta \in (0,1]$; maximum iterations $\overline{n}$; tolerance parameter `tol`.

Set $n \leftarrow 0$.

while $n < \overline{n}$ and $\max_j\{\frac{1}{N}\sum_{i=1}^N \pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n) - 1\} >$ `tol` **do**

`/* FR-step: reweighting */`

for $j = 1, \ldots, m$ **do**

Compute posterior weights $\pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n) = \frac{\ell(Y_i|X_i,\theta_n^j)\,\omega_n^j}{\sum_{k=1}^m \ell(Y_i|X_i,\theta_n^k)\,\omega_n^k}, \quad i = 1, \ldots, N$. Update $\omega_{n+1}^j = (1-\eta)\,\omega_n^j + \frac{\eta}{N}\sum_{i=1}^N \pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_n)$.

end

`/* W-step: transport */`

for $j = 1, \ldots, m$ **do**

Compute updated posteriors (with new weights) $\pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_{n+1}) = \frac{\ell(Y_i|X_i,\theta_n^j)\,\omega_{n+1}^j}{\sum_{k=1}^m \ell(Y_i|X_i,\theta_n^k)\,\omega_{n+1}^k}, \quad i = 1, \ldots, N$.

Take a gradient step $\tilde{\theta}_{n+1}^j = \theta_n^j + \frac{\eta}{N}\sum_{i=1}^N \pi_{ij}(\boldsymbol{\theta}_n, \boldsymbol{\omega}_{n+1})\,\frac{\partial}{\partial \theta_n^j}\log \ell(Y_i \mid X_i, \theta_n^j)$.

Project to the feasible set $\theta_{n+1}^j \leftarrow P_\Theta(\tilde{\theta}_{n+1}^j)$.

end

Set $n \leftarrow n+1$.

end

Output: Final grid and weights $(\theta_{\hat{n}}^j, \omega_{\hat{n}}^j)_{j=1}^m$, where $\hat{n}$ denotes the stopping index (the first n satisfying the convergence criterion, or $\overline{n}$ if it is not met).

(i) Sample a set of indices $\mathcal{I}_j = \{i(1), \ldots, i(B)\}$ uniformly at random from $\{1, \ldots, N\}$.

(ii) Compute the subsample MLE using $\mathcal{I}_j$:

$$\theta_0^j = \arg\max_{\theta \in \Theta} \sum_{b=1}^B \log \ell(Y_{i(b)} \mid X_{i(b)}, \theta)$$

and assign a uniform weight $\omega_0^j = 1/m$.

The resulting collection $(\theta_0^j, \omega_0^j)_{j=1}^m$ serves as the initial grid and weights.

6.2 Comparison with the EM Algorithm

The EM algorithm is often used to find a (local) solution to mixture problems of the form (16). Let $D_{ij} = \mathbf{1}[\theta_i = \theta^j]$ denote an unobserved indicator of unit i's membership in component j. If the true D_{ij} were known, the "complete-data" maximum likelihood problem would be

$$\arg\max_{\boldsymbol{\theta}\in\Theta^m} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} D_{ij}\log \ell(Y_i \,|\, X_i, \theta^j) = \arg\max_{\boldsymbol{\theta}\in\Theta^m} \sum_{j=1}^{m}\sum_{i:D_{ij}=1} \log \ell(Y_i \,|\, X_i, \theta^j),$$

which can be optimized directly with respect to $\boldsymbol{\theta}$. However, this is infeasible because the true D_{ij} are unobserved. The EM algorithm circumvents this difficulty by replacing D_{ij} with its conditional expectation,

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\omega}}[D_{ij} \,|\, Y_i, X_i] = \frac{\ell(Y_i \,|\, X_i, \theta^j)\,\omega^j}{\sum_{k=1}^{m} \ell(Y_i \,|\, X_i, \theta^k)\,\omega^k} = \pi_{ij}(\boldsymbol{\theta},\boldsymbol{\omega}),$$

evaluated at the current iterate $(\boldsymbol{\theta},\boldsymbol{\omega})$. This step is referred to as the E-step. Let $(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}})$ denote the nth iterate of the support points and their weights. Given these values, the expected log-likelihood can be written as

$$Q_N(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}}) := \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N} \pi_{ij}(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}}) \log \ell(Y_i \,|\, X_i, \theta^j).$$

In the subsequent M-step, $(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}})$ are updated by solving

$$\begin{aligned}
\theta_{n+1}^{j,\mathrm{EM}} &= \arg\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \pi_{ij}(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}}) \log \ell(Y_i \,|\, X_i, \theta), \\
\omega_{n+1}^{j,\mathrm{EM}} &= \frac{1}{N}\sum_{i=1}^{N} \pi_{ij}(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}}), \qquad j = 1,\ldots,m.
\end{aligned}$$

The E- and M-steps are iterated until convergence, $(\boldsymbol{\theta}_n^{\mathrm{EM}},\boldsymbol{\omega}_n^{\mathrm{EM}}) \to (\hat{\boldsymbol{\theta}}^{\mathrm{EM}},\hat{\boldsymbol{\omega}}^{\mathrm{EM}})$, yielding the EM estimator $\hat{G}^{\mathrm{EM}} = G^{\hat{\boldsymbol{\theta}}^{\mathrm{EM}},\hat{\boldsymbol{\omega}}^{\mathrm{EM}}}$.

When η in our algorithm is set to 1, the FR-step coincides with the weight update in the EM algorithm. For smaller values of η, the weights are adjusted more conservatively, leading to more stable updates than in EM. Similarly, the W-step can be

viewed as moving $\boldsymbol{\theta}_n$ incrementally along the gradient of the Q_N-function. Because it is formulated as a gradient-descent procedure, the WFR algorithm enjoys stronger convergence guarantees than EM. By contrast, the EM algorithm generally requires more careful initialization to reach the global maximum (Balakrishnan, Wainwright, and Yu, 2017).

7 Empirical Analysis of Income Dynamics

There is a large body of literature seeking to understand heterogeneity in earnings. See Altonji, Hynsjö, and Vidangos (2023) and Browning and Ejrnaes (2013) for recent reviews of this literature. We highlight several closely related methodological contributions. Chamberlain and Hirano (1999) consider a parametric Bayesian approach that allow for heterogeneity in σ_i^2. Geweke and Keane (2000) also adopt a Bayesian framework with a finite mixture model for the composite errors $\sigma_i e_{it}$, while Hirano (2002) extends this framework using an infinite mixture model. Gu and Koenker (2017) propose an EB approach that was closest in spirit to ours. These four papers did not allow for random (b_i, ρ_i). Giacomini, Lee, and Sarpietro (2025) develop an individual-weight shrinkage estimator that optimizes unit-level (rather than aggregate) accuracy by exploiting each individual's own past history, with feasible weights motivated by minimax regret. Moon, Schorfheide, and Zhang (2023) consider random components in all $(a_i, b_i, \rho_i, \sigma_i^2)$ within a parametric Bayesian framework. Specifically, they adopt a spike-and-slab prior that accommodate either dense or sparse structures but assume independence across random components. In contrast, we adopt a nonparametric EB approach for the prior distribution and allow for dependence among the random components.

In this section, we re-examine the application in Gu and Koenker (2017, GK hereafter), which analyzes log real earnings data studied in Meghir and Pistaferri (2004). The extract consists of 938 individuals who have continuous earnings records from age 25 onward in the Panel Study of Income Dynamics (PSID) for the period 1968–1993. The panel is unbalanced with varying numbers of time periods T_i.[11] As in Gu and Koenker (2017), Y_{it} denotes residuals obtained from year-specific regressions of log

[11] The only minor difference between the current model and the original HIVDX model in (1) is that T_i may vary across i. The likelihood function can be easily modified to accommodate unbalanced panels.

real earnings on a vector of covariates. Though the HIVDX model motivates GK's analysis, they assume ρ_i is homogeneous and $b_i = 0$ in the application. Furthermore, their likelihood, expressed in terms of sufficient statistics, is conditioned on the initial observation Y_{i1}, which implicitly assumes that the random component is independent of Y_{i1}. Thus, their marginal likelihood depends on Y_{i1} only through $Y_{i2} - \rho_* Y_{i1}$.

Using GK's PSID sample extract and their definition of Y_{it} (residualized incomes), we estimate the HIVDX model with $\theta_i := (a_i, b_i, \sigma_i^2, \rho_i)$ and $X_{2,it} := \text{Exp}_{it}/10$, so that b_i is scaled up by a factor of 10, and $X_{2,it}$ is potential experience constructed as $\text{Exp}_{it} := \text{age} - \max\{\text{years of schooling}, 12\} - 6$.[12] Furthermore, our analysis differs from GK in three other respects. First, we model Y_{i1} under a stationarity assumption without assuming independence between Y_{i1} and θ_i. Second, the likelihood is directly evaluated from the observed data since sufficient statistics are not available in the presence of $X_{2,it}$ and ρ_i. Third, we fit the model and estimate the prior distribution of θ_i using the WFR algorithm described in Section 6.1. In empirical analyses, we used $\overline{n} = 100,000$ as the maximum number of iterations and did not implement the stopping criterion. The initialization algorithm with $B = 1$ and $m = 500$ and the step size of $\eta = 0.0005$ are used.

Table 1: First and Second Moments of the Estimated Prior $\hat{G}$

$\mathbb{V}_{\hat{G}}(\theta_i)$	θ_i			
	a_i	b_i	σ_i^2	ρ_i
a_i	0.2052	−0.1003	−0.0105	−0.0039
b_i	−0.1003	0.1007	−0.0104	0.0074
σ_i^2	−0.0105	−0.0104	0.0229	−0.0013
ρ_i	−0.0039	0.0074	−0.0013	0.1052
$\mathbb{E}_{\hat{G}}[\theta_i]$	0.0134	−0.0166	0.0787	0.4219

Notes: The first four rows form a 4×4 covariance matrix. The last row presents means.

Table 1 shows the first and second moments from the estimated prior $\hat{G}$. The marginal variances suggest that the coefficients are heterogeneous in all four dimen-

[12] An equivalent form used in the literature is $\text{Exp}_{it} = \min\{\text{age} - \text{years of schooling} - 6, \text{age} - 18\}$.

sions. The variance of a_i is substantial, especially relative to the variance of Y_{it} across all units and time (0.22). Regarding b_i, the variance of $b_i X_{2,it}$ when $X_{2,it} = x$ is approximately $0.1x^2$. The value of x ranges over $[0.2, 3.2]$ within the sample, indicating notable heterogeneity in experience profiles. The variance of σ_i^2 is 0.0229 and is non-negligible. Finally, the distribution of ρ_i is also spread out, with approximately 73% of probability mass lying in $[0.2, 0.8]$.

Turning to the covariances of the random coefficients, the covariance between individual intercepts (a_i) and variances (σ_i^2) is negative, as in GK. However, allowing for additional sources of heterogeneity leads to some new findings. We find a negative correlation between σ_i^2 and b_i, and weak associations between ρ_i with the other parameters (a_i, b_i, σ_i^2). The most interesting result is the strong negative correlation between the individual intercepts (a_i) and slopes (b_i).

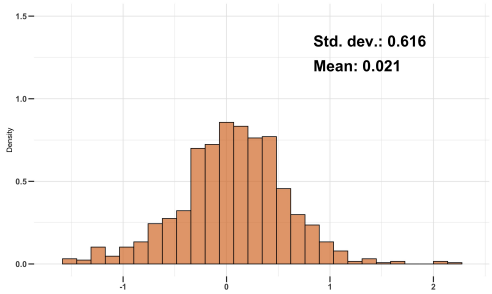
A side-by-side comparison between the individual MLEs and EB estimates of θ_i is presented in Figure 1. As we work with residual earnings, it is not surprising that the means of the individual MLEs and EB estimates of a_i and b_i are close to zero. A notable feature is that the individual MLEs of ρ_i are highly dispersed, with a substantial fraction taking negative values, whereas the EB estimates are mostly positive and considerably more concentrated around their mean value of 0.422. The cross-sectional variances of $\hat{a}_i^{\text{EB}}$, $\hat{b}_i^{\text{EB}}$, and $\hat{\rho}_i^{\text{EB}}$ are reduced by approximately 63%, 72%, and 48%, respectively, relative to the corresponding individual MLEs. While no conspicuous shrinkage is observed for $\hat{\sigma}_i^{2,\text{EB}}$, these reductions suggest substantial compound mean squared error gains from the EB approach.

Lastly, we consider one-period-ahead predictions of $(Y_{i,T_i})_{i=1}^N$ using a sample that holds out final-period observations. Figure 2 compares the resulting predictions and prediction errors from the individual MLE and the EB methods. The one-period-ahead prediction of Y_{i,T_i} is computed as

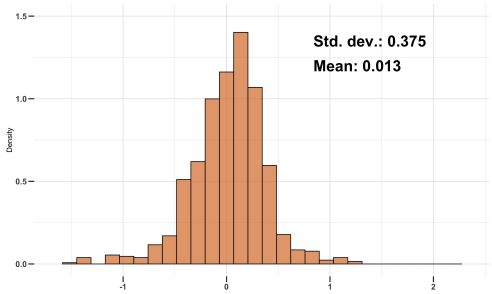
$$\hat{Y}_{i,T_i} = \hat{a}_i + X_{2,iT_i}\hat{b}_i + \hat{\rho}_i Y_{iT_i-1} - \widehat{\rho_i a_i} - X_{2,iT_i-1}\widehat{\rho_i b_i}, \tag{17}$$

where the hats indicate either EB or individual-level MLE estimates. As shown in Figure 1, the EB method exhibits substantial shrinkage, reducing the variance of the predicted values by approximately 25% compared to the MLE. This translates into improved forecasting performance, leading to a variance reduction of approximately 20% in the prediction errors.

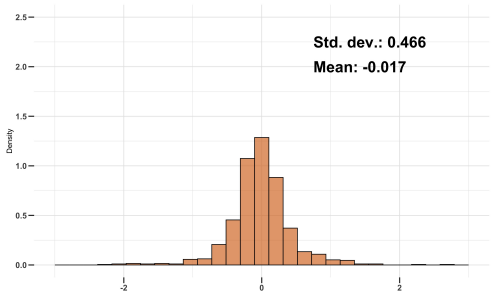
Figure 1: Shrinkage Effects: Comparison of $(\hat{\theta}_i^{\text{MLE}})_{i=1}^N$ and $(\hat{\theta}_i^{\text{EB}})_{i=1}^N$



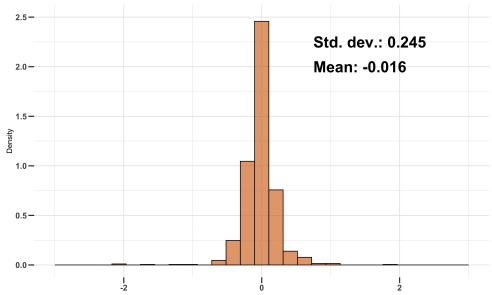
(a) MLEs of a_i.



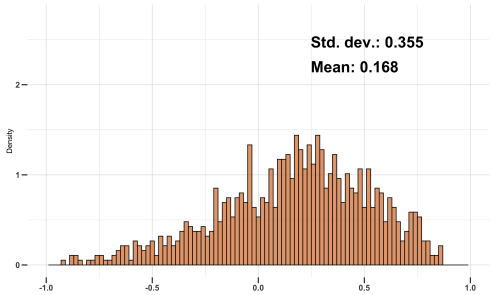
(b) EB estimates of a_i.



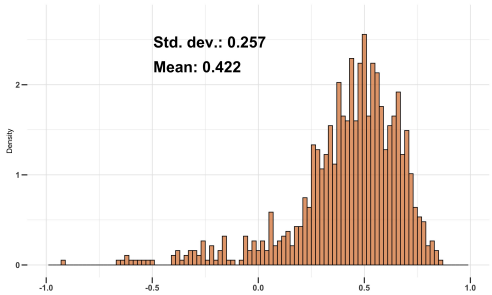
(c) MLEs of b_i.



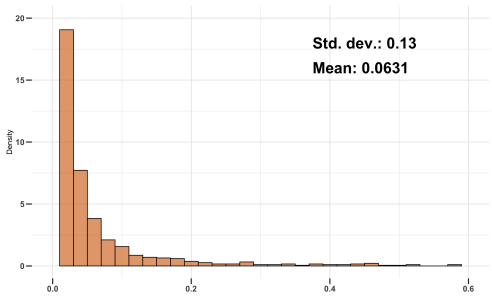
(d) EB estimates of b_i.



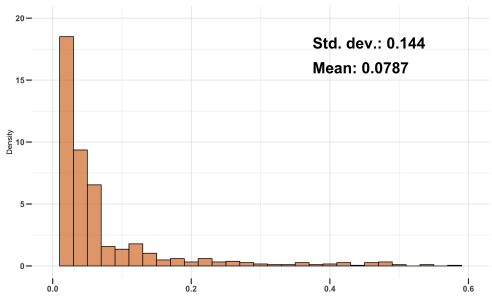
(e) MLEs of ρ_i.



(f) EB estimates of ρ_i.



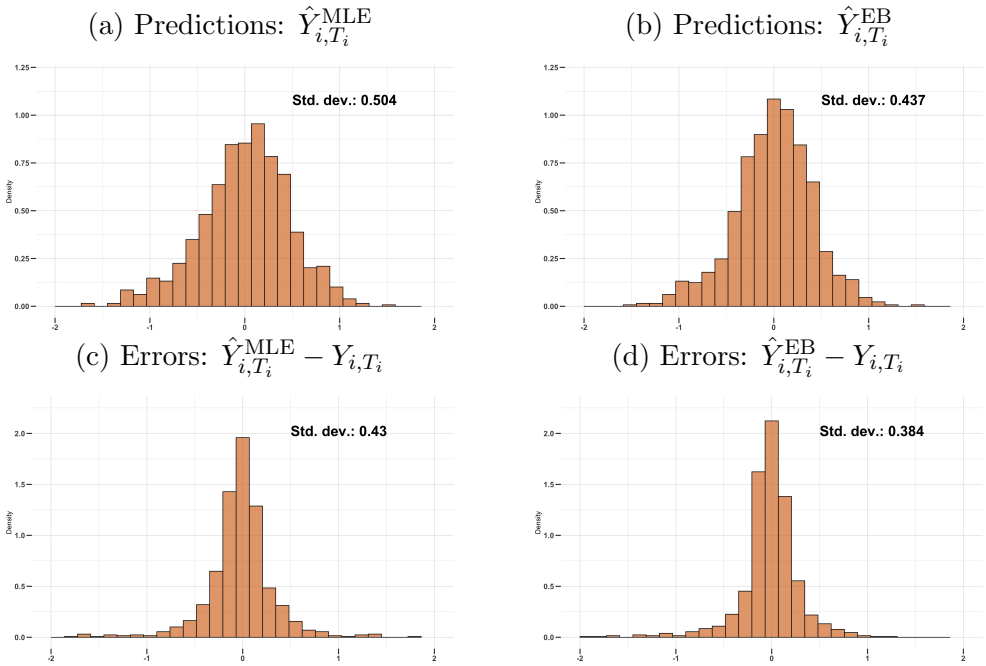
(g) MLEs of σ_i^2.



(h) EB estimates of σ_i^2.

Note: The left and right panels display the histograms of the individual MLE and EB estimates for each component of $(\theta_i)_{i=1}^N$.

Figure 2: Pseudo-out-of-sample predictions and prediction errors

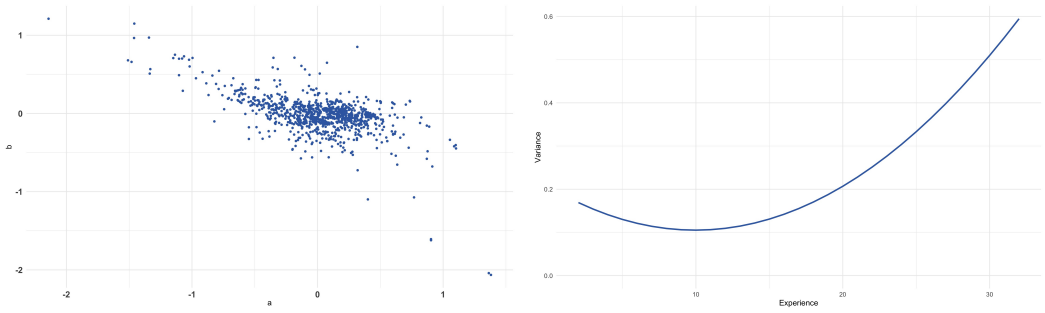


Notes: The reported standard deviations of the prediction errors are computed after excluding a single outlier observation whose prediction error falls below -4.5 under both methods; if this observation is included, the standard deviations for the MLE and EB errors are 0.456 and 0.411, respectively.

As shown in Table 1, a notable feature is the negative correlation between a_i and b_i. To further examine this relationship, the left panel of Figure 3 plots the EB estimate pairs $(\hat{a}_i^{\text{EB}}, \hat{b}_i^{\text{EB}})$ for $N = 938$ individuals, revealing a clear negative association between the individual intercepts $\hat{a}_i^{\text{EB}}$ and slopes $\hat{b}_i^{\text{EB}}$.[13] To explore the experience profile in the cross-section, we use the estimated moments in Table 1 to compute the cross-sectional prior variance of $a_i + b_i x$, which captures the deviation of an individual's income trajectory at $X_{2,it} = x$ from the average. The mapping $x \mapsto \mathbb{V}_{\hat{G}}(a_i + b_i x)$, shown in the right panel of Figure 3, indicates that the variance declines up to about $x = 1$ (ten years of experience) and increases thereafter. This non-monotonic pattern arises from the negative covariance between a_i and b_i. Interestingly, a similar U-shaped relationship between the log variance of residual earnings and experience is documented in Mincer (1974, Chapter 6, p.104), who attributes it to "a weak correlation between post-school investment ratios and earning capacity."

[13]For completeness, Figure 4 in Appendix O-2 of the arXiv version of this paper (Song, Lee, and Ng, 2026) presents all pairwise scatter plots of the EB estimates for θ_i.

Figure 3: EB Estimates of (a_i, b_i) and Cross-Sectional Variance of Earnings



(a) Pairs $(\hat{a}_i^{\text{EB}}, \hat{b}_i^{\text{EB}})$. (b) Variance profile by experience.

Note: The left panel plots the EB estimates $(\hat{a}_i^{\text{EB}}, \hat{b}_i^{\text{EB}})_{i=1}^N$. The right panel displays the cross-sectional variance of earnings as a function of years of experience, $\text{Exp} \mapsto \mathbb{V}_{\hat{G}}(a_i + b_i \text{Exp})$.

For completeness, we also computed estimates using the EM algorithm discussed in Section 6.2, initializing it with the same starting values used for WFR (as described in Section 6.1). We find that the EM estimates are very similar to the WFR solution. While the EM algorithm attains a slightly higher likelihood value by optimizing weights over the dense initial grid, the WFR algorithm yields an approximate solution with fewer mass points. Furthermore, although the EM algorithm generally requires fewer iterations to converge, the computational cost per iteration is higher. These results are encouraging, as they suggest WFR is well-suited for higher-dimensional problems where fixed-grid algorithms become computationally challenging.

8 Monte Carlo Experiments

We conduct Monte Carlo experiments to evaluate the performance of the EB estimator in dynamic panel models relative to (i) the oracle decision rule and (ii) the individual MLE.

The following DGPs are considered.

1. The HIVD model with $\theta_i = (a_i, \sigma_i^2, \rho_i)$. The individual intercepts $(a_i)_{i=1}^N$ are drawn from the Gamma$(1/2, \sqrt{2})$ distribution independently of (σ_i^2, ρ_i). The joint distribution of (σ_i^2, ρ_i) assigns probability mass $1/6$ to $(\sigma_i^2, \rho_i) = (0.1, 0.8)$ and $(0.3, 0.2)$, and probability mass $1/3$ to $(\sigma_i^2, \rho_i) = (0.1, 0.2)$ and $(0.3, 0.8)$.

This design features a negative correlation of -0.33 between ρ_i and σ_i^2. The initial u_{i0} is drawn from the stationary distribution $\mathcal{N}(0, \sigma_i^2/(1-\rho_i^2))$. The sample size for estimation is $(N, T) = (1000, 5)$. An additional observation Y_{iT+1} is generated for each unit to compute prediction errors.

2. The Restricted HIVD model follows the HIVD model above but imposes $\rho_i \equiv 0.5$ and assumes σ_i^2 has a two-point distribution satisfying $G_*(\sigma_i^2 = 0.1) = G_*(\sigma_i^2 = 0.3) = 0.5$. The initial distribution of u_{i0} is $\mathcal{N}(0, \sigma_i^2/(1-0.5^2))$. The distribution of a_i and the sample size is the same as in HIVD. This DGP mimics a scenario in which a researcher does not know that ρ_i is a constant and continue to estimate the HIVD model.

3. The HIDVX model generates $\theta_i = (a_i, b_i, \sigma_i^2, \rho_i)$ from the prior distribution estimated from the PSID sample. The covariate (interpreted as experience divided by 10) is assumed to evolve according to $X_{2,it} = X_{2,i,t-1} + 0.1$, with $(N, T) = (1000, 13)$, approximately as in the PSID sample. The initial value $X_{2,i1}$ is drawn from a discrete distribution with probabilities $\mathbb{P}(X_{2,i1} = 0.3) = 0.25$, $\mathbb{P}(X_{2,i1} = 0.4) = 0.05$, $\mathbb{P}(X_{2,i1} = 0.5) = \mathbb{P}(X_{2,i1} = 0.6) = 0.1$, and $\mathbb{P}(X_{2,i1} = 0.7) = 0.5$.

For the first two models, the WFR parameters are $\eta = 0.1$, $\overline{n} = 2000$, and $\texttt{tol} = 10^{-4}$. For the more complex HIVDX, $\eta = 0.005$, $\overline{n} = 10000$ without applying the early stopping criterion to ensure stable and reliable convergence of the algorithm.

Let $\hat{\varepsilon}_i = \hat{\tau}_i - \tau_i$ denote the estimation error for unit i, where τ_i is one of the model parameters, or the one step ahead prediction. For each method and replication, we compute

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i, \quad \text{SD} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\varepsilon}_i - \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i\right)^2\right)^{1/2}, \quad \text{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i^2\right)^{1/2}.$$

Additionally, a measure of predictability, denoted by R^2, is computed as one minus the ratio of the mean squared prediction error to the sample variance of Y_{iT+1}.

Table 2 reports results for the HIVD model and its restricted variant. In both designs, the oracle estimator achieves the lowest estimation error, but the EB estimator improves upon the individual MLE with uniformly smaller RMSE across all parameters. The gains are particularly pronounced for ρ_i. In line with this theoretical

prediction, the gains from EB are even larger in the restricted HIVD design, where the prior for ρ_i is more informative. Most of the RMSEs of the EB estimator are within 10% of those of the oracle.

Table 3 presents the results for the HIVDX model. Similarly to Table 2, the EB estimator outperforms the individual MLE across all parameters. For both a_i and b_i, the EB method achieves MSEs that are approximately 55% of those of the MLE. The EB estimator substantially reduces the MSE for ρ_i as well, primarily attributable to bias reduction. It is interesting to note that the Monte Carlo results for the bias of the MLE of ρ_i echo the pattern observed in the empirical application in Section 7. In terms of out-of-sample prediction accuracy, R^2 increases from 0.591 for the MLE to 0.662 for EB.

Finally, Table 4 shows the prior moment estimation results in the HIVDX design. The first moments of θ_i are generally well estimated, with the exception of a modest downward bias in the estimate of $\mathbb{E}_{G_*}[\rho_i]$. The variances of θ_i tend to be somewhat overestimated, which is expected given that estimation noise generally inflates the estimated variance of θ_i relative to the true variance. The negative covariance between a_i and b_i is estimated with a downward bias but the sign is reliably recovered.

9 Conclusions

We have proposed a nonparametric empirical Bayes framework for analyzing short-panel data with rich forms of unobserved heterogeneity. The analysis is a significant generalization of classical G-modeling to accommodate heterogeneous slopes and non-spherical error structures. Identification, consistency, and optimality results are established under general conditions, with primitive sufficient conditions derived for specific cases of interest. However, we have not developed any accompanying inference methods. Recent contributions such as Armstrong, Kolesár, and Plagborg-Møller (2022) and Ignatiadis and Wager (2022) can be useful for future research.

References

Adusumilli, K., J. Gu, and J. Tao (2025): "Empirical Bayes for Compound Adaptive Experiments," Available at `https://sites.google.com/site/adusumik/research`.

Table 2: Errors from 500 Monte Carlo Replications

Errors	DGP	Metric	Estimator		
			Oracle	MLE	EB
$\hat{a}_i - a_i$	HIVD	Bias	−0.004	0.000	−0.002
		SD	0.378	0.449	0.414
		RMSE	0.378	0.449	0.414
	HIVDR	Bias	0.001	0.000	−0.003
		SD	0.303	0.346	0.339
		RMSE	0.303	0.346	0.339
$\hat{\sigma}_i^2 - \sigma_i^2$	HIVD	Bias	0.000	−0.077	−0.006
		SD	0.078	0.117	0.085
		RMSE	0.078	0.141	0.086
	HIVDR	Bias	0.000	−0.077	−0.007
		SD	0.079	0.118	0.085
		RMSE	0.079	0.141	0.086
$\hat{\rho}_i - \rho_i$	HIVD	Bias	−0.003	−0.552	−0.062
		SD	0.273	0.482	0.323
		RMSE	0.273	0.733	0.330
	HIVDR	Bias	0.000	−0.546	−0.048
		SD	0.000	0.447	0.174
		RMSE	0.000	0.706	0.182
$\hat{Y}_{iT+1} - Y_{iT+1}$	HIVD	Bias	−0.001	0.000	0.000
		SD	0.485	0.560	0.490
		RMSE	0.485	0.560	0.490
		R^2	0.825	0.766	0.821
	HIVDR	Bias	−0.001	0.000	−0.001
		SD	0.471	0.562	0.477
		RMSE	0.471	0.562	0.477
		R^2	0.824	0.749	0.819

Notes: Each row reports averages computed from 500 replications. Oracle refers to the true posterior mean, and MLE corresponds to the individual MLE.

Table 3: Errors from 500 Monte Carlo Replications (HIVDX)

		Estimator		
Errors	Metric	Oracle	MLE	EB
$\hat{a}_i - a_i$	Bias	0.001	−0.001	−0.001
	SD	0.244	0.426	0.318
	RMSE	0.244	0.426	0.318
$\hat{b}_i - b_i$	Bias	−0.001	0.001	0.001
	SD	0.188	0.338	0.251
	RMSE	0.188	0.338	0.251
$\hat{\sigma}_i^2 - \sigma_i^2$	Bias	0.000	−0.020	−0.006
	SD	0.047	0.066	0.064
	RMSE	0.047	0.069	0.064
$\hat{\rho}_i - \rho_i$	Bias	−0.002	−0.326	−0.082
	SD	0.212	0.308	0.275
	RMSE	0.212	0.448	0.288
$\hat{Y}_{iT+1} - Y_{iT+1}$	Bias	0.000	0.001	0.001
	SD	0.299	0.341	0.314
	RMSE	0.299	0.341	0.314
	R^2	0.692	0.591	0.662

Notes: Each row reports averages computed from 500 replications. Oracle refers to the true posterior mean, and MLE corresponds to the individual MLE.

Table 4: Monte Carlo Results for Moment Estimates of G_* (HIVDX)

Panel A. Means				
	$\mathbb{E}_{\hat{G}}[a_i]$	$\mathbb{E}_{\hat{G}}[b_i]$	$\mathbb{E}_{\hat{G}}[\sigma_i^2]$	$\mathbb{E}_{\hat{G}}[\rho_i]$
Truth	0.013	−0.017	0.079	0.422
Bias	−0.001	0.001	−0.006	−0.082
SD	0.022	0.017	0.005	0.016
RMSE	0.023	0.018	0.011	0.080

Panel B. Variances				
	$\mathbb{V}_{\hat{G}}(a_i)$	$\mathbb{V}_{\hat{G}}(b_i)$	$\mathbb{V}_{\hat{G}}(\sigma_i^2)$	$\mathbb{V}_{\hat{G}}(\rho_i)$
Truth	0.205	0.101	0.023	0.105
Bias	0.054	0.030	−0.007	0.017
SD	0.037	0.025	0.006	0.009
RMSE	0.065	0.039	0.009	0.020

Panel C. Pairwise Covariances						
	(a_i, b_i)	(a_i, σ_i^2)	(a_i, ρ_i)	(b_i, σ_i^2)	(b_i, ρ_i)	(ρ_i, σ_i^2)
Truth	−0.100	−0.011	−0.004	−0.010	0.007	−0.001
Bias	−0.038	−0.003	0.005	0.007	−0.004	0.003
SD	0.028	0.011	0.009	0.010	0.006	0.002
RMSE	0.047	0.010	0.011	0.012	0.008	0.003

Notes: The row labeled 'Truth' reports the true values of the corresponding moments. Results are based on 500 replications.

Altonji, J. G., D. M. Hynsjö, and I. Vidangos (2023): "Individual earnings and family income: Dynamics and distribution," *Review of Economic Dynamics*, 49, 225–250.

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2022): "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 90(6), 2567–2602.

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017): "Statistical Guarantees for the EM Algorithm: From Population to Sample-Based Analysis," *Annals of Statistics*, 45(1).

Browning, M., and M. Ejrnaes (2013): "Heterogeneit in the Dynamics of Labor Earnings," *Annual Review of Economics*, 5, 219–245.

Bruni, C., and G. Koch (1985): "Identifiability of continuous mixtures of unknown Gaussian distributions," *Annals of Probability*, pp. 1341–1357.

Bunting, J., P. Diegert, and A. Maurel (2024): "Heterogeneity, Uncertainty and Learning: Semiparametric Identification and Estimation," arXiv:2402.08575 [econ.EM], available at `https://arxiv.org/abs/2402.08575`.

Chamberlain, G., and K. Hirano (1999): "Predictive Distributions Based on Longitudinal Earnings Data," *Annales d'Économie et de Statistique*, (55/56), 211–242.

Chen, J. (2024): "Empirical Bayes When Estimation Precision Predicts Parameters," arXiv:2212.14444 [econ.EM], available at `https://arxiv.org/abs/2212.14444`.

Cheng, X., S. C. Ho, and F. Schorfheide (2025): "Optimal Estimation of Two-Way Effects under Limited Mobility," Available at `https://arxiv.org/abs/2506.21987` (arXiv.2506.21987 [econ.EM]).

Compiani, G., and Y. Kitamura (2016): "Using Mixtures in Econometric Models: A Brief Review and Some New Results," *Econometrics Journal*, 19(3), C95–C127.

de Paula, A., C. Gualdani, E. Pastorino, and S. Salgado (2025): "On the Identification of Models of Uncertainty, Learning, and Human Capital Acquisition with Sorting," working paper available at `https://www.cristinagualdani.com/`.

Efron, B. (2014): "Two Modeling Strategies for Empirical Bayes Estimation," *Statistical Science*, 29:2, 285–301.

Efron, B., and C. Morris (1973): "Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach," *Journal of the American Statistical Association*, 68(341), 117–130.

Fan, Z., L. Guan, Y. Shen, and Y. Wu (2023): "Gradient flows for empirical Bayes in high-dimensional linear models," arXiv:2312.12708 [math.ST], available at `https://arxiv.org/abs/2312.12708`.

Gaillac, C. (2024): "Predicting Unobserved Individual-level Causal Effects," Working paper, accessed June 19, 2025, available at `www.cgaillac.com`.

Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnigns Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96, 293–356.

Giacomini, R., S. Lee, and S. Sarpietro (2025): "Individual Shrinkage for Random Effects," Available at `https://arxiv.org/abs/2308.01596` (arXiv:2308.01596 [econ.EM]).

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

Hirano, K. (2002): "Semiparametric Estimation of Regression Models for Panel Data," *Econometrica*, 70, 781–799.

Ignatiadis, N., and B. Sen (2025): "Empirical Partially Bayes Multiple Testing and Compound χ^2 Decisions," *Annals of Statistics*, 53(1), 1–36.

Ignatiadis, N., and S. Wager (2022): "Confidence intervals for nonparametric empirical Bayes analysis," *Journal of the American Statistical Association*, 117(539), 1149–1166.

James, W., and C. Stein (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379, Berkeley, CA. University of California Press.

Jiang, H., and A. Guntuboyina (2025): "A Nonparametric Maximum Likelihood Approach to Mixture of Regression," Available at https://arxiv.org/abs/2108.09816 (arXiv:2108.09816 [stat.ME]).

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

Kallenberg, O. (2021): *Foundations of Modern Probability*, vol. 99. Springer International Publishing, Cham.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics*, 27(4), 887–906.

Kim, Y., P. Carbonetto, M. Stephens, and M. Anitescu (2020): "A Fast Algorithm for Maximum Likelihood Estimation of Mixture Proportions Using Sequential Quadratic Programming," *Journal of Computational and Graphical Statistics*, 29(2), 261–273.

Koenker, R., and J. Gu (2024): "Empirical Bayes for the Reluctant Frequentist," arXiv:2404.03422 [stat.ME], available at https://arxiv.org/abs/2404.03422.

——— (2025): *Empirical Bayes: Tools, Rules and Duals*, Econometric Society Monograph Series. Cambridge University Press, Cambridge.

Koenker, R., and I. Mizera (2014): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *Journal of the American Statistical Association*, 109(506).

Kwon, S. (2025): "Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models," Available at https://arxiv.org/abs/2308.12485 (arXiv:2308.12485 [econ.EM]).

Lindsay, B. G. (1983): "The geometry of mixture likelihoods: a general theory," *Annals of Statistics*, pp. 86–94.

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

Meghir, C., and L. Pistaferri (2004): "Income variance dynamics and heterogeneity," *Econometrica*, 72(1), 1–32.

Mincer, J. A. (1974): *Schooling, Experience, and Earnings*. National Bureau of Economic Research, available at https://www.nber.org/books-and-chapters/schooling-experience-and-earnings.

Moon, H. R., F. Schorfheide, and B. Zhang (2023): "Bayesian Estimation of Panel Models Under Potentially Sparse Heterogeneity," arXiv:2310.13785 [econ.EM], available at https://arxiv.org/abs/2310.13785.

Mukherjee, S., B. Sen, and S. Sen (2023): "A Mean Field Approach to Empirical Bayes Estimation in High-dimensional Linear Regression," arXiv:2309.16843 [math.ST], available at https://arxiv.org/abs/2309.16843.

Pastorino, E. (2024): "Careers in Firms: The Role of Learning about Ability and Human Capital Acquisition," *Journal of Political Economy*, 132(6), 1994–2073.

Polyanskiy, Y., and M. Sellke (2025): "Nonparametric MLE for Gaussian Location Mixtures: Certified Computation and Generic Behavior," Available at https://arxiv.org/abs/2503.20193 (arXiv:2503.20193 [math.ST]).

Robbins, H. (1956): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, p. 157–163, Berkeley and Los Angeles. University of California Press.

Royden, H. L., and P. Fitzpatrick (2010): *Real Analysis*. Prentice Hall, Upper Saddle River, NJ, 4 edn.

Shen, Y., and Y. Wu (2025): "Poisson Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?," *Annals of Statistics*, forthcoming, available at https://imstat.org/journals-and-publications/annals-of-statistics/annals-of-statistics-future-papers/,.

Soloff, J. A., A. Guntuboyina, and B. Sen (2025): "Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1), 1–32.

Song, M., S. Lee, and S. Ng (2026): "Empirical Bayes Estimation in Heterogeneous Coefficient Panel Models," Available at `https://arxiv.org/abs/2601.07059` (arXiv:2601.07059 [econ.EM]).

Stein, C. (1956): "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 197–206, Berkeley and Los Angeles, CA. University of California Press.

Walters, C. (2024): "Chapter 3 - Empirical Bayes methods in labor economics," in *Handbook of Labor Economics*, ed. by C. Dustmann, and T. Lemieux, vol. 5, pp. 183–260. Elsevier.

Wang, H., S. Ibrahim, and R. Mazumder (2025): "Nonparametric Finite Mixture Models with Possible Shape Constraints: A Cubic Newton Approach," *SIAM Journal on Mathematics of Data Science*, 7(1), 163–188.

Yan, Y., K. Wang, and P. Rigollet (2024): "Learning Gaussian mixtures using the Wasserstein–Fisher–Rao gradient flow," *Annals of Statistics*, 52(4), 1774–1795.

Zhang, C.-H. (1997): "Empirical Bayes and Compound Estimation of Normal Means," *Statistica Sinica*, 7(1), 181–193.

Zhang, Y., Y. Cui, B. Sen, and K.-C. Toh (2024): "On Efficient and Scalable Computation of the Nonparametric Maximum Likelihood Estimator in Mixture Models," *Journal of Machine Learning Research*, 25(8), 1–46.

Supplement to "Empirical Bayes Estimation in Heterogeneous Coefficient Panel Models"

S-1 Proofs

S-1.1 Proofs of Main Results

Proof of Theorem 1. We divide the proof into two parts. Part (i) establishes identification of the marginal distribution of δ_i; part (ii) builds on this to show that the full mixing distribution G_* is identified.

Part (i). Take any solution $G_o \in \mathcal{G}$ to (7), which must satisfy $f_{G_o}(Y_i, X_i) = f_{G_*}(Y_i, X_i)$ almost surely. Let H_* and H_o denote the marginal distributions of δ_i induced by G_* and G_o, respectively. In what follows, we show that $H_* = H_o$.

Let $\boldsymbol{X}$ and $\boldsymbol{M}$ be given as in Assumption 2. Conditioning on $X_i = \boldsymbol{X}$ and premultiplying both sides of (2) by $\boldsymbol{M}'$, we obtain $\boldsymbol{M}'Y_i = \boldsymbol{M}'P_i e_i$. Taking conditional characteristic functions of $\boldsymbol{M}'Y_i$ according to $f_{G_o}(\cdot \,|\, \boldsymbol{X})$ and $f_{G_*}(\cdot \,|\, \boldsymbol{X})$, respectively, we have

$$\mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{M}'Y_i}|X_i = \boldsymbol{X}] = \int e^{-t'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}t/2} dH_o(\delta),$$
$$\mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{M}'Y_i}|X_i = \boldsymbol{X}] = \int e^{-t'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}t/2} dH_*(\delta).$$

Equating $\mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{M}'Y_i}|X_i = \boldsymbol{X}]$ with $\mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{M}'Y_i}|X_i = \boldsymbol{X}]$, we have

$$\begin{aligned}
&\int_{\mathcal{K}_\delta} \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{1/2}}{(2\pi)^{(T-d_\beta)/2}} e^{-t'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}t/2} \cdot |\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2} dH_o(\delta) \\
= &\int_{\mathcal{K}_\delta} \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{1/2}}{(2\pi)^{(T-d_\beta)/2}} e^{-t'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}t/2} \cdot |\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2} dH_*(\delta), \qquad (18)
\end{aligned}$$

which can be rewritten as

$$\mathcal{T}[0, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \tilde{H}_o](t) = \mathcal{T}[0, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \tilde{H}_*](t), \quad \forall t \in \mathbb{R}^{T-d_\beta},$$

with $\mathcal{T}$ defined before Lemma 1. The probability measures $\tilde{H}_o$ and $\tilde{H}_*$ are defined

via

$$d\tilde{H}_o(\delta) := \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2}}{C} dH_o(\delta), \quad d\tilde{H}_*(\delta) := \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2}}{C} dH_*(\delta),$$

where the norming constant $C > 0$ is

$$C = \int_{\mathcal{K}_\delta} |\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2} dH_*(\delta) = \int_{\mathcal{K}_\delta} |\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2} dH_*(\delta).$$

This equality follows from integrating both sides of (18) with respect to $t \in \mathbb{R}^{T-d_\beta}$, ensuring that both $\tilde{H}_*$ and $\tilde{H}_o$ are probability measures. We are prepared to apply Lemma 1. Since $\delta \mapsto (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}$ is injective by Assumption 2(ii), (30) is satisfied by $\Lambda(\mathcal{K}_\delta)$ which consists only of $(\lambda_1(\delta), \lambda_2(\delta)) := (0, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1})$. By Lemma 1, it follows that $\tilde{H}_* = \tilde{H}_o$, and hence $H_* = H_o$.

Part (ii). Next, we show that the conditional distribution of $\boldsymbol{X}\beta_i$ given $\delta_i = \delta$ is identified for each δ in the support of H_*.

Let $\boldsymbol{X}$ and $\boldsymbol{M}$ be again given as in Assumption 2. Computing the conditional characteristic functions of Y_i given $X_i = \boldsymbol{X}$ according to the densities $f_{G_o}(\cdot \,|\, \boldsymbol{X})$ and $f_{G_*}(\cdot \,|\, \boldsymbol{X})$, respectively, we obtain:

$$\varphi_o(t) := \mathbb{E}_{G_o}[e^{\mathrm{i}t'Y_i} | X_i = \boldsymbol{X}] = \int \mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i} | \delta_i = \delta] e^{-t'P(\delta)P(\delta)'t/2} dH_o(\delta),$$
$$\varphi_*(t) := \mathbb{E}_{G_*}[e^{\mathrm{i}t'Y_i} | X_i = \boldsymbol{X}] = \int \mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i} | \delta_i = \delta] e^{-t'P(\delta)P(\delta)'t/2} dH_*(\delta).$$

These function coincide with each other since $f_{G_o} = f_{G_*}$ By Part (i), we also have $H_o = H_*$. Fix $t \in \mathbb{R}^T$, and let

$$\psi(\delta) := \mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i} | \delta_i = \delta] - \mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i} | \delta_i = \delta].$$

Equating $\varphi_o(t + \boldsymbol{M}s)$ with $\varphi_*(t + \boldsymbol{M}s)$, we then obtain

$$\int \psi(\delta) \exp\left(-\frac{1}{2}\|P(\delta)'(\boldsymbol{M}s + t)\|^2\right) dH_*(\delta) = 0 \tag{19}$$

for all $s \in \mathbb{R}^{T-d_\beta}$.

For each $\delta \in \mathcal{K}_\delta$, we write

$$t = \boldsymbol{M}\alpha_\delta + e_\delta,$$

where $\alpha_\delta = (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}\boldsymbol{M}'P(\delta)P(\delta)'t$ is chosen so that $\boldsymbol{M}'P(\delta)P(\delta)'e_\delta = 0$. This allows us to rewrite (19) as

$$\begin{aligned} 0 &= \int \psi(\delta) \exp\left(-\frac{1}{2}\|P(\delta)'e_\delta\|^2 - \frac{1}{2}(s+\alpha_\delta)'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}(s+\alpha_\delta)\right) dH_*(\delta) \\ &= \int (2\pi)^{(T-d_\beta)/2}|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2}\psi(\delta)\exp\left(-\frac{1}{2}\|P(\delta)'e_\delta\|^2\right) \\ &\qquad \times \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{1/2}}{(2\pi)^{(T-d_\beta)/2}} \exp\left(-\frac{1}{2}(s+\alpha_\delta)'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}(s+\alpha_\delta)\right) dH_*(\delta) \end{aligned} \tag{20}$$

for all $s \in \mathbb{R}^r$. Define a complex Borel measure

$$d\mu(\delta) := (2\pi)^{(T-d_\beta)/2}|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{-1/2}\psi(\delta)\exp\left(-\frac{1}{2}\|P(\delta)'e_\delta\|^2\right) dH_*(\delta).$$

Our next goal is to establish that the real and imaginary parts of μ are both zero. By construction of μ, (20) implies that

$$\begin{aligned} &\mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu](s) \\ &= \int \frac{|\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}|^{1/2}}{(2\pi)^{(T-d_\beta)/2}} \exp\left(-\frac{1}{2}(s+\alpha_\delta)'\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M}(s+\alpha_\delta)\right) d\mu(\delta) = 0 \end{aligned} \tag{21}$$

for all $s \in \mathbb{R}^{T-d_\beta}$. Using the Jordan decomposition theorem, we decompose μ into

$$\mu = \mu_+^R - \mu_-^R + \mathrm{i}(\mu_+^I - \mu_-^I),$$

where $\mu_\pm^R$ are mutually singular non-negative Borel measures, representing the positive and negative real parts of μ, respectively. The same holds for the positive and negative imaginary parts $\mu_\pm^I$. By linearity of $\mathcal{T}$ with respect to μ, it follows from (21) that

$$\mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_+^R](s) = \mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_-^R](s), \tag{22}$$

$$\mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_I^+](s) = \mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_I^-](s). \tag{23}$$

Integrating both sides of (22) with respect to $s \in \mathbb{R}^{T-d_\beta}$, by Fubini's theorem, it follows

$$\begin{aligned}\mu_+^R(\mathcal{K}_\delta) &= \int_{\mathbb{R}^{T-d_\beta}} \mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_+^R](s)ds \\ &= \int_{\mathbb{R}^{T-d_\beta}} \mathcal{T}[-\alpha_\delta, (\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M})^{-1}, \mu_-^R](s)ds = \mu_-^R(\mathcal{K}_\delta).\end{aligned}$$

We examine two possible cases. If $\mu_-^R(\mathcal{K}_\delta) = \mu_+^R(\mathcal{K}_\delta) = 0$, the real part of μ vanishes trivially. Otherwise, we may without loss of generality assume that both μ_-^R and μ_+^R are probability measures. Lemma 1 then implies that $\mu_+^R = \mu_-^R$. However, since μ_+^R and μ_-^R are mutually singular, this is only possible when both μ_+^R and μ_-^R have total mass of zero, which leads to a contradiction. Thus, we conclude $\mu_-^R = \mu_+^R \equiv 0$, and hence the real part of μ is identically zero. The same applies to μ_I^+ and μ_I^-. It follows $\mu \equiv 0$, thus implying that $\psi(\delta) = 0$ H_*-a.e. on $\mathcal{K}_\delta$.

Thus far, we have established that

$$\mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta] - \mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta] = 0$$

for H_*-a.e. δ for each fixed $t \in \mathbb{R}^T$. Define the function

$$\omega(t, \delta) := \mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta] - \mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta]$$

which admits a jointly measurable version, and hence can be regarded as Borel-measurable in (t, δ). By Fubini's theorem, we have

$$\begin{aligned}0 = \int_{\mathbb{R}^T} H_*\left(\{\delta \in \mathcal{K}_\delta : \omega(t, \delta) \neq 0\}\right) dt &= (\mathrm{Leb} \times H_*)\left(\{(t, \delta) \in \mathbb{R}^T \times \mathcal{K}_\delta : \omega(t, \delta) \neq 0\}\right) \\ &= \int_{\mathcal{K}_\delta} \mathrm{Leb}\left(\{t \in \mathbb{R}^T : \omega(t, \delta) \neq 0\}\right) dH_*(\delta),\end{aligned}$$

where $\mathrm{Leb} \times H_*$ denotes the product measure of the Lebesgue measure and H_*. It follows that $\mathrm{Leb}(\{t : \omega(t, \delta) \neq 0\}) = 0$ for H_*-a.e. $\delta \in \mathcal{K}_\delta$. Observe that $\omega(\cdot, \delta)$ is a uniformly continuous function of $t \in \mathbb{R}^T$ for $\delta \in \mathcal{K}_\delta$. Since the set of all zeros of $\omega(\cdot, \delta)$ is dense in $\mathbb{R}^T$, $\omega(\cdot, \delta)$ must vanish everywhere. It follows that for H_*-a.e. δ,

$$\mathbb{E}_{G_o}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta] = \mathbb{E}_{G_*}[e^{\mathrm{i}t'\boldsymbol{X}\beta_i}|\delta_i = \delta], \quad \forall t \in \mathbb{R}^T.$$

This establishes that the conditional distribution of $\boldsymbol{X}\beta_i$ given δ_i is identified. Finally, since $\boldsymbol{X}$ has full column rank, this is equivalent to identifying the conditional distribution of β_i given δ_i, implying that G_* is identified from the conditional distribution $\beta_i \,|\, \delta_i$ and the marginal distribution of δ_i. □

Proof of Theorem 2. We verify a multivariate version of Assumptions 1-5 in Kiefer and Wolfowitz (1956).

KW Assumption 1 is trivially met, since $Y \mapsto f_G(Y, X)$ is a legitimate density with respect to the Lebesgue measure on $\mathbb{R}^T$. KW Assumption 2 requires

$$f_{\tilde{\mu}}(Y, X) \to f_{\mu}(Y, X) \quad \text{as} \quad d(\tilde{\mu}, \mu) \to 0$$

for any μ and $\mathbb{P}$-a.e. (Y, X), which is established in Lemma 2. KW Assumptions 3 and 4 are verified in Lemma 3(ii) and (iii), respectively. Finally, KW Assumption 5 follows from the fact that $\mathbb{E}[|\log f_{G_*}(Y_i, X_i)|] < \infty$ and the random variable $\log \bar{f}_{\mu,\varepsilon}(Y_i, X_i)$ is uniformly bounded above for all $\mu \in \bar{\mathcal{G}}$ and $\varepsilon > 0$, as established in Lemma 3(i) and (ii). □

Proof of Theorem 3. We first show that $\mathbb{E}[N^{-1}\sum_{i=1}^{N} |\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*|^p] \to 0$ as $N \to \infty$. Let $\epsilon > 0$ be fixed, and let $M > 0$ be given as in Assumption 4(ii). Consider the following decomposition

$$\tau_i = \tau_{M,i} + r_{M,i},$$

where

$$\tau_{M,i} \equiv \tau_M(\theta_i) := \frac{M}{\max\{|\tau(\theta_i)|, M\}}\tau(\theta_i) \equiv \begin{cases} \tau(\theta_i) & \text{if } |\tau(\theta_i)| \le M \\ \frac{M}{|\tau(\theta_i)|}\tau(\theta_i) & \text{if } |\tau(\theta_i)| > M \end{cases}$$

denotes the truncation of τ_i at level M, and $r_{M,i} = r_M(\theta_i) := \tau(\theta_i) - \tau_M(\theta_i)$ is the remainder term. By construction, part (i) of Assumption 4 implies that $\tau_M(\theta)$ is a bounded continuous function of θ, and part (ii) implies

$$\limsup_{N\to\infty} \mathbb{E} \int_{\Theta} \|r_M(\theta)\|^p d\hat{G}_N(\theta) \le \epsilon.$$

Write $\tau_{M,i}^*$ and $\hat{\tau}_{M,i}^{\mathrm{EB}}$ for the oracle and the EB decision rules for $\tau_{M,i}$, and similarly for

$r^*_{M,i}$ and $\hat{r}_{M,i}$. Using the fact that $\tau^*_i = \tau^*_{M,i} + r^*_{M,i}$ and $\hat{\tau}^{\mathrm{EB}}_i = \hat{\tau}^{\mathrm{EB}}_{M,i} + \hat{r}^{\mathrm{EB}}_{M,i}$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}^{\mathrm{EB}}_i - \tau^*_i|^p &\le C_p \left(\frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}^{\mathrm{EB}}_{M,i} - \tau^*_{M,i}|^p + \frac{1}{N}\sum_{i=1}^{N} |\hat{r}_{M,i} - r^*_{M,i}|^p \right) \\ &=: C_p(I_1 + I_2)\end{aligned}$$

for some $C_p < \infty$ that only depends on p. In the following steps, we show that both $\mathbb{E}[I_1]$ and $\mathbb{E}[I_2]$ converge to zero.

Let $M_0 > 0$ be a sufficiently large number such that $\mathbb{P}(\|Y_i\| \le M_0,\, M_0^{-1} I_{d_\beta} \le X_i' X_i \le M_0 I_{d_\beta}) \ge 1 - \epsilon/(2M)^p$. Let $\mathcal{R} = \{(Y, X) : \|Y\| \le M_0,\, M_0^{-1} I_{d_\beta} \le X'X \le M_0 I_{d_\beta}\}$, and let $\mathcal{R}_i = 1\{\|Y_i\| \le M_0,\, M_0^{-1} I_{d_\beta} \le X_i' X_i \le M_0 I_{d_\beta}\}$ be the corresponding indicator for unit i. We decompose I_1 into two terms:

$$I_{11} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}^{\mathrm{EB}}_{M,i} - \tau^*_{M,i}|^p (1 - \mathcal{R}_i), \quad I_{12} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}^{\mathrm{EB}}_{M,i} - \tau^*_{M,i}|^p \mathcal{R}_i.$$

It is straightforward to see that

$$\mathbb{E}[I_{11}] \le (2M)^p \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} (1 - \mathcal{R}_i) \right] \le \epsilon$$

by construction of $\mathcal{R}_i$. On the other hand,

$$\mathbb{E}[I_{12}] \le \mathbb{E} \sup_{i:(Y_i,X_i)\in\mathcal{R}} |\hat{\tau}^{\mathrm{EB}}_{M,i} - \tau^*_{M,i}|^p \le \mathbb{E} \sup_{(Y,X)\in\mathcal{R}} |\hat{\tau}^{\mathrm{EB}}_M(Y, X) - \tau^*_M(Y, X)|^p \to 0$$

by Lemma 4 and the bounded convergence theorem (Royden and Fitzpatrick, 2010). These results imply that $\limsup_{N\to\infty} \mathbb{E}[I_1] \le \epsilon$.

To show $\limsup_{N\to\infty} \mathbb{E}[I_2] \le \epsilon$, first observe that

$$I_2 \le C_p \left(\frac{1}{N}\sum_{i=1}^{N} |\hat{r}^{\mathrm{EB}}_{M,i}|^p + \frac{1}{N}\sum_{i=1}^{N} |r^*_{M,i}|^p \right) =: C_p(I_{21} + I_{22}).$$

For the first term, we have, by the Jensen's inequality,

$$|\hat{r}^{\mathrm{EB}}_{M,i}|^p \le \mathbb{E}_{\hat{G}}[|r_M(\theta_i)|^p \,|\, Y_i, X_i] = \frac{\int_\Theta |r_M(\theta)|^p \ell(Y_i \,|\, X_i, \theta) d\hat{G}_N(\theta)}{f_{\hat{G}}(Y_i, X_i)}, \quad i = 1, \ldots, N.$$

This implies

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{r}_{M,i}^{\mathrm{EB}}|^p \leq \frac{1}{N}\sum_{i=1}^{N}\frac{\int_\Theta |r_M(\theta)|^p \ell(Y_i \mid X_i,\theta) d\hat{G}_N(\theta)}{f_{\hat{G}}(Y_i,X_i)} = \int_\Theta \frac{1}{N}\sum_{i=1}^{N}\frac{\ell(Y_i \mid X_i,\theta)}{f_{\hat{G}}(Y_i,X_i)}|r_M(\theta)|^p d\hat{G}_N(\theta).$$

The first-order conditions for the NPMLE state that $N^{-1}\sum_{i=1}^{N}\ell(Y_i \mid X_i,\theta)/f_{\hat{G}}(Y_i,X_i) = 1$ for $\hat{G}_N$-a.e. θ, which imply

$$\int_\Theta \frac{1}{N}\sum_{i=1}^{N}\frac{\ell(Y_i \mid X_i,\theta)}{f_{\hat{G}}(Y_i,X_i)}|r_M(\theta)|^p d\hat{G}_N(\theta) = \int_\Theta |r_M(\theta)|^p d\hat{G}_N(\theta).$$

It follows that $\limsup_{N\to\infty}\mathbb{E}[I_{21}] \leq \epsilon$ by Assumption 4(ii).

Turning to the term I_{22}, we have, by Jensen's inequality,

$$I_{22} \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{G_*}[|r_M(\theta_i)|^p \mid Y_i, X_i],$$

which implies that $\limsup_{N\to\infty}\mathbb{E}[I_{22}] \leq \mathbb{E}_{G_*}[|r_M(\theta_i)|^p]$. Since $\theta \mapsto |r_M(\theta)|^p$ is a nonnegative, continuous function of θ and $\hat{G}_N$ converges weakly to G_* almost surely, Portmanteau theorem implies that

$$\int_\Theta |r_M(\theta)|^p dG_*(\theta) \leq \liminf_{N\to\infty}\int_\Theta |r_M(\theta)|^p d\hat{G}_N(\theta) \quad \text{a.s.}$$

By Fatou's lemma, $\mathbb{E}[\liminf_{N\to\infty}\int_\Theta |r_M(\theta)|^p d\hat{G}_N(\theta)] \leq \liminf_{N\to\infty}\mathbb{E}[\int_\Theta |r_M(\theta)|^p d\hat{G}_N(\theta)]$. By Assumption 4(ii), this implies that $\limsup_{N\to\infty}\mathbb{E}[I_{22}] \leq \epsilon$.

Putting these pieces together, we conclude that

$$\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*|^p\right] \leq 3C_p\epsilon.$$

Since $\epsilon > 0$ was arbitrarily chosen, we have $\mathbb{E}[N^{-1}\sum_{i=1}^{N}|\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*|^p] \to 0$.

To prove the regret consistency of $\hat{\tau}_i^{\mathrm{EB}}$, observe that Assumption 4 holding for some $p \in [2,\infty)$ automatically implies that it also holds for $p = 2$. By the same argument, we have $\mathbb{E}[N^{-1}\sum_{i=1}^{N}|\hat{\tau}_i^{\mathrm{EB}} - \tau_i^*|^2] \to 0$, from which regret consistency follows. □

Proof of Proposition 1. It suffices to verify Assumption 2 and it is verified in the main text. See Section 3.1.1 for details. □

Proof of Proposition 2. Again, it suffices to verify Assumption 2. See Appendix O-1.1 in the arXiv version of this paper for details. □

Proof of Proposition 3. By assumption, there exists $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4) \in \text{supp}(X_{2,i})$ such that either $\Delta\boldsymbol{x}_2 = \pm\Delta\boldsymbol{x}_3 \neq 0$ or $\Delta\boldsymbol{x}_2 \neq \Delta\boldsymbol{x}_4$ where $\Delta\boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$. Upon inspecting the proof of Proposition 4, we find that the assumption $T \geq 5$ is explicitly invoked only in Cases 1(c), 2(c), and 3(c); in the other cases, $T = 4$ suffices for identification. Further inspection reveals that Case 1(c) requires $(\Delta\boldsymbol{x}_2, \Delta\boldsymbol{x}_3, \Delta\boldsymbol{x}_4) = (0, z, 0)$ for some $z \neq 0$. Similarly, Cases 2(c) and 3(c) require that $(\Delta\boldsymbol{x}_2, \Delta\boldsymbol{x}_3, \Delta\boldsymbol{x}_4)$ take the forms $(z, 0, z)$ for some $z \neq 0$, and (z, z', z) for some $z \neq \pm z'$, $z, z' \neq 0$, respectively. Since the given $\boldsymbol{x}$ belongs to none of these cases, the identification follows from the proof of Proposition 4. □

Proof of Proposition 4. We prove Proposition 4 by verifying Assumption 2. Let $\Delta a_t := a_{t+1} - a_t$ denote the forward first difference of a_t. By the assumption that $\Delta X_{2,it}$ is not identically zero, there is $\boldsymbol{x} = (\boldsymbol{x}_t)_{t=1}^T \in \text{supp}(X_{2,i})$ such that $\Delta\boldsymbol{x} = (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)_{t=1}^{T-1} \neq 0$. For notational ease, write $\boldsymbol{z} = \Delta\boldsymbol{x} \in \mathbb{R}^{T-1}$. Conditioning on $X_{2,i} = \boldsymbol{x}$, we obtain, after first-differencing,

$$\Delta Y_{it} = b_i \boldsymbol{z}_t + \Delta u_{it},$$

which yields

$$\frac{\Delta Y_{it}}{\boldsymbol{z}_t} = b_i + \frac{\Delta u_{it}}{\boldsymbol{z}_t} \quad \text{if} \quad \boldsymbol{z}_t \neq 0,$$

and

$$\Delta Y_{it} = \Delta u_{it} \quad \text{if} \quad \boldsymbol{z}_t = 0.$$

Let $\mathcal{T} := \{t : \boldsymbol{z}_t \neq 0\}$ be the set of indices at which $\boldsymbol{z}_t \neq 0$, with $t(1) < \cdots < t(K)$ being the increasing enumeration of $\mathcal{T}$ and $K = |\mathcal{T}|$. The annihilator matrix $\boldsymbol{M}$ can be chosen so that

$$(\boldsymbol{M}' Y_i)_k = \frac{\Delta u_{it(k+1)}}{\boldsymbol{z}_{t(k+1)}} - \frac{\Delta u_{it(k)}}{\boldsymbol{z}_{t(k)}} \quad \text{for } k = 1, \ldots, K-1,$$

for the first $K-1$ elements, and

$$((\boldsymbol{M}'Y_i)_K, \ldots, (\boldsymbol{M}'Y_i)_{T-1}) = (\Delta u_{it})_{t \in \mathcal{T}^c}.$$

Our objective is to determine $\delta = (\sigma^2, \rho)$ from the covariance $\boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M} = \mathbb{V}(\boldsymbol{M}'Y_i \,|\, \delta_i = \delta)$ of the normalized second differences $\frac{\Delta u_{it(k+1)}}{\boldsymbol{z}_{t(k+1)}} - \frac{\Delta u_{it(k)}}{\boldsymbol{z}_{t(k)}}$ for $k = 1, \ldots, K-1$ and the first differences Δu_{it} for $t \in \mathcal{T}^c$.

The argument proceeds by nested case analysis. We first consider Case 1.

Case 1: *(Two or more zeros in $\boldsymbol{z}$.)* Assume $|\mathcal{T}^c| \geq 2$. For any $s > t$ satisfying $\boldsymbol{z}_t = 0$ and $\boldsymbol{z}_s = 0$, we find that

$$\mathbb{V}\left(\Delta u_{it} \,|\, \delta_i = \delta\right) = \frac{2\sigma^2}{1+\rho}$$

and

$$\mathrm{Cov}\left(\Delta u_{it}, \Delta u_{is} \,|\, \delta_i = \delta\right) = -\frac{\rho^{s-t-1}(1-\rho)\sigma^2}{1+\rho}.$$

Using these equations, we can pin down the values of

$$\frac{\sigma^2}{1+\rho} \quad \text{and} \quad \rho^{|s-t|-1}(1-\rho) \quad \forall t, s \in \mathcal{T}^c. \tag{24}$$

We further divide into three subcases and show that δ can be uniquely determined in each subcase.

– *Case 1(a):* When s and t are consecutive, i.e., $s = t+1$, it follows from (24) that both $1-\rho$ and $\sigma^2/(1+\rho)$ are uniquely determined. These together identify $\delta = (\sigma^2, \rho)$.

– *Case 1(b):* Suppose there are at least three elements $t_1 < t_2 < t_3$ in $\mathcal{T}^c$ with $d_1 := t_2 - t_1 \geq 1$ and $d_2 := t_3 - t_2 \geq 1$. Then, the values of $\rho^{d_1-1}(1-\rho)$, $\rho^{d_2-1}(1-\rho)$, and $\rho^{d_1+d_2-1}(1-\rho)$ are identified from (24). If all three are zero, this implies $\rho = 0$, and vice versa. Thus, ρ is identified in this case, and σ^2 is also identified from $\sigma^2/(1+\rho)$. Thus, δ is identified.

Otherwise, all three have nonzero values, and one can pin down

$$\frac{\rho^{d_1-1}(1-\rho) \cdot \rho^{d_2-1}(1-\rho)}{\rho^{d_1+d_2-1}(1-\rho)} = \rho^{-1}(1-\rho) = \frac{1}{\rho} - 1,$$

implying that $\delta = (\sigma^2, \rho)$ is identified in this case as well.

– *Case 1(c):* Suppose that $|\mathcal{T}^c| = 2$ and the elements $t, s \in \mathcal{T}^c$ $(t < s)$ are not

consecutive. Note that $|\mathcal{T}| = T - 1 - |\mathcal{T}^c| = T - 3 \geq 2$, and hence there are at least two time indices for which $\boldsymbol{z}_t \neq 0$. It must therefore be either $t-1, t+1 \in \mathcal{T}$ or $s-1, s+1 \in \mathcal{T}$. Without loss of generality, assume that $t-1, t+1 \in \mathcal{T}$. We then observe

$$\operatorname{Cov}\left(\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}} - \frac{\Delta u_{it-1}}{\boldsymbol{z}_{t-1}}, \Delta u_{it} \,|\, \delta_i = \delta\right) = -\frac{(1-\rho)\sigma^2}{1+\rho}\left(\frac{1}{\boldsymbol{z}_{t+1}} - \frac{1}{\boldsymbol{z}_{t-1}}\right),$$

and

$$\operatorname{Cov}\left(\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}} - \frac{\Delta u_{it-1}}{\boldsymbol{z}_{t-1}}, \Delta u_{is} \,|\, \delta_i = \delta\right) = -\frac{(1-\rho)\rho^{s-t-2}\sigma^2}{1+\rho}\left(\frac{1}{\boldsymbol{z}_{t+1}} - \frac{\rho^2}{\boldsymbol{z}_{t-1}}\right).$$

If $\boldsymbol{z}_{t+1} \neq \boldsymbol{z}_{t-1}$, the first equation identifies $(1-\rho)\sigma^2/(1+\rho)$. Since $\sigma^2/(1+\rho)$ is already identified by (24), these together uniquely determine ρ and σ^2, and hence δ. If $\boldsymbol{z}_{t+1} = \boldsymbol{z}_{t-1} \neq 0$, we use the second equation to identify

$$\frac{(1-\rho)\rho^{s-t-2}\sigma^2}{1+\rho}(1-\rho^2) = \frac{\sigma^2}{1+\rho}\rho^{s-t-1}(1-\rho)\frac{1-\rho^2}{\rho}.$$

Together with (24), this further identifies

$$\frac{\rho}{1-\rho^2},$$

which is a strictly increasing function of $\rho \in (-1, 1)$. Hence, ρ is identified, and so is δ.

Case 2: *(Only one zero in $\boldsymbol{z}$.)* Let t be the only index for which $\boldsymbol{z}_t = 0$. As noted earlier in Case 1, we can then identify

$$\mathbb{V}\left(\Delta u_{it} \,|\, \delta_i = \delta\right) = \frac{2\sigma^2}{1+\rho}.$$

Again consider three subcases as follows.

– *Case 2(a):* Assume that $t = 1$ or $t = T-1$, i.e., that the only zero of $\boldsymbol{z}$ occurs either the first or last time period. Without loss of generality, let $t = 1$. Since $2, 3 \in \mathcal{T}$, we observe that

$$\mathbb{V}\left(\frac{\Delta u_{i3}}{\boldsymbol{z}_3} - \frac{\Delta u_{i2}}{\boldsymbol{z}_2} \,|\, \delta_i = \delta\right) = \frac{\sigma^2}{1+\rho}\left(\frac{2}{\boldsymbol{z}_3^2} + \frac{2}{\boldsymbol{z}_2^2} + \frac{2(1-\rho)}{\boldsymbol{z}_3\boldsymbol{z}_2}\right),$$

from which we can identify

$$\left(\frac{2}{\boldsymbol{z}_3^2}+\frac{2}{\boldsymbol{z}_2^2}+\frac{2(1-\rho)}{\boldsymbol{z}_3\boldsymbol{z}_2}\right).$$

Clearly, this uniquely determines ρ, and σ^2 subsequently follows from $\sigma^2/(1+\rho)$.

– *Case 2(b):* Assume that $1<t<T-1$ and that $\boldsymbol{z}_{t+1}\neq\boldsymbol{z}_{t-1}$, which must be nonzero. Observe that

$$\mathrm{Cov}\left(\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}}-\frac{\Delta u_{it-1}}{\boldsymbol{z}_{t-1}},\Delta u_{it}\,|\,\delta_i=\delta\right)=-\frac{(1-\rho)\sigma^2}{1+\rho}\left(\frac{1}{\boldsymbol{z}_{t+1}}-\frac{1}{\boldsymbol{z}_{t-1}}\right).$$

From this equation, we identify $(1-\rho)\sigma^2/(1+\rho)$. Combined with $\sigma^2/(1+\rho)$, we identify δ.

– *Case 2(c):* Assume that $1<t<T-1$ and $\boldsymbol{z}_{t+1}=\boldsymbol{z}_{t-1}\neq 0$. Since t is the only zero, either $\boldsymbol{z}_{t+2}\neq 0$ or $\boldsymbol{z}_{t-2}\neq 0$ must hold, depending on whether $t\geq 2$ or $t\leq T-2$. (For this, we invoke $T\geq 5$.) Without loss of generality, assume the first. Let $\boldsymbol{z}\neq 0$ denote the value shared by $\boldsymbol{z}_{t+1}=\boldsymbol{z}_{t-1}$. We use

$$\mathbb{V}\left(\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}}-\frac{\Delta u_{it-1}}{\boldsymbol{z}_{t-1}}\,|\,\delta_i=\delta\right)=\frac{2\sigma^2}{\boldsymbol{z}^2(1+\rho)}\left(2+\rho(1-\rho)\right),$$

along with $\sigma^2/(1+\rho)$, to identify $2+\rho(1-\rho)$. Next, we employ

$$\mathrm{Cov}\left(\frac{\Delta u_{it+2}}{\boldsymbol{z}_{t+2}}-\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}},\Delta u_{it}\,|\,\delta_i=\delta\right)=-\frac{(1-\rho)\sigma^2}{1+\rho}\left(\frac{\rho}{\boldsymbol{z}_{t+2}}-\frac{1}{\boldsymbol{z}_{t+1}}\right),$$

together with $\sigma^2/(1+\rho)$, to identify $\frac{(1-\rho)\rho}{\boldsymbol{z}_{t+2}}-\frac{(1-\rho)}{\boldsymbol{z}_{t+1}}$. Lastly, $\frac{(1-\rho)\rho}{\boldsymbol{z}_{t+2}}-\frac{(1-\rho)}{\boldsymbol{z}_{t+1}}$, combined with $2+\rho(1-\rho)$ identified earlier, can be used to identify ρ. σ^2 is subsequently identified from $\sigma^2/(1+\rho)$.

<u>*Case 3*</u>*: (No zeros in $\boldsymbol{z}$.)* We show that δ can be identified using the covariance structure of

$$\left(\frac{\Delta u_{i2}}{\boldsymbol{z}_2}-\frac{\Delta u_{i1}}{\boldsymbol{z}_1},\quad\frac{\Delta u_{i3}}{\boldsymbol{z}_3}-\frac{\Delta u_{i2}}{\boldsymbol{z}_2},\quad\frac{\Delta u_{i4}}{\boldsymbol{z}_4}-\frac{\Delta u_{i3}}{\boldsymbol{z}_3}\right).$$

Although the last term requires $T\geq 5$, Cases 3(a) and(b) remain applicable under $T=4$, in which the first two terms are sufficient for identification.

We begin by noting that

$$\mathbb{V}\left(\frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}} - \frac{\Delta u_{it}}{\boldsymbol{z}_t} \,|\, \delta_i = \delta\right) = \frac{2\sigma^2}{1+\rho}\left(1 + \frac{\boldsymbol{z}_{t+1}}{\boldsymbol{z}_t} + \frac{\boldsymbol{z}_t}{\boldsymbol{z}_{t+1}} - \rho\right)\frac{1}{\boldsymbol{z}_{t+1}\boldsymbol{z}_t}, \quad t = 1, 2, 3.$$

Consider the following three subcases.

– *Case 3(a):* Assume that there exist $s < t$ such that $\frac{\boldsymbol{z}_{s+1}}{\boldsymbol{z}_s} + \frac{\boldsymbol{z}_s}{\boldsymbol{z}_{s+1}} \neq \frac{\boldsymbol{z}_{t+1}}{\boldsymbol{z}_t} + \frac{\boldsymbol{z}_t}{\boldsymbol{z}_{t+1}}$. We may assume $s = 1$ and $t = 2$. Then, we can use the equation

$$\frac{\boldsymbol{z}_2\boldsymbol{z}_1\mathbb{V}\left(\frac{\Delta u_{i2}}{\boldsymbol{z}_2} - \frac{\Delta u_{i1}}{\boldsymbol{z}_1} \,|\, \delta_i = \delta\right)}{\boldsymbol{z}_3\boldsymbol{z}_2\mathbb{V}\left(\frac{\Delta u_{i3}}{\boldsymbol{z}_3} - \frac{\Delta u_{i2}}{\boldsymbol{z}_2} \,|\, \delta_i = \delta\right)} = \frac{1 + \frac{\boldsymbol{z}_2}{\boldsymbol{z}_1} + \frac{\boldsymbol{z}_1}{\boldsymbol{z}_2} - \rho}{1 + \frac{\boldsymbol{z}_3}{\boldsymbol{z}_2} + \frac{\boldsymbol{z}_2}{\boldsymbol{z}_3} - \rho} = 1 + \frac{(\frac{\boldsymbol{z}_2}{\boldsymbol{z}_1} + \frac{\boldsymbol{z}_1}{\boldsymbol{z}_2}) - (\frac{\boldsymbol{z}_3}{\boldsymbol{z}_2} + \frac{\boldsymbol{z}_2}{\boldsymbol{z}_3})}{1 + \frac{\boldsymbol{z}_3}{\boldsymbol{z}_2} + \frac{\boldsymbol{z}_2}{\boldsymbol{z}_3} - \rho}$$

to pin down ρ, which then identifies δ.

If Case 3(a) does not hold, $\frac{\boldsymbol{z}_{t+1}}{\boldsymbol{z}_t} + \frac{\boldsymbol{z}_t}{\boldsymbol{z}_{t+1}}$ must be constant in t, which is equivalent to the existence of some $\kappa \neq 0, -1, 1$ such that

$$r_t := \frac{\boldsymbol{z}_{t+1}}{\boldsymbol{z}_t} \in \left\{\kappa, \frac{1}{\kappa}\right\} \text{ for } t = 1, 2, 3.$$

This implies that if Case 3(a) does not hold, we can identify

$$\frac{\sigma^2}{1+\rho}\left(1 + \kappa + \frac{1}{\kappa} - \rho\right) \tag{25}$$

from the equation above. Further using covariances, we identify

$$\begin{aligned}
&\operatorname{Cov}\left(\frac{\Delta u_{it+2}}{\boldsymbol{z}_{t+2}} - \frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}}, \frac{\Delta u_{it+1}}{\boldsymbol{z}_{t+1}} - \frac{\Delta u_{it}}{\boldsymbol{z}_t} \,|\, \delta_i = \delta\right) \\
&= \frac{1}{\boldsymbol{z}_{t+1}^2}\operatorname{Cov}\left(\frac{1}{r_{t+1}}\Delta u_{it+2} - \Delta u_{it+1}, \Delta u_{it+1} - r_t\Delta u_{it} \,|\, \delta_i = \delta\right) \\
&= -\frac{1}{\boldsymbol{z}_{t+1}^2}\frac{\sigma^2}{1+\rho}\left(2 + (1-\rho)\left(\frac{1}{r_{t+1}} + r_t - \frac{r_t}{r_{t+1}}\rho\right)\right).
\end{aligned} \tag{26}$$

Now proceed to Cases 3(b) and 3(c).

– *Case 3(b):* Assume that there are consecutive equal terms in r_t. Without loss of generality, assume that $r_1 = r_2 = \kappa$. Substituting r_1 and r_2 into (26) with $t = 1$, the

ratio of (26) to (25) identifies

$$\frac{-\frac{1}{\boldsymbol{z}_{t+1}^2}\frac{\sigma^2}{1+\rho}\left(2+(1-\rho)\left(\frac{1}{r_{t+1}}+r_t-\frac{r_t}{r_{t+1}}\rho\right)\right)}{\frac{\sigma^2}{1+\rho}\left(1+\kappa+\frac{1}{\kappa}-\rho\right)}=C\frac{(1-\rho)(\kappa+1/\kappa-\rho)+2}{1+\kappa+1/\kappa-\rho}, \tag{27}$$

where $C \neq 0$ is a known constant. Define $\alpha := \kappa + \frac{1}{\kappa}$, which satisfies $\alpha > 2$ when $\kappa > 0$, and $\alpha < -2$ when $\kappa < 0$. From the expression above, we can identify

$$-\frac{(1-\rho)(\kappa+1/\kappa-\rho)+2}{1+\kappa+1/\kappa-\rho}=\rho+\frac{\alpha+2}{\rho-(\alpha+1)},$$

whose value is denoted by A. If $\alpha < -2$, the function $\rho \mapsto \rho + \alpha + 2/(\rho - (\alpha+1))$ is strictly increasing in $\rho \in (-1, 1)$, which identifies ρ, and subsequently σ^2 from $\sigma^2/(1+\rho)(1+\kappa+1/\kappa-\rho)$. Otherwise, if $\alpha > 2$, the equation

$$\rho+\frac{\alpha+2}{\rho-(\alpha+1)}=A \tag{28}$$

admits at most two distinct solutions ρ_1 and ρ_2, which are linked by

$$\rho_2=\alpha+1+\frac{\alpha+2}{\rho_1-(\alpha+1)},$$

and vice versa symmetrically. Suppose ρ_1 lies in $(-1, 1)$. Consequently,

$$\rho_2=\alpha+1+\frac{\alpha+2}{\rho_1-(\alpha+1)}\geq\alpha+1+\frac{\alpha+2}{1-(\alpha+1)}=\alpha-\frac{2}{\alpha}\geq 1,$$

which necessarily falls outside $(-1, 1)$. Hence, (28) yields at most one solution in $(-1, 1)$, thereby identifying ρ, and subsequently δ.

– *Case 3(c):* Consider the case in which no consecutive terms in (r_1, r_2, r_3) are equal, i.e., (r_1, r_2, r_3) are alternating between κ and $1/\kappa$ with $\kappa \neq \pm 1$. Without loss of generality, let $r_1 = r_3 = \kappa$ and $r_2 = 1/\kappa$. We observe that

$$\begin{aligned}
\boldsymbol{z}_3\boldsymbol{z}_2\,\mathrm{Cov}\left(\frac{\Delta u_{i3}}{\boldsymbol{z}_3}-\frac{\Delta u_{i2}}{\boldsymbol{z}_2},\frac{\Delta u_{i2}}{\boldsymbol{z}_2}-\frac{\Delta u_{i1}}{\boldsymbol{z}_1}\,|\,\delta_i=\delta\right)&=-\frac{\sigma^2}{1+\rho}\left(\frac{2}{\kappa}+(1-\rho)\,(2-\kappa\rho)\right),\\
\boldsymbol{z}_4\boldsymbol{z}_3\,\mathrm{Cov}\left(\frac{\Delta u_{i4}}{\boldsymbol{z}_4}-\frac{\Delta u_{i3}}{\boldsymbol{z}_3},\frac{\Delta u_{i3}}{\boldsymbol{z}_3}-\frac{\Delta u_{i2}}{\boldsymbol{z}_2}\,|\,\delta_i=\delta\right)&=-\frac{\sigma^2}{1+\rho}\left(2\kappa+(1-\rho)\,(2-\rho/\kappa)\right).
\end{aligned}$$

Taking the difference of these equations, we identify

$$\left(\kappa - \frac{1}{\kappa}\right) \frac{\sigma^2}{1+\rho} \left(2 + (1-\rho)\rho\right),$$

where $\kappa - 1/\kappa \neq 0$ is known. Additionally, we know the value of

$$\boldsymbol{z}_4\boldsymbol{z}_2 \operatorname{Cov}\left(\frac{\Delta u_{i4}}{\boldsymbol{z}_4} - \frac{\Delta u_{i3}}{\boldsymbol{z}_3}, \frac{\Delta u_{i2}}{\boldsymbol{z}_2} - \frac{\Delta u_{i1}}{\boldsymbol{z}_1} \mid \delta_i = \delta\right) = \frac{\sigma^2}{1+\rho}(1-\rho)(\kappa\rho - 1)(\rho - \kappa).$$

Combined with (25) identified earlier, these two terms are used to identify the following:

$$-\frac{2 + (1-\rho)\rho}{\rho - (\alpha+1)} = \rho + \alpha + \frac{(\alpha+2)(\alpha-1)}{\rho - (\alpha+1)}$$

and

$$-\frac{(1-\rho)(\kappa\rho - 1)(\rho - \kappa)}{\rho - (\alpha+1)} = \rho^2 + \alpha - 1 + \frac{\alpha(\alpha+2)}{\rho - (\alpha+1)}.$$

Since α is known, this information is equivalent to knowing $\rho + (\alpha+2)(\alpha-1)/(\rho - (\alpha+1))$ and $\rho^2 + \alpha(\alpha+2)/(\rho - (\alpha+1))$, which can be used to identify

$$\left[\rho^2 + \frac{\alpha(\alpha+2)}{\rho - (\alpha+1)}\right] - \frac{\alpha}{\alpha - 1}\left[\rho + \frac{(\alpha+2)(\alpha-1)}{\rho - (\alpha+1)}\right] = \rho^2 - \frac{\alpha}{\alpha-1}\rho.$$

Label the known values of $\rho + (\alpha+2)(\alpha-1)/(\rho - (\alpha+1))$ and $\rho^2 - \alpha\rho/(\alpha-1)$ as B and C, respectively. Viewed as rational equations for ρ, this leads to the following system of quadratic equations:

$$\begin{aligned} &\rho^2 - (B+1)\rho + B(\alpha+1) - 2 = 0, \\ &\rho^2 - \frac{\alpha}{\alpha-1}\rho - C = 0. \end{aligned} \tag{29}$$

We show by way of contradiction that (29) can be consistent with at most one ρ in $(-1, 1)$, thus uniquely determining ρ. For the quadratic system (29) to admit multiple solutions, note that the two equations must share the same roots, and hence coincide, i.e.,

$$B = \frac{1}{\alpha - 1}, \quad C = 1 - \frac{2}{\alpha - 1}.$$

Plugging $C = 1 - \frac{2}{\alpha-1}$ into the second equation, the proof is complete once we verify

that

$$\rho^2 - \frac{\alpha}{\alpha-1}\rho + \frac{3-\alpha}{\alpha-1} = 0$$

admits at most one solution in $(-1, 1)$. To show this, evaluate this function at $\rho = 1$:

$$1 - \frac{\alpha}{\alpha-1} + \frac{3-\alpha}{\alpha-1} = \frac{2-\alpha}{\alpha-1}.$$

Similarly, at $\rho = -1$, we obtain

$$1 + \frac{\alpha}{\alpha-1} + \frac{3-\alpha}{\alpha-1} = \frac{\alpha+2}{\alpha-1}.$$

Since $\alpha = \kappa + 1/\kappa$ satisfies $|\alpha| > 2$, we have

$$\frac{2-\alpha}{\alpha-1}\frac{\alpha+2}{\alpha-1} = \frac{4-\alpha^2}{(\alpha-1)^2} < 0,$$

which shows that the function $\rho \mapsto \rho^2 - \frac{\alpha}{\alpha-1}\rho + \frac{3-\alpha}{\alpha-1}$ takes values of opposite signs at $\rho = 1$ and $\rho = -1$. Hence, it must have a unique root in $(-1, 1)$, leading to a contradiction. This concludes Case 3(c).

□

Proof of Proposition 5. We verify parts (i), (ii), and (iii) sequentially.
Proof of part (i): The regret consistency of $\hat{\sigma}_i^{2,\mathrm{EB}}$ and $\hat{\rho}_i^{\mathrm{EB}}$ is an immediate consequence of Theorem 3 because (σ_i^2, ρ_i) have compact support by Assumption 1.

Proof of part (ii): We show that Assumption 4(ii) holds for $\tau_i \in \{a_i, b_i\}$ with $p = 2$. By Markov's inequality, we have

$$\int_\Theta \|\beta\|^2 1\{\|\beta\| \geq M\} d\hat{G}_N(\theta) \leq \frac{1}{M^\varepsilon} \int_\Theta \|\beta\|^{2+\varepsilon} d\hat{G}_N(\theta).$$

It follows from Lemma 5 that

$$\int_\Theta \|\beta\|^2 1\{\|\beta\| \geq M\} d\hat{G}_N(\theta) \leq \frac{1}{M^\varepsilon} \frac{A^{2+\varepsilon}}{N} \sum_{i=1}^N \|Y_i\|^{2+\varepsilon},$$

and thus,

$$\limsup_{N\to\infty} \mathbb{E}\left[\int_\Theta \|\beta\|^2 1\{\|\beta\| \geq M\} d\hat{G}_N(\theta)\right] \leq \frac{A^{2+\varepsilon}}{M^\varepsilon} \mathbb{E}[\|Y_i\|^{2+\varepsilon}].$$

One can make the right-hand side arbitrarily close to zero by choosing a sufficiently large M. Since a_i and b_i are dominated in absolute value by $\|\beta_i\|$, this proves Assumption 4(ii) for these parameters. It follows from Theorem 3 that $\hat{a}_i^{\mathrm{EB}}$ and $\hat{b}_i^{\mathrm{EB}}$ are regret-consistent.

Proof of part (iii): By Cauchy-Schwarz inequality, we observe that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_{iT+1}^{\mathrm{EB}} - Y_{iT+1}^{*})^2/5 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}(\hat{a}_i^{\mathrm{EB}} - a_i^{*})^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}} - b_i^{*})^2|X_{2,iT+1}|^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{\rho}_i^{\mathrm{EB}} - \rho_i^{*})^2|Y_{iT}|^2 \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(\widehat{a\rho}_i^{\mathrm{EB}} - (a\rho)_i^{*})^2 + \frac{1}{N}\sum_{i=1}^{N}(\widehat{b\rho}_i^{\mathrm{EB}} - (b\rho)_i^{*})^2|X_{2,i,T}|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}(\hat{a}_i^{\mathrm{EB}} - a_i^{*})^2 + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}} - b_i^{*})^4}\sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_{2,iT+1}|^4} \\
&\quad + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\rho}_i^{\mathrm{EB}} - \rho_i^{*})^4}\sqrt{\frac{1}{N}\sum_{i=1}^{N}|Y_{iT}|^4} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(\widehat{a\rho}_i^{\mathrm{EB}} - (a\rho)_i^{*})^2 + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{b\rho}_i^{\mathrm{EB}} - (b\rho)_i^{*})^4}\sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_{2,i,T}|^4},
\end{aligned}
$$

where $(a\rho)_i := a_i\rho_i$ and $(b\rho)_i := b_i\rho_i$. Taking expectations on both sides, it suffices to show that the expectations of the terms on the RHS converge to zero.

Observe that $\tau_i \in \{a_i, b_i, a_i\rho_i, b_i\rho_i\}$ are all dominated in absolute value by $\|\beta_i\|$ since $|\rho_i| \leq 1$. Combining the moment condition $\mathbb{E}[\|Y_i\|^{4+\varepsilon}] < \infty$ with the argument from Part (ii), we find that Assumption 4(ii) holds for $\tau_i \in \{a_i, b_i, a_i\rho_i, b_i\rho_i\}$ with $p = 4$. By Theorem 3 (with $p = 2$), we have $\mathbb{E}[N^{-1}\sum_{i=1}^{N}(\hat{a}_i^{\mathrm{EB}} - a_i^{*})^2] \to 0$ and $\mathbb{E}[N^{-1}\sum_{i=1}^{N}(\widehat{a\rho}_i^{\mathrm{EB}} - (a\rho)_i^{*})^2] \to 0$ immediately. This addresses the first and fourth terms on the RHS. Applying Cauchy-Schwarz inequality, the second term on the RHS

is bounded above by

$$\mathbb{E}\left[\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}}-b_i^*)^4}\sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_{2,iT+1}|^4}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}}-b_i^*)^4\right]}\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|X_{2,iT+1}|^4\right]}.$$

Theorem 3 implies that $\mathbb{E}[N^{-1}\sum_{i=1}^{N}(\hat{b}_i^{\mathrm{EB}}-b_i^*)^4]\to 0$, while $\mathbb{E}[N^{-1}\sum_{i=1}^{N}|X_{2,iT+1}|^4]=\mathbb{E}[|X_{2,iT+1}|^4]$ is finite by assumption. Thus, the second term converges to zero. The third and fifth terms can be handled in an analogous way. We omit the details to avoid repetition. Each term on the right-hand side has expectation converging to zero. Hence $\mathbb{E}[N^{-1}\sum_{i=1}^{N}(\hat{Y}_{iT+1}^{\mathrm{EB}}-Y_{iT+1}^*)^2]\to 0$ as desired, which establishes the regret consistency of $\hat{Y}_{iT+1}^{\mathrm{EB}}$.

□

S-1.2 Technical Lemmas

S-1.2.1 Identification Lemma (Theorem 1)

For the sake of completeness, we reproduce the key identification result from Bruni and Koch (1985) as Lemma 1, which is primarily used in this paper. Let $n\in\mathbb{N}$, $d\in\mathbb{N}$, and let $\Theta\subseteq\mathbb{R}^n$ denote a compact parameter space for θ. Following the notation in Bruni and Koch (1985), we define the following class of functions:

$$\Lambda(\Theta):=\left\{(\lambda_1,\lambda_2):\Theta\to\mathbb{R}^d\times S_{+,d}\;\middle|\;\begin{array}{l}\|\lambda_1(\theta)\|+\|\dot{\lambda}_1(\theta)\|\leq K_1,\\ \|\lambda_2(\theta)\|+\|\dot{\lambda}_2(\theta)\|\leq K_1,\ K_2I_T\leq\lambda_2(\theta)\leq K_3I_T\end{array}\right\},$$

where $S_{+,d}$ denotes the set of $d\times d$ positive definite matrices, and K_j, $j=1,2,3$ are some positive constants. Each function $(\lambda_1(\theta),\lambda_2(\theta))$ represents the conditional mean and covariance of $Y_i\,|\,\theta_i\sim\mathcal{N}(\lambda_1(\theta_i),\lambda_2(\theta_i))$, where $\theta_i\in\Theta$ represents a random parameter. We also need the following restriction as in Bruni and Koch (1985): for any $(\lambda_1,\lambda_2),(\tilde{\lambda}_1,\tilde{\lambda}_2)\in\Lambda(\Theta)$ and $\theta,\tilde{\theta}\in\Theta$, it holds that

$$\lambda_1(\theta)=\tilde{\lambda}_1(\tilde{\theta}),\lambda_2(\theta)=\tilde{\lambda}_2(\tilde{\theta})\quad\Longrightarrow\quad\theta=\tilde{\theta}. \tag{30}$$

The condition in (30) requires that, for any $(\mu, \Sigma) \in \mathbb{R}^d \times S_{+,d}$, there exists at most one $\theta = \theta(\mu, \Sigma) \in \Theta$ such that

$$\exists(\lambda_1, \lambda_2) \in \Lambda(\Theta) : (\mu, \Sigma) = (\lambda_1(\theta), \lambda_2(\theta)),$$

thus preventing multiple labels within Θ. This condition is trivially satisfied when $\Lambda(\Theta)$ is a singleton consisting of a one-to-one function.

We write $\mathcal{M}(\Theta)$ for the set of complex Borel measures on Θ, and $\mathcal{P}(\Theta)$ for its subset consisting of probability measures. Finally, define $\mathcal{T} : \Lambda(\Theta) \times \mathcal{M}(\Theta) \to C(\mathbb{R}^d, \mathbb{C})$ by

$$\mathcal{T}[\lambda_1, \lambda_2, \mu](y) := \int \frac{1}{(2\pi)^{d/2}|\lambda_2(\theta)|^{1/2}} \exp\left(-\frac{1}{2}\|\lambda_2(\theta)^{-1/2}(y - \lambda_1(\theta))\|^2\right) d\mu(\theta),$$

where $C(\mathbb{R}^d, \mathbb{C})$ denotes the class of all complex-valued continuous functions on $\mathbb{R}^d$. Note that, for $\mu \in \mathcal{P}(\Theta)$, $\mathcal{T}[\lambda_1, \lambda_2, \mu](x)$ represents the density of $\int_\Theta \mathcal{N}(\lambda_1(\theta), \lambda_2(\theta))d\mu(\theta)$, i.e., the marginal density of

$$Y_i = \lambda_1(\theta_i) + \lambda_2(\theta_i)^{1/2} e_i$$

under $\theta_i \sim \mu$ and $e_i \,|\, \theta_i \sim \mathcal{N}(0, I_d)$. The following lemma from Bruni and Koch (1985) shows that $\mathcal{T}$ is an injective mapping, meaning that the mean and variance mixture components, as well as the mixing distribution, are identifiable from general Gaussian mixture distributions under the assumption of compact support.

Lemma 1 (Identification of Gaussian Mixture Models, Bruni and Koch, 1985) *Assume that* $\mathcal{T}[\lambda_1, \lambda_2, \mu](y) = \mathcal{T}[\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\mu}](y)$ *for all* $y \in \mathbb{R}^d$, *where* $(\lambda_1, \lambda_2), (\tilde{\lambda}_1, \tilde{\lambda}_2) \in \Lambda(\Theta)$ *and* $\mu, \tilde{\mu} \in \mathcal{P}(\Theta)$. *Then,* $\lambda_1 = \tilde{\lambda}_1$, $\lambda_2 = \tilde{\lambda}_2$, *and* $\mu = \tilde{\mu}$.

S-1.2.2 Consistency Lemmas (Theorem 2)

Let $\mathcal{P}_{\text{sub}}$ denote the class of all subprobability measures on Θ. We extend the metric d to $\mathcal{P}_{\text{sub}}$ by defining

$$d(\mu_0, \mu_1) := \sum_{r=1}^{\infty} \frac{1}{2^r} \left| \int_\Theta h_r(\theta) d\mu_0(\theta) - \int_\Theta h_r(\theta) d\mu_1(\theta) \right|, \tag{31}$$

where $(h_r)_{r=1}^{\infty}$ is the same as in Section 4. Let $\bar{\mathcal{G}}$ denote the closure of $\mathcal{G}$ relative to $\mathcal{P}_{\text{sub}}$ with respect to the metric d. Since d induces the vague topology on $\mathcal{P}_{\text{sub}}$, and $\mathcal{P}_{\text{sub}}$ is compact under this topology, $(\bar{\mathcal{G}}, d)$ is itself a compact metric space (Kallenberg, 2021). We extend $f_G(Y, X)$, originally defined for $(G, Y, X) \in \mathcal{G} \times \mathbb{R}^T \times \text{supp}(X_i)$, to $(\mu, Y, X) \in \bar{\mathcal{G}} \times \mathbb{R}^T \times \text{supp}(X_i)$ by

$$f_\mu(Y, X) := \int_\Theta \ell(Y \mid X, \theta) d\mu(\theta). \tag{32}$$

The next lemma establishes that the extended likelihood is continuous in $\mu \in \bar{\mathcal{G}}$.

Lemma 2 *Let Assumptions 1, 2, and 3 hold. Then, the mapping $\mu \mapsto f_\mu(Y, X) : \bar{\mathcal{G}} \to \mathbb{R}$ is continuous for all $(Y, X) \in \mathbb{R}^T \times \text{supp}(X_i)$.*

Proof of Lemma 2. Note that

$$\ell(Y \mid X, \theta) = \frac{1}{(2\pi)^{T/2}|P(\delta)|} \exp\left(-\frac{1}{2}\|P(\delta)^{-1}(Y - X\beta)\|^2\right) \leq C_1 \exp\left(-C_2\|Y - X\beta\|^2\right)$$

for some constants $C_1, C_2 > 0$ by Assumption 2(ii). Since $\text{rank}(X) = d_\beta$ for all $X \in \text{supp}(X_i)$ by Assumption 3(ii), it follows that $\|Y - X\beta\|^2 \to \infty$ as $\|\beta\| \to \infty$, and hence $\lim_{\|\theta\| \to \infty} |\ell(Y \mid X, \theta)| = 0$. Let $\varphi_{Y,X}(\theta) := \ell(Y \mid X, \theta)$. This implies that $\varphi_{Y,X} \in C_0(\Theta)$ for all pairs $(Y, X) \in \mathbb{R}^T \times \text{supp}(X_i)$, where $C_0(\Theta)$ denotes the class of continuous functions $h : \Theta \to \mathbb{R}$ such that $\lim_{\|\theta\| \to \infty} |h(\theta)| = 0$. It is well-known that $C_0(\Theta)$ is the completed space of $C_c(\Theta)$ with respect to the uniform norm $\|h\|_\infty = \sup_{\theta \in \Theta} |h(\theta)|$ (see, e.g., Section 21.4 of Royden and Fitzpatrick, 2010).

Now, consider a sequence $\mu_n \in \bar{\mathcal{G}}$ such that μ_n vaguely converges to $\mu \in \bar{\mathcal{G}}$, i.e., $\lim_{n \to \infty} d(\mu_n, \mu) = 0$. Fix a pair $(Y, X) \in \mathbb{R}^T \times \text{supp}(X_i)$. Since $(h_r)_{r=1}^{\infty}$ in (31) spans a dense subset of $C_c(\Theta)$, and hence of $C_0(\Theta)$, for any $\varepsilon > 0$, there exists a real sequence $(a_r)_{r=1}^{\bar{r}}$ with $\bar{r} < \infty$ such that $\left\|\varphi_{Y,X} - \sum_{r=1}^{\bar{r}} a_r h_r\right\|_\infty < \varepsilon$. This implies

$$\begin{aligned}
|f_{\mu_n}(Y, X) - f_\mu(Y, X)| &= \left| \int_\Theta \varphi_{Y,X}(\theta) d\mu_n(\theta) - \int_\Theta \varphi_{Y,X}(\theta) d\mu(\theta) \right| \\
&\leq \left| \int_\Theta \sum_{r=1}^{\bar{r}} a_r h_r(\theta) d\mu_n(\theta) - \int_\Theta \sum_{r=1}^{\bar{r}} a_r h_r(\theta) d\mu(\theta) \right| + 2\varepsilon \\
&\leq \sup_{1 \leq r \leq \bar{r}} |a_r 2^r| \cdot d(\mu_n, \mu) + 2\varepsilon
\end{aligned}$$

by construction of d, and hence

$$\limsup_{n\to\infty} |f_{\mu_n}(Y,X) - f_{\mu}(Y,X)| \leq 2\varepsilon.$$

Since $\varepsilon > 0$ can be made arbitrarily small, it follows that $|f_{\mu_n}(Y,X) - f_{\mu}(Y,X)| \to 0$, completing the proof.

□

Lemma 3 *Let Assumptions 1, 2, and 3 hold. Then, the following hold.*

(i) $\mathbb{E}_{G_*}[|\log f_{G_*}(Y_i, X_i)|] < \infty$.

(ii) For any $\varepsilon > 0$ *and* $\mu \in \bar{\mathcal{G}}$,

$$\bar{f}_{\mu,\varepsilon}(Y_i, X_i) := \sup\{f_{\tilde{\mu}}(Y_i, X_i) : d(\tilde{\mu}, \mu) \leq \varepsilon\}$$

is a uniformly bounded random variable. Moreover,

$$\lim_{\varepsilon\to 0} \mathbb{E}[\log \bar{f}_{\mu,\varepsilon}(Y_i, X_i)] = \mathbb{E}[\log f_{\mu}(Y_i, X_i)].$$

(iii) For any $\mu \in \bar{\mathcal{G}}$ *with* $\mu \neq G_*$, $f_{\mu}(Y_i, X_i) \neq f_{G_*}(Y_i, X_i)$ *with positive probability.*

Proof of Lemma 3. Part (i): Since $\underline{c} I_T \leq P(\delta)P(\delta)' \leq \bar{c} I_T$ for all $\delta \in \mathcal{K}_\delta$ by Assumption 1(ii), we have

$$C_1 \int \exp(-C_2 \|Y_i - X_i\beta\|^2) dG_*(\theta) \leq f_{G_*}(Y_i, X_i) \leq C_3$$

for some absolute constants $C_1, C_2, C_3 > 0$. Taking logarithms on both sides and applying Jensen's inequality, it follows that

$$\log(C_1) - C_2 \int \|Y_i - X_i\beta\|^2 dG_*(\theta) \leq \log f_{G_*}(Y_i, X_i) \leq \log(C_3),$$

which, in turn, implies

$$|\log f_{G_*}(Y_i, X_i)| \leq |\log(C_1)| + |\log(C_3)| + C_2 \int \|Y_i - X_i\beta\|^2 dG_*(\theta).$$

Using $\|Y_i - X_i\beta\|^2 \leq 2\|Y_i\|^2 + 2\|X_i\beta\|^2$, it follows that

$$\mathbb{E}[|\log f_{G_*}(Y_i, X_i)|] \leq |\log(C_1)| + |\log(C_3)| + 2\mathbb{E}[\|Y_i\|^2] + 2\mathbb{E}_{G_*}[\|X_i\beta_i\|^2] < \infty,$$

since $\mathbb{E}_{G_*}[\|X_i\beta_i\|^2] \leq \mathbb{E}_{G_*}[\|X_i\beta_i\|^2] + \mathbb{E}_{G_*}[\|P_i e_i\|^2] = \mathbb{E}[\|Y_i\|^2] < \infty$ by Assumption 3(i).

Part (ii): Let $\varepsilon > 0$ and $\mu \in \bar{\mathcal{G}}$ be given. Since $(\bar{\mathcal{G}}, d)$ is a compact metric space, $\{\tilde{\mu} \in \bar{\mathcal{G}} : d(\tilde{\mu}, \mu) \leq \varepsilon\}$ is closed and hence compact, and therefore admits a countable dense subset $\{\tilde{\mu}_j\}_{j \geq 1}$. We thus have

$$\bar{f}_{\mu,\varepsilon}(Y_i, X_i) = \sup_{j \in \mathbb{N}} f_{\tilde{\mu}_j}(Y_i, X_i)$$

by the continuity of $\mu \mapsto f_\mu$ established in Lemma 2. Since each $f_{\tilde{\mu}_j}(Y_i, X_i)$ is a random variable, we find that $\bar{f}_{\mu,\varepsilon}(Y_i, X_i)$ is as well. The uniform boundedness of $\bar{f}_{\mu,\varepsilon}(Y_i, X_i)$ follows from that of $f_\mu(Y, X)$ across all (Y, X) and $\mu \in \mathcal{P}_{\text{sub}}$.

Part (iii): The case $\mu \in \mathcal{G}$ with $\mu \neq G_*$ is established in Theorem 1. Consider instead the case $\mu \in \bar{\mathcal{G}} \setminus \mathcal{G}$. Suppose, for contradiction, $f_\mu(Y_i, X_i) = f_{G_*}(Y_i, X_i)$ a.s. This implies

$$\int_{\mathbb{R}^T} f_\mu(y, X_i)dy = \int_{\mathbb{R}^T} \frac{f_\mu(y, X_i)}{f_{G_*}(y, X_i)} f_{G_*}(y, X_i)dy = \mathbb{E}\left[\left.\frac{f_\mu(Y_i, X_i)}{f_{G_*}(Y_i, X_i)}\right| X_i\right] = 1.$$

This leads to a contradiction, since by Fubini's theorem,

$$\int_{\mathbb{R}^T} f_\mu(y, X_i)dy = \int_\Theta \left(\int_{\mathbb{R}^T} \ell(y \,|\, X_i, \theta)dy\right) d\mu(\theta) = \mu(\Theta) < 1.$$

□

S-1.2.3 Regret Consistency Lemmas (Theorem 3, Proposition 5)

Lemma 4 *Let* $\tau^*(Y, X) = \mathbb{E}_{G_*}[\tau_i \,|\, (Y_i, X_i) = (Y, X)]$ *and* $\hat{\tau}^{\text{EB}}(Y, X) = \mathbb{E}_{\hat{G}}[\tau_i \,|\, (Y_i, X_i) = (Y, X)]$ *denote the oracle and EB decision rules for* τ_i*, respectively. Let* $\mathcal{R} = \{(Y, X) \in \mathbb{R}^T \times \text{supp}(X_i) : \|Y\| \leq M,\ M^{-1}I_{d_\beta} \leq X'X \leq MI_{d_\beta}\}$ *for a positive constant* $M > 1$*. Let Assumptions 1, 2, 3, and 4 hold. Then,* $\sup_{(Y,X)\in\mathcal{R}} |\hat{\tau}^{\text{EB}}(Y, X) - \tau^*(Y, X)| \xrightarrow{\text{a.s.}} 0$.

Proof of Lemma 4. We first observe that

$$\hat{\tau}^{\mathrm{EB}}(Y,X) = \frac{\int_\Theta \tau(\theta)\ell(Y \mid X,\theta)d\hat{G}(\theta)}{\int_\Theta \ell(Y \mid X,\theta)d\hat{G}(\theta)} = \frac{\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)d\hat{G}(\theta)}{\int_\Theta \varphi_{Y,X}(\theta)d\hat{G}(\theta)},$$

where $\varphi_{Y,X}(\theta) = \ell(Y \mid X,\theta)$ is defined in the proof of Lemma 2. In that proof, we have established that $\varphi_{Y,X} \in C_0(\Theta)$ for each $(Y,X) \in \mathbb{R}^T \times \mathrm{supp}(X_i)$, and hence that

$$\int_\Theta \varphi_{Y,X}(\theta)d\hat{G}(\theta) \xrightarrow{\text{a.s.}} \int_\Theta \varphi_{Y,X}(\theta)dG_*(\theta)$$

as $d(\hat{G}, G_*) \xrightarrow{\text{a.s.}} 0$. To further establish that the above convergence occurs uniformly in $(Y,X) \in \mathbb{R}^T \times \mathrm{supp}(X_i)$, we verify that the family $\{\varphi_{Y,X} : (Y,X) \in \mathcal{R}\}$ forms a relatively compact subset of $C_0(\Theta)$ with respect to the sup-norm.

Since $X'P(\delta)^{-1}P(\delta)^{-1\prime}X \geq (M^{-1}\underline{c})I_T$ for all $\delta \in \mathcal{K}_\delta$, by Assumption 1 and the construction of $\mathcal{R}$, there exist some positive constants c_1 and c_2 such that

$$\sup_{(Y,X)\in\mathcal{R}} |\varphi_{Y,X}(\theta)| \leq c_1 \exp\left(-c_2\|\beta\|^2\right).$$

This implies that there exists $D > 0$ such that

$$\sup_{\theta:\|\beta\|\geq D} \sup_{(Y,X)\in\mathcal{R}} |\varphi_{Y,X}(\theta)| < \epsilon/2. \tag{33}$$

Consider the restriction of $\varphi_{Y,X}$ to $B_D \times \mathcal{K}_\delta$, where $B_D := \{\beta : \|\beta\| \leq D\}$. Arzela-Ascoli theorem implies that the class $\{\varphi_{Y,X}|_{B_D\times\mathcal{K}_\delta} : (Y,X) \in \mathcal{R}\}$ is relatively compact in $C(B_D \times \mathcal{K}_\delta)$ provided it is uniformly bounded and equi-continuous (Royden and Fitzpatrick, 2010). The uniform boundedness is obvious from the previous argument. The uniform equicontinuity follows from the fact that

$$\sup_{(Y,X)\in\mathcal{R}} \|\nabla_\theta \varphi_{Y,X}(\theta)\| \leq c_1' \exp\left(-c_2\|\beta\|^2\right)(\|\beta\|^2 + 1),$$

for some $c_1' > 0$ and the same c_2 as before, where $\nabla_\theta = \partial/\partial\theta$. Taking the supremum over θ, this implies that $\sup_{\theta\in B_D\times\mathcal{K}_\delta}\left(\sup_{(Y,X)\in\mathcal{R}} \|\nabla_\theta\varphi_{Y,X}(\theta)\|\right) < \infty$. Consequently, Arzela-Ascoli theorem implies that there exist $(Y_1,X_1),\ldots,(Y_K,X_K) \in \mathcal{R}$, $K < \infty$

such that, for all $(Y, X) \in \mathcal{R}$,

$$\min_{k=1,\ldots,K} \sup_{\theta \in B_D \times \mathcal{K}_\delta} |\varphi_{Y,X}(\theta) - \varphi_{Y_k,X_k}(\theta)| < \epsilon$$

from which, combined with (33), it follows

$$\min_{k=1,\ldots,K} \sup_{\theta \in \Theta} |\varphi_{Y,X}(\theta) - \varphi_{Y_k,X_k}(\theta)| < \epsilon.$$

Therefore, $\{\varphi_{Y,X}\}_{(Y,X)\in\mathcal{R}}$ is relatively compact in $C_0(\Theta)$. The same line of arguments establish that $\{\tau \cdot \varphi_{Y,X}\}_{(Y,X)\in\mathcal{R}}$ is also relatively compact, whose details are omitted for brevity.

Since a relatively compact set admits a finite ϵ-net for every $\epsilon > 0$, by the standard approximation argument, we find that

$$\sup_{(Y,X)\in\mathcal{R}} \left| \int_\Theta \varphi_{Y,X}(\theta) d\hat{G}(\theta) - \int_\Theta \varphi_{Y,X}(\theta) dG_*(\theta) \right| \xrightarrow{\text{a.s.}} 0$$

and

$$\sup_{(Y,X)\in\mathcal{R}} \left| \int_\Theta \tau(\theta)\varphi_{Y,X}(\theta) d\hat{G}(\theta) - \int_\Theta \tau(\theta)\varphi_{Y,X}(\theta) dG_*(\theta) \right| \xrightarrow{\text{a.s.}} 0.$$

By the continuity of f_{G_*} and the compactness of $\mathcal{R}$, we observe that

$$\inf_{(Y,X)\in\mathcal{R}} \int_\Theta \varphi_{Y,X}(\theta) dG_*(\theta) = \inf_{(Y,X)\in\mathcal{R}} f_{G_*}(Y, X) \geq c > 0$$

for some $c > 0$, and hence

$$\liminf_{N\to\infty} \inf_{(Y,X)\in\mathcal{R}} \int_\Theta \varphi_{Y,X}(\theta) d\hat{G}(\theta) = \liminf_{N\to\infty} \inf_{(Y,X)\in\mathcal{R}} f_{\hat{G}}(Y, X) \geq c/2 > 0 \quad \text{a.s.}$$

Moreover, we have

$$\sup_{(Y,X)\in\mathcal{R}} \left(\left| \int_\Theta \tau(\theta)\varphi_{Y,X}(\theta) dG_*(\theta) \right| + |f_{G_*}(Y, X)| \right) < \infty.$$

This implies that

$$\begin{aligned}\sup_{(Y,X)\in\mathcal{R}}\left|\hat{\tau}^{\mathrm{EB}}(Y,X)-\tau^*(Y,X)\right| &= \sup_{(Y,X)\in\mathcal{R}}\left|\frac{\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)d\hat{G}(\theta)}{f_{\hat{G}}(Y,X)}-\frac{\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)dG_*(\theta)}{f_{G_*}(Y,X)}\right| \\ &\le \sup_{(Y,X)\in\mathcal{R}}\left|\frac{\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)d\hat{G}(\theta)-\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)dG_*(\theta)}{f_{\hat{G}}(Y,X)}\right| \\ &\quad+\sup_{(Y,X)\in\mathcal{R}}\left|\frac{\int_\Theta \tau(\theta)\varphi_{Y,X}(\theta)dG_*(\theta)}{f_{G_*}(Y,X)}\frac{f_{\hat{G}}(Y,X)-f_{G_*}(Y,X)}{f_{\hat{G}}(Y,X)}\right| \\ &\xrightarrow{\text{a.s.}} 0.\end{aligned}$$

□

Lemma 5 *Assume the HIVDX model in* (1)*. Let Assumption 1 hold. Further suppose that* $|\bar{X}_{2,i}|\le M$ *and* $\sum_{t=1}^T(X_{2,it}-\bar{X}_{2,i})^2\ge c$ *for some constants* $M,c>0$ *and all* $i=1,\dots,N$*. Then, there exists an absolute constant* $A<\infty$ *depending only on* $\bar{c}$*,* $\underline{c}$*,* M*,* c*, and* T*, such that*

$$\left(\int_\Theta \|\beta\|^p d\hat{G}_N(\theta)\right)^{1/p}\le A\left(\frac{1}{N}\sum_{i=1}^N\|Y_i\|^p\right)^{1/p}\qquad \text{for all } p\ge 2,$$

where $\hat{G}_N$ *denotes the NPMLE obtained from the sample* $(Y_i,X_i)_{i=1}^N$*.*

Proof of Lemma 5. By the first-order conditions for the NPMLE, we have

$$\frac{1}{N}\sum_{i=1}^N\frac{\ell(Y_i\,|\,X_i,\theta)}{f_{\hat{G}}(Y_i,X_i)}\le 1\quad\text{for all }\theta\in\Theta,$$

and $\hat{G}(\{\theta\in\Theta: N^{-1}\sum_{i=1}^N\ell(Y_i\,|\,X_i,\theta)/f_{\hat{G}}(Y_i,X_i)=1\})=1$. This implies

$$\frac{1}{N}\sum_{i=1}^N\frac{\ell(Y_i\,|\,X_i,\theta)}{f_{\hat{G}}(Y_i,X_i)}\frac{\partial\log\ell(Y_i\,|\,X_i,\theta)}{\partial\beta}=0$$

for $\hat{G}$-a.e. θ, and hence

$$\frac{1}{N}\sum_{i=1}^N\frac{\ell(Y_i\,|\,X_i,\theta)}{f_{\hat{G}}(Y_i,X_i)}X_i'(P(\delta)P(\delta)')^{-1}(Y_i-X_i\beta)=0.$$

Rearranging the equation above, we obtain

$$\beta = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} X_i' (P(\delta) P(\delta)')^{-1} X_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} X_i' (P(\delta) P(\delta)')^{-1} Y_i \right).$$

for $\hat{G}$-a.e. θ. Regarded as a weighted regression of $P(\delta)^{-1} Y_i$ on $P(\delta)^{-1} X_i$, this yields the following inequality:

$$\beta' \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} X_i' (P(\delta) P(\delta)')^{-1} X_i \right) \beta \leq \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \| P(\delta)^{-1} Y_i \|^2.$$

By Assumption 1, we have $\bar{c}^{-1} I_T \leq (P(\delta) P(\delta)')^{-1} \leq \underline{c}^{-1} I_T$. Let $\mu_i = (1/T) \sum_{t=1}^{T} X_{2,it}$ and $s_i^2 = (1/T) \sum_{t=1}^{T} (X_{2,it} - \mu_i)^2$. Then, the smallest eigenvalue of $X_i' X_i$ satisfies

$$\lambda_{\min}(X_i' X_i) \geq \frac{T s_i^2}{1 + \mu_i^2 + s_i^2} \geq \underline{\lambda} > 0 \quad \text{for all } i = 1, , \ldots, N,$$

where $\underline{\lambda} = \underline{\lambda}(M, c) = c/(1 + M^2 + c/T)$ is a positive constant. This implies that

$$\|\beta\|^2 = \|\beta\|^2 \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \leq \frac{\bar{c}}{\underline{c}\underline{\lambda}} \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \|Y_i\|^2 = A^2 \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \|Y_i\|^2$$

where $A^2 = \bar{c}/(\underline{c}\underline{\lambda}(M, c))$. By the Jensen's inequality, it follows that, for $p \geq 2$,

$$\|\beta\|^p \leq A^p \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \|Y_i\|^2 \right)^{p/2} \leq A^p \frac{1}{N} \sum_{i=1}^{N} \frac{\ell(Y_i \,|\, X_i, \theta)}{f_{\hat{G}}(Y_i, X_i)} \|Y_i\|^p.$$

Integrating both sides with respect to $\hat{G}$ yields

$$\int_{\Theta} \|\beta\|^p d\hat{G}(\theta) \leq A^p \frac{1}{N} \sum_{i=1}^{N} \|Y_i\|^p.$$

□

Online Appendix only for arXiv (do not include this for "Supplemental Material")

O-1 Auxiliary Identification Results

O-1.1 Verification of Assumption 2 for ARMA(1,1) Errors

We specify the choice of $(\boldsymbol{X}, \boldsymbol{M})$ for Assumption 2 in the context of the HIVD model with ARMA(1,1) errors. For simplicity, we assume $T = 4$. To illustrate the construction, let $\boldsymbol{X}$ be a column of ones (corresponding to the random intercept) and define the differencing matrix

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

This yields the differenced vector $\boldsymbol{M}'Y_i = (\Delta u_{i2}, \Delta u_{i3}, \Delta u_{i4})'$. Consequently, the transformed covariance matrix $\mathcal{V} := \boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M} = \mathbb{V}(\boldsymbol{M}'Y_i \mid \delta_i = \delta)$ has elements:

$$\begin{aligned}
\mathcal{V}_{11} &= 2\,\frac{1+\varphi^2-\varphi(1-\rho)}{1+\rho}\,\sigma^2, \\
\mathcal{V}_{12} &= \frac{\varphi^2(\rho-1)+\varphi(\rho^2-\rho+2)+(\rho-1)}{1+\rho}\,\sigma^2, \\
\mathcal{V}_{13} &= \frac{(\rho-1)(\rho+\varphi)(1+\rho\varphi)}{1+\rho}\,\sigma^2.
\end{aligned}$$

Using these moments, ρ is explicitly identified as

$$\rho = 1 + \frac{\mathcal{V}_{13}}{\mathcal{V}_{11}/2 + \mathcal{V}_{12}},$$

where the condition $\rho + \varphi \neq 0$ guarantees that the denominator is non-zero. Once ρ is determined, φ and σ^2 are recovered from the following system:

$$\frac{1+\rho}{2}\mathcal{V}_{11} + (1-\rho)(\rho\mathcal{V}_{12} - \mathcal{V}_{13}) = (1+\varphi^2)\sigma^2, \tag{34}$$

$$\rho\mathcal{V}_{12} - \mathcal{V}_{13} = \varphi\sigma^2. \tag{35}$$

We distinguish two cases based on the left-hand side of (35):

1. If $\rho\mathcal{V}_{12} - \mathcal{V}_{13} = 0$, then $\varphi = 0$ (since $\sigma^2 > 0$). In this scenario, σ^2 is directly identified by (34).

2. If $\rho\mathcal{V}_{12} - \mathcal{V}_{13} \neq 0$, then $\varphi \neq 0$. We take the ratio of (34) to (35) to eliminate σ^2:

$$\frac{1+\varphi^2}{\varphi} = \frac{\frac{1+\rho}{2}\mathcal{V}_{11} + (1-\rho)(\rho\mathcal{V}_{12} - \mathcal{V}_{13})}{\rho\mathcal{V}_{12} - \mathcal{V}_{13}} \equiv \mathcal{R}.$$

   Rearranging this expression leads to the quadratic equation $\varphi^2 - \mathcal{R}\varphi + 1 = 0$. From the properties of quadratic equations, $\mathcal{R}$ represents the sum of the two roots, while the constant 1 represents their product. This implies that the roots are reciprocals, say z and $1/z$. The condition $|\mathcal{R}| \geq 2$ (which holds by construction for any real φ) ensures that the roots are real. Finally, the assumption $|\varphi| \leq 1$ allows us to uniquely identify φ as the root that is less than or equal to 1 in absolute value.

Having identified φ, σ^2 is determined by $\sigma^2 = (\rho\mathcal{V}_{12} - \mathcal{V}_{13})/\varphi$. Therefore, we have verified the required one-to-one mapping between $\mathcal{V}$ and δ.

O-1.2 Identification Failure of the HIVDX model with $T = 4$

Proposition 3 indicates that G_* is not identified when $T = 4$ and the sequence of first differences $(\Delta X_{2,it})_{t=2}^4$ takes the symmetric form

$$(\Delta X_{2,i2}, \Delta X_{2,i3}, \Delta X_{2,i4}) = (x, \tilde{x}, x),$$

with $|\tilde{x}| \neq |x|$. To see this, consider the case where $X_{2,it} = \mathbf{1}\{t \geq 3\}$ for all i:

$$Y_{it} = a_i + b_i\mathbf{1}\{t \geq 3\} + u_{it}, \quad t = 1, \ldots, 4, \tag{36}$$

with u_{it} following the structure in (1b). Here, the difference sequence is $(0, 1, 0)$. To eliminate the random slope b_i (which appears only in the difference ΔY_{i3}), we

construct $\boldsymbol{M}$ to select only the second and fourth differences:

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{pmatrix}.$$

This yields $\boldsymbol{M}'Y_i = (\Delta u_{i2}, \Delta u_{i4})' = (u_{i2} - u_{i1}, u_{i4} - u_{i3})'$. The corresponding covariance matrix $\mathcal{V} := \boldsymbol{M}'P(\delta)P(\delta)'\boldsymbol{M} = \mathbb{V}(\boldsymbol{M}'Y_i \mid \delta_i = \delta)$ is 2×2 and imposes the following two restrictions on $\delta = (\sigma^2, \rho)$:

$$\begin{aligned} \mathcal{V}_{11} = \mathcal{V}_{22} &= \frac{2\sigma^2}{1+\rho}, \\ \mathcal{V}_{12} = \mathcal{V}_{21} &= -\frac{\rho(1-\rho)\sigma^2}{1+\rho}. \end{aligned}$$

These relations are insufficient to uniquely determine δ. For instance, the distinct parameter pairs $(1.04, 0.2)$ and $(1.56, 0.8)$ yield identical covariance matrices $\mathcal{V}$.

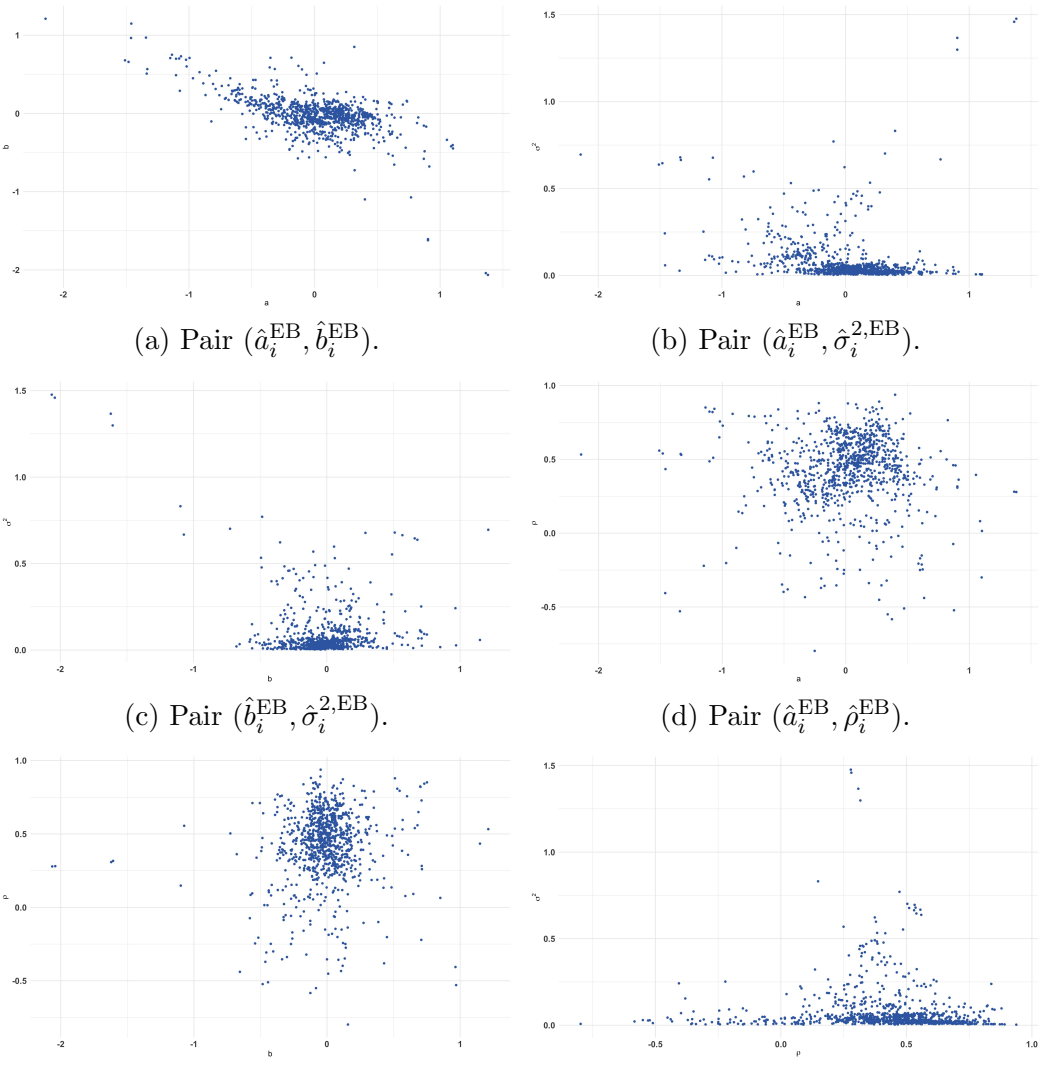
The intuition for this identification failure is straightforward. Because the differenced regressor sequence $(0, 1, 0)$ is symmetric and the error process u_{it} is stationary, the joint distribution of the utilized differences $(\Delta Y_{i2}, \Delta Y_{i4})$ is symmetric. Specifically, $\mathcal{V}_{11} = \mathcal{V}_{22}$ provides the scale of the variance, while the correlation depends on the ratio $\mathcal{V}_{12}/\mathcal{V}_{11} = -\rho(1-\rho)/2$. Since the function $f(\rho) = \rho(1-\rho)$ is not injective on $(-1, 1)$, solving for ρ involves a quadratic equation with potentially two distinct valid roots, precluding point identification.

This failure highlights that sufficient time variation in the covariates is a prerequisite for identification in short panels. However, as established in Proposition 4, the additional variation afforded by a fifth time period breaks this symmetry and ensures that δ is uniquely determined.

O-2 Additional Empirical Results

Figure 4 shows the pairwise scatter plots of the EB estimates for θ_i computed from the sample of $N = 938$ individuals, supporting the heterogeneity pattern reported in

Figure 4: Pairwise scatter plots of $\hat{\theta}_i^{\mathrm{EB}}$



(a) Pair $(\hat{a}_i^{\mathrm{EB}}, \hat{b}_i^{\mathrm{EB}})$.

(b) Pair $(\hat{a}_i^{\mathrm{EB}}, \hat{\sigma}_i^{2,\mathrm{EB}})$.

(c) Pair $(\hat{b}_i^{\mathrm{EB}}, \hat{\sigma}_i^{2,\mathrm{EB}})$.

(d) Pair $(\hat{a}_i^{\mathrm{EB}}, \hat{\rho}_i^{\mathrm{EB}})$.

(e) Pair $(\hat{b}_i^{\mathrm{EB}}, \hat{\rho}_i^{\mathrm{EB}})$.

(f) Pair $(\hat{\rho}_i^{\mathrm{EB}}, \hat{\sigma}_i^{2,\mathrm{EB}})$.

Note: Each panel shows the scatter plot of a pair of components of $(\hat{\theta}_i^{\mathrm{EB}})_{i=1}^N$.

Table 1.[14] Panel (a) illustrates that individual intercepts ($\hat{a}_i^{\text{EB}}$) and slopes ($\hat{b}_i^{\text{EB}}$) are negatively correlated. Panels (b) and (c) also reveal negative associations between $\hat{a}_i^{\text{EB}}$ and $\hat{\sigma}_i^{2,\text{EB}}$, and between $\hat{b}_i^{\text{EB}}$ and $\hat{\sigma}_i^{2,\text{EB}}$, respectively. In the remaining panels, individual persistence levels appear only weakly correlated with the other individual-specific parameters.

[14]To be precise, by the law of total variance, $\mathbb{V}(\theta_i) = \mathbb{V}(\hat{\theta}_i^{\text{EB}}) + \mathbb{E}[\mathbb{V}(\theta_i|(Y_{it}, X_{2,it})_{t=1}^T)]$, where the second term represents the expected posterior variance, reflecting the average variance of the EB estimator.