Forecasting with Dynamic Panel Data Models

Laura Liu
Indiana University

Hyungsik Roger Moon
University of Southern California
USC Dornsife INET, and Yonsei

Frank Schorfheide*
University of Pennsylvania
CEPR, NBER, and PIER

May 29, 2019

*Correspondence: L. Liu: Department of Economics, Indiana University, 100 South Woodlawn Ave., Bloomington, IN 47405. Email: lauraliu@iu.edu. H.R. Moon: Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089. E-mail: moonr@usc.edu. F. Schorfheide: Department of Economics, University of Pennsylvania, 133 S. 36th Street, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu. We thank Ulrich Müller (co-editor), three anonymous referees, Xu Cheng, Frank Diebold, Peter Phillips, Akhtar Siddique, and participants at various seminars and conferences for helpful comments and suggestions. Moon and Schorfheide gratefully acknowledge financial support from the National Science Foundation under Grants SES 1625586 and SES 1424843, respectively.

Abstract

This paper considers the problem of forecasting a collection of short time series using cross sectional information in panel data. We construct point predictors using Tweedie's formula for the posterior mean of heterogeneous coefficients under a correlated random effects distribution. This formula utilizes cross-sectional information to transform the unit-specific (quasi) maximum likelihood estimator into an approximation of the posterior mean under a prior distribution that equals the population distribution of the random coefficients. We show that the risk of a predictor based on a non-parametric kernel estimate of the Tweedie correction is asymptotically equivalent to the risk of a predictor that treats the correlated-random-effects distribution as known (ratio-optimality). Our empirical Bayes predictor performs well compared to various competitors in a Monte Carlo study. In an empirical application we use the predictor to forecast revenues for a large panel of bank holding companies and compare forecasts that condition on actual and severely adverse macroeconomic conditions.

JEL CLASSIFICATION: C11, C14, C23, C53, G21

KEY WORDS: Bank Stress Tests, Empirical Bayes, Forecasting, Panel Data, Ratio Optimality, Tweedie's Formula

1 Introduction

The main goal of this paper is to forecast a collection of short time series. Examples are the performance of start-up companies, developmental skills of small children, and revenues and leverage of banks after significant regulatory changes. In these applications the key difficulty lies in the efficient implementation of the forecast. Due to the short time span, each time series, taken by itself, provides insufficient sample information to precisely estimate unit-specific parameters. We will use the cross-sectional information in the sample to make inference about the distribution of heterogeneous parameters. This distribution can then serve as a prior for the unit-specific coefficients to sharpen posterior inference based on the short time series.

More specifically, we consider a linear dynamic panel model in which the unobserved individual heterogeneity, which we denote by the vector λ_i, interacts with some observed predictors W_{it-1}:

$$Y_{it} = \lambda_i' W_{it-1} + \rho' X_{it-1} + \alpha' Z_{it-1} + U_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \tag{1}$$

X_{it-1} is a vector of predetermined variables that may include lags of Y_{it}, Z_{it-1} is a vector of strictly exogenous covariates, and U_{it} is an unpredictable shock. Throughout this paper we adopt a correlated random effects approach in which the λ_is are treated as random variables that are possibly correlated with some of the predictors. An important special case is the linear dynamic panel data model in which $W_{it-1} = 1$, λ_i is a heterogeneous intercept, and the sole predictor is the lagged dependent variable: $X_{it-1} = Y_{it-1}$.

We develop methods to generate point forecasts of Y_{iT+1} which are evaluated under a quadratic loss function. Our paper builds on the dynamic panel literature that has developed consistent estimators of the common parameters (ρ, α) and focuses on the estimation of λ_i, which is essential for the prediction of Y_{it}. The benchmark for our prediction methods is the forecast that is based on the knowledge of the common coefficients (ρ, α) and the distribution $\pi(\lambda_i|\cdot)$ of the heterogeneous coefficients λ_i, but not the values λ_i themselves. This forecast is called oracle forecast. A natural notion of risk is that of compound risk, which is a (possibly weighted) cross-sectional average of expected losses. In a correlated random-effects setting, this averaging is done under the distribution $\pi(\lambda_i|\cdot)$, which means that the compound risk associated with the forecasts of the N units is the same as the integrated risk for the forecast of a particular unit i. It is well known, that the integrated risk is minimized by the Bayes predictor that minimizes the posterior expected loss conditional on time T information for unit i. Thus, the oracle replaces λ_i by its posterior mean under the prior distribution $\pi(\lambda_i|\cdot)$.

In practice neither the common coefficients (ρ, α) nor the distribution of the unit-specific coefficients $\pi(\lambda_i|\cdot)$ are known. Thus, we first replace the unknown common parameters by a consistent ($N \longrightarrow \infty$, T is fixed) estimator. Second, rather than computing the posterior mean of λ_i based on the likelihood function and an estimate of $\pi(\lambda_i|\cdot)$, we use a formula – attributed to the astronomer Arthur Eddington and the statistician Maurice Tweedie – that expresses the posterior mean of λ_i as a function of the cross-sectional density of certain sufficient statistics. We estimate this density either parametrically or non-parametrically from the cross-sectional information and plug it into what is in the statistics literature commonly referred to as Tweedie's formula. This leads to an empirical Bayes estimate of λ_i and an empirical Bayes predictor of Y_{iT+1}. The posterior mean predictor shrinks the estimates of the unit-specific coefficients toward a common prior mean, which reduces its sampling variability.

Our paper makes three contributions. First, we show in the context of the basic dynamic panel data model that an empirical Bayes predictor based on a consistent estimator of (ρ, α) and a kernel estimator of the cross-sectional density of the relevant sufficient statistics can asymptotically ($N \longrightarrow \infty$, T is fixed) achieve the same compound risk as the oracle predictor. Our main theorem provides a significant extension of the central result in Brown and Greenshtein (2009) from a vector-of-means model to a panel data model with estimated common coefficients. Importantly, the convergence result is uniform over families Π of correlated random effects distributions $\pi(\lambda_i|y_{i0})$ that contain point masses, i.e., span parameter homogeneity. As in Brown and Greenshtein (2009), we are able to show that the rate of convergence to the oracle risk accelerates in the case of homogeneous λ coefficients. Second, we provide a detailed Monte Carlo study that compares the performance of various implementations, both non-parametric and parametric, of our predictor. Third, we use our techniques to forecast pre-provision net-revenues (PPNRs) of a panel of banks.[1] Such forecasts are of interest to bank regulators and supervisors such as the Federal Reserve Board of Governors and the Office of the Comptroller of the Currency in the U.S.

Our empirical Bayes predictor can be compared to two easily implementable benchmark predictors whose implicit assumptions about the distribution of the λ_is correspond to special

[1] For a bank PPNR is the sum of net interest income and non-interest income less expenses, before making provisions for future losses on loans.

cases of our $\pi(\lambda_i|\cdot)$. The first predictor, which we call plug-in predictor, is obtained by estimating λ_i conditional on $(\hat{\rho}, \hat{\alpha})$ by maximum likelihood for each unit i. This predictor can be viewed as an approximation to the empirical Bayes predictor if $\pi(\lambda_i|\cdot)$ is very uninformative relative to the likelihood function associated with unit i. The second predictor, which we call pooled-OLS predictor, assumes homogeneity, i.e., $\lambda_i = \lambda$ for all i, and is based on a joint maximum likelihood estimate of (ρ, α, λ). It approximates the empirical Bayes estimator if $\pi(\lambda_i|\cdot)$ is very concentrated relative to unit i's likelihood. Asymptotically, the plug-in predictor and the pooled-OLS predictor are suboptimal because they do not converge to the oracle predictor. However, in finite samples at least one of them, depending on the amount of heterogeneity in the data, may work quite well because they do not rely on (potentially noisy) density estimates. In our Monte Carlo simulations and in the empirical analysis we document that in practice the empirical Bayes predictor dominates, either weakly or strictly, both the plug-in predictor and the pooled-OLS predictor.[2]

In our empirical application we forecast PPNRs of bank holding companies. The stress tests that have become mandatory under the Dodd-Frank Act require banks to establish how revenues vary in stressed macroeconomic and financial scenarios. We capture the effect of macroeconomic conditions on bank performance by including the unemployment rate, an interest rate, and an interest rate spread in the vector W_{it-1} in (1). Our analysis consists of two steps. We first document the superior forecast accuracy of the empirical Bayes predictor developed in this paper under the actual economic conditions, meaning that we set the aggregate covariates to their observed values. In a second step, we replace the observed values of the macroeconomic covariates by counterfactual values that reflect severely adverse macroeconomic conditions. According to our estimates, the effect of stressed macroeconomic conditions on bank revenues is heterogeneous, but typically small relative to the cross-sectional dispersion of revenues across holding companies.

Our paper is related to several strands of the literature. For $\alpha = \rho = 0$ and $W_{it} = 1$ the problem analyzed in this paper reduces to the classic problem of estimating a vector of means. In this context, Tweedie's formula has been used, for instance, by Robbins (1951) and more recently by Brown and Greenshtein (2009) and Efron (2011) in a "big data" application. Throughout this paper we are adopting an empirical Bayes approach, that uses cross-sectional information to estimate aspects of the prior distribution of the correlated random effects and then conditions on these estimates. Empirical Bayes methods date back

[2] While our theoretical results are based on a kernel implementation of the empirical Bayes predictor, the Monte Carlo experiments also include results for finite-mixture and nonparametric maximum likelihood estimates of the Tweedie correction.

to Robbins (1956); see Robert (1994) for a textbook treatment. We use compound decision theory as in Robbins (1964), Brown and Greenshtein (2009), Jiang and Zhang (2009) to state our optimality result. Because our setup nests the linear dynamic panel data model, we utilize results on the consistent estimation of ρ in dynamic panel data models with fixed effects when T is small, e.g., Anderson and Hsiao (1981), Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998), Alvarez and Arellano (2003).

The papers that are most closely related to ours are Gu and Koenker (2017a,b). They also consider a linear panel data model and use Tweedie's formula to construct an approximation to the posterior mean of the heterogeneous regression coefficients. However, their main object of interest is the estimation of the heterogeneous coefficients, and not out-of-sample forecasting. Their papers implement the empirical Bayes predictor based on a nonparametric maximum likelihood estimator, following Kiefer and Wolfowitz (1956), of the cross-sectional distribution of the sufficient statistics, but do not provide any theoretical optimality result. A key contribution of our work is to establish a rate at which the regret associated with the empirical Bayes predictor vis-a-vis the oracle predictor vanishes uniformly across families of correlated random-effects distributions Π that include point masses. Moreover, we also provide novel Monte Carlo and empirical evidence on the performance of the empirical Bayes procedures.

Building on the Bayesian literature on dynamic panel data models, e.g., Chamberlain and Hirano (1999), Hirano (2002), and Lancaster (2002), the paper by Liu (2018) develops a fully Bayesian (as opposed to empirical Bayes) approach to generate panel data forecasts. She uses mixtures of Gaussian linear regressions to construct a prior for the correlated random effects, which then is updated in view of the observed panel data. While the fully Bayesian approach is more suitable for density forecasting and can be more easily extended to nonlinear panel data models (see Liu, Moon, and Schorfheide (2018a) for an extension to a panel Tobit model), it is also a lot more computationally intensive. Moreover, it is much more difficult to establish convergence rates. Liu (2018) shows that her posterior predictive density converges strongly to the oracle's predictive density, but does not establish uniform bounds on the regret associated with the Bayes predictor.

There is an earlier panel forecast literature on the best linear unbiased prediction (BLUP) proposed by Goldberger (1962); see the survey article by Baltagi (2008). Compared to the BLUP-based forecasts, our forecasts based on Tweedie's formula have several advantages. First, it is known that the estimator of the unobserved individual heterogeneity parameter based on the BLUP method corresponds to the Bayes estimator based on a Gaussian prior (see, for example, Robinson (1991)), while our estimator based on Tweedie's formula is consistent with much more general prior distributions. Second, the BLUP method finds the forecast that minimizes the expected quadratic loss in the class of linear (in $(Y_{i0}, ..., Y_{iT})'$) and unbiased forecasts. Therefore, it is not necessarily optimal in our framework. Third, the existing panel forecasts based on the BLUP were developed for panel regressions with random effects and do not apply to correlated random effects settings.

The empirical application is related to the literature on "top-down" stress testing, which relies on publicly available bank-level income and capital data. Some authors, e.g. Covas, Rump, and Zakrajsek (2014), use time series quantile regression techniques to analyze revenue and balance sheet data for the relatively small set of bank holding companies with consolidated assets of more than 50 billion dollars. There are slightly more than 30 of these companies and they are subject to the Comprehensive Capital Analysis and Review (CCAR) conducted by the Federal Reserve Board of Governors. Because of mergers and acquisitions a lot of care is required to construct sufficiently long synthetic data sets that are amenable to time series analysis.

Closer to our work are the studies by Hirtle, Kovner, Vickery, and Bhanot (2016) and Kapinos and Mitnik (2016), which analyze PPNR components for a broader panel of banks. The former paper considers, among other techniques, pooled OLS estimation of models for bank income components and then uses the models to compute predictions under stressed macroeconomic conditions. The latter paper compares a standard fixed effect approach, a bank-by-bank time series approach, and fixed effects with optimal grouping in terms of out-of-sample forecasts, and finds that the bank-by-bank time-series approach does not perform well while the grouping approach provides better performance, which parallels our findings that the empirical Bayes methods usually outperform pooled OLS and plug-in predictors.

The remainder of the paper is organized as follows. Section 2 specifies the forecasting problem in the context of the basic dynamic panel data model. We introduce the compound decision problem, present the oracle forecast and its implementation through Tweedie's formula, and discuss the practical implementation. Section 3 establishes the ratio optimality of the empirical Bayes predictor and contains the main theoretical result of the paper. Monte Carlo simulation results are presented in Section 4. Section 5 discusses extensions to multi-step forecasting and a more general linear panel data model with covariates. The empirical application is presented in Section 6 and Section 7 concludes. Technical derivations, proofs, and the description of the data set used in the empirical analysis are relegated to the Appendix.

2 The Basic Dynamic Panel Data Model

We consider a panel with observations for cross-sectional units $i = 1, \ldots, N$ in periods $t = 1, \ldots, T$. The goal is to generate a point forecast for each unit i for period $t = T + h$. For now, we set $h = 1$ and assume that the observations Y_{it} are generated from a basic dynamic panel data model with homoskedastic Gaussian innovations:

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \quad U_{it} \sim iidN(0, \sigma^2), \tag{2}$$

where (λ_i, Y_{i0}) is independently and identically distributed (iid) with density $\pi(\lambda, y_0)$. This model is a restricted version of (1), where $W_{it-1} = 1$, $X_{it-1} = Y_{it-1}$, and $\alpha = 0$. Thus, λ_i is a unit-specific (heterogeneous) intercept. We combine the homogeneous parameters into the vector

$$\theta = (\rho, \sigma^2).$$

The purpose of investigating the simple model is to develop a full econometric theory of optimal forecasting.

In Section 2.1 we define the loss function under which the forecasts are evaluated and specify how we take expectations to construct our measure of risk. We construct an infeasible benchmark forecast in Section 2.2. This so-called oracle forecast is based on the posterior mean of λ_i under the prior $\pi(\lambda, y_0)$. The posterior mean can be conveniently evaluated using Tweedie's formula, which is discussed in Section 2.3. Finally, Section 2.4 describes how the infeasible oracle forecast can be turned into a feasible forecast by replacing unknown objects with estimates based on the cross-sectional information contained in the panel. Later in Section 5, we discuss extensions to the more general forecasting model (1), which is more relevant to empirical applications.

2.1 Compound Risk

The unit-specific forecasts $\hat{Y}_{iT+1}$ are evaluated under the conventional quadratic loss function and the forecast error losses are summed over the units i to obtain the compound loss

$$L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1}) = \sum_{i=1}^{N} (\widehat{Y}_{iT+1} - Y_{iT+1})^2, \tag{3}$$

where $Y_{T+1}^N = (Y_{1T+1}, \ldots, Y_{NT+1})$. We define the compound risk by taking expectations over (indicated by superscripts) the observed trajectories $\mathcal{Y}^N = (Y_1^{0:T}, \ldots, Y_N^{0:T})$ with $Y_i^{0:T} = (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$, the unobserved heterogeneous coefficients $\lambda^N = (\lambda_1, \ldots, \lambda_N)$, and future shocks $U_{T+1}^N = (U_{1T+1}, \ldots, U_{NT+1})$:

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N,\lambda^N,U_{T+1}^N}\left[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)\right]. \tag{4}$$

We use the subscripts to indicate that the expectation is conditional on the homogeneous parameter θ and the correlated random effects distribution π. Upper case variables, e.g., Y_{it}, are generally used to denote random variables, and lower case variables, e.g., y_{it}, to denote their realizations.

2.2 Oracle Forecast

In order to develop an optimality theory for the panel data forecasts, we begin by characterizing an infeasible benchmark forecast, which is called the oracle forecast. In the compound decision theory it is assumed that the oracle knows the distribution of the heterogeneous coefficients $\pi(\lambda_i, h_i)$, but it does not know the specific λ_i for unit i. In addition, the oracle knows θ and has observed the trajectories $\mathcal{Y}^N$.

Conditional on θ, the compound risk takes the form of an integrated risk that can be expressed as

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\mathbb{E}_{\theta,\pi,\mathcal{Y}^N}^{\lambda^N,U_{T+1}^N}[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)]\right]. \tag{5}$$

The inner expectation can be interpreted as posterior risk, which is obtained by conditioning on the observations $\mathcal{Y}^N$ and integrating over the heterogeneous parameter λ^N and the shocks U_{T+1}^N. The outer expectation averages over the possible trajectories $\mathcal{Y}^N$. It is well known that the integrated risk is minimized by choosing the forecast that minimizes the posterior risk (with the understanding that we are conditioning on (θ, π) throughout) for each realization $\mathcal{Y}^N$.

Using the independence across i, the posterior risk can be written as follows:

$$\mathbb{E}_{\theta,\pi,\mathcal{Y}^N}^{\lambda^N,U_{T+1}^N}[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)] = \sum_{i=1}^N \left\{\left(\widehat{Y}_{iT+1} - \mathbb{E}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i,U_{iT+1}}[Y_{iT+1}]\right)^2 + \mathbb{V}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i,U_{iT+1}}[Y_{iT+1}]\right\}, \tag{6}$$

where $\mathbb{V}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i,U_{iT+1}}[\cdot]$ is the posterior predictive variance of Y_{iT+1}. The decomposition of the risk into a squared bias term and the posterior variance of Y_{iT+1} implies that the optimal predictor is the mean of the posterior predictive distribution. Because U_{iT+1} is mean-independent of λ_i and $\mathcal{Y}_i$, we obtain

$$\widehat{Y}_{iT+1}^{opt} = \mathbb{E}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i,U_{iT+1}}[Y_{iT+1}] = \mathbb{E}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] + \rho Y_{iT}. \tag{7}$$

The compound risk associated with the oracle forecast is

$$R_N^{\mathrm{opt}} = \mathbb{E}_\theta^{\mathcal{Y}^N}\left[\sum_{i=1}^N \left\{\mathbb{V}_{\theta,\pi,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] + \sigma^2\right\}\right]. \tag{8}$$

According to (8), the compound oracle risk has two components. The first component reflects uncertainty with respect to the heterogeneous coefficient λ_i and the second component captures uncertainty about the error term U_{iT+1}. Unfortunately, the direct implementation of the oracle forecast is infeasible because neither the parameter vector θ nor the correlated random effect distribution (or prior) $\pi(\cdot)$ are known. Thus, the oracle risk R_N^{opt} provides a lower bound for the risk that is attainable in practice.

2.3 Tweedie's Formula

The posterior mean $\mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ that appears in (7) can be evaluated using a formula which is named after the statistician Maurice Tweedie (though it had been previously derived by the astronomer Arthur Eddington). This formula is convenient for our purposes, because it expresses the posterior mean not as a function of the in practice unknown correlated random effects density $\pi(\lambda_i, y_{i0})$ but instead in terms of the marginal distribution of a sufficient statistic, which can be estimated from the cross-sectional information.

The contribution of unit i to the likelihood function associated with the basic dynamic panel data model in (2) is given by

$$p(y_i^{1:T}|y_{i0},\lambda_i,\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^T (y_{it} - \rho y_{it-1} - \lambda_i)^2\right\} \propto \exp\left\{-\frac{T}{2\sigma^2}\big(\hat{\lambda}_i(\rho) - \lambda_i\big)^2\right\}, \tag{9}$$

where $\propto$ denotes proportionality and the sufficient statistic $\hat{\lambda}_i(\rho)$ is

$$\hat{\lambda}_i(\rho) = \frac{1}{T}\sum_{t=1}^T (y_{it} - \rho y_{it-1}). \tag{10}$$

Using Bayes Theorem, the posterior distribution of λ_i can be expressed as

$$p(\lambda_i|y_i^{0:T},\theta) = p(\lambda_i|\hat{\lambda}_i, y_{i0},\theta) = \frac{p(\hat{\lambda}_i|\lambda_i, y_{i0},\theta)\pi(\lambda_i|y_{i0})}{\exp\big\{\ln p(\hat{\lambda}_i|y_{i0})\big\}}, \tag{11}$$

where $p(\hat{\lambda}_i|\lambda_i, y_{i0},\theta)$ is proportional to the right-hand side of (9).

To obtain a representation for the posterior mean, we now differentiate the equation $\int p(\lambda_i|\hat{\lambda}_i, y_{i0},\theta)d\lambda = 1$ with respect to $\hat{\lambda}_i$. Exchanging the order of integration and differentiation and using the properties of the exponential function, we obtain

$$\begin{aligned} 0 &= \frac{T}{\sigma^2}\int(\lambda_i - \hat{\lambda}_i)p(\lambda_i|\hat{\lambda}_i, y_{i0},\theta)d\lambda_i - \frac{\partial}{\partial\hat{\lambda}_i}\ln p(\hat{\lambda}_i|y_{i0},\theta) \\ &= \frac{T}{\sigma^2}\big(\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] - \hat{\lambda}_i\big) - \frac{\partial}{\partial\hat{\lambda}_i}\ln p(\hat{\lambda}_i|y_{i0},\theta). \end{aligned}$$

Solving this equation for the posterior mean yields Tweedie's formula:[3]

$$\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] = \hat{\lambda}_i(\rho) + \frac{\sigma^2}{T}\frac{\partial}{\partial\hat{\lambda}_i(\rho)}\ln p(\hat{\lambda}_i(\rho), y_{i0}). \tag{12}$$

Tweedie's formula was used by Robbins (1951) to estimate a vector of means λ^N for the model $Y_i|\lambda_i \sim N(\lambda_i, 1)$, $\lambda_i \sim \pi(\cdot)$, $i = 1,\dots,N$. Recently, it was extended by Efron (2011) to the family of exponential distributions, allowing for an unknown finite-dimensional parameter θ. The posterior mean takes the form of the sum of the sufficient statistic $\hat{\lambda}_i(\theta)$ and a correction term that captures the effect of the prior distribution of λ_i on the posterior. The correction term is expressed as a function of the marginal density of the sufficient statistic $\hat{\lambda}_i(\theta)$ conditional on Y_{i0} and θ. Thus, to evaluate the posterior mean, it is not necessary to explicitly solve a deconvolution problem that separates the prior density $\pi(\lambda_i|y_{i0})$ from the distribution of the error terms U_{it}.

Provided that the vector of homogeneous parameters θ is identifiable, it is possible to identify the cross-sectional distribution of the sufficient statistic $\hat{\lambda}(\theta)$. This, through Tweedie's formula, also identifies the posterior mean $\mathbb{E}^{\lambda}_{\theta,\pi,\mathcal{Y}}[\lambda]$ and leads to a unique Bayes predictor.

[3] We replaced the conditional log density $\ln p(\hat{\lambda}_i(\rho)|y_{i0})$ by the joint log density $\ln p(\hat{\lambda}_i(\rho), y_{i0})$ because the two differ only by a constant which drops out after the differentiation.

2.4 Implementation

We approximate the oracle forecast using an empirical Bayes approach that replaces the unknown objects θ and $p(\hat{\lambda}_i(\rho), Y_{i0})$ in (12) by estimates that exploit the cross-sectional information. A key requirement for an estimator of the homogeneous parameter θ is that it is consistent. In our basic dynamic panel data model, consistency can be achieved by various types of generalized method of moments (GMM) estimators, e.g., Arellano and Bond (1991), Arellano and Bover (1995), or Blundell and Bond (1998), or by a quasi-maximum-likelihood estimator (QMLE) that integrates out the heterogeneous λ_is under the misspecified correlated random effects distribution $\lambda_i | Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$.[4] The density $p(\hat{\lambda}_i(\rho), Y_{i0})$ could be estimated using kernel methods, a mixture approximation, or nonparametric maximum likelihood. In the next section, we provide an optimality result for the kernel estimator and we illustrate the performance of other estimators in a Monte Carlo study in Section 4.

3 Ratio Optimality

We now will prove that the predictor $\widehat{Y}^N_{T+1}$ that is constructed by replacing θ with a consistent estimator $\hat{\theta}$ and by replacing $p(\hat{\lambda}_i(\rho), Y_{i0})$ with a kernel density estimator achieves ϵ_0-ratio optimality uniformly for priors $\pi \in \Pi$. That is, for any $\epsilon_0 > 0$

$$\limsup_{N \to \infty} \sup_{\pi \in \Pi} \frac{R_N(\widehat{Y}^N_{T+1}; \pi) - R_N^{\text{opt}}(\pi)}{N \mathbb{E}^{\mathcal{Y}^N}_{\theta, \pi}\left[\mathbb{V}^{\lambda_i}_{\theta, \pi, \mathcal{Y}_i}[\lambda_i]\right] + N^{\epsilon_0}} \leq 0. \tag{13}$$

The convergence is uniform with respect to the correlated random effects distributions π in some set Π that we will characterize in more detail below. The uniformity holds in the neighborhood of point masses for which the prior and posterior variances of λ_i are zero. Thus, the convergence statement covers the case of λ_i being homogeneous across i.

First, consider the numerator in (13). The autoregressive coefficient in basic dynamic panel model can be $\sqrt{N}$-consistently estimated, which suggests that $\sum_{i=1}^N (\hat{\rho} - \rho)^2 Y_{iT}^2 = O_p(1)$. Thus, whether a predictor $\widehat{Y}_{iT+1}$ attains ratio optimality crucially depends on the rate at which the discrepancy between $\mathbb{E}^{\lambda_i}_{\theta, \pi, \mathcal{Y}_i}[\lambda_i]$ and $\widehat{\mathbb{E}}^{\lambda_i}_{\theta, \pi, \mathcal{Y}_i}[\lambda_i]$ vanishes, where the latter is a function of a nonparametric estimate of $p(\hat{\lambda}_i(\rho), Y_{i0})$. Second, note that the denominator of the ratio in (13) is strictly positive and divergent due to the N^{ϵ_0} term. The rate of divergence

[4] This estimator is described in more detail in Section 4 and its consistency is proved in the Online Appendix.

depends on the posterior variance of λ_i. If the posterior variance is strictly greater than zero, then the denominator is of order $O(N)$. Because the posterior is based on a finite number of observations T, the posterior variance is zero only if the prior density $\pi(\lambda)$ is a point mass. In this case the definition of ratio optimality requires that the regret vanishes at a faster rate, because the rate of the numerator drops from $O(N)$ to N^{ϵ_0}.[5]

The proof of the ratio-optimality result presented below in Theorem 3.7 below is a significant generalization of the proof in Brown and Greenshtein (2009), allowing for the presence of estimated parameters in the sufficient statistic $\hat{\lambda}(\cdot)$ and uniformity with respect to the correlated random effect density $\pi(\cdot)$, which is allowed to have a unbounded support. The remainder of this section is organized as follows. The kernel estimator of $p(\hat{\lambda}_i(\rho), Y_{i0})$ and the resulting formula for the predictor $\widehat{Y}_{iT+1}$ are presented in Section 3.1. In Section 3.2 we provide high-level assumptions that lead to the main theorem. In Section 3.3 we show that the high-level conditions are satisfied if $\pi(\cdot)$ belongs to a collection of finite-mixtures of Gaussian random variables with bounded means and variances.

3.1 Kernel Estimation and Truncation

To facilitate the theoretical analysis, we use a leave-one-out kernel density estimator of the form:

$$\hat{p}^{(-i)}(\hat{\lambda}_i(\rho), y_{i0}) = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0}-y_{i0}}{B_N}\right), \tag{14}$$

where $\phi(\cdot)$ is the pdf of a $N(0,1)$ and B_N is the kernel bandwidth. Using the fact that the observations are cross-sectionally independent and conditionally normally distributed, one can directly compute the expected value of the leave-one-out estimator:

$$\begin{aligned}\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] \quad = \quad & \int \frac{1}{\sqrt{\sigma^2/T+B_N^2}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T+B_N^2}}\right) \\ & \times\left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0}-\tilde{y}_{i0}}{B_N}\right)p(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}\right]p(\lambda_i)d\lambda_i.\end{aligned} \tag{15}$$

Taking expectations of the kernel estimator leads to a variance adjustment for the conditional distribution of $\hat{\lambda}_i|\lambda_i$ ($\sigma^2/T+B_N^2$ instead of σ^2/T) and the density of $Y_{i0}|\lambda_i$ is replaced by a

[5]If it were known that the λ_is are in fact homogeneous and the model is estimated with a common intercept, then $\sqrt{N}(\hat{\lambda}-\lambda) = O_p(1)$ and the regret in the numerator would be $O(1)$. Thus, the convergence result could also be achieved by standardizing with sequences that diverge at a rate slower than N^{ϵ_0}.

convolution. We define:

$$\begin{aligned} \pi_*(\lambda, y_0) &= \int \frac{1}{B_N}\phi\left(\frac{y_0-\tilde{y}_0}{B_N}\right)\pi(\lambda,\tilde{y}_0)d\tilde{y}_0 \\ p_*(\hat{\lambda}, y_0; \pi) &= \int \frac{1}{\sqrt{\sigma^2/T+B_N^2}}\phi\left(\frac{\hat{\lambda}-\lambda}{\sqrt{\sigma^2/T+B_N^2}}\right)\pi_*(\lambda,y_0)d\lambda, \end{aligned} \tag{16}$$

such that we can write

$$\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] = p_*(\hat{\lambda}_i, y_{i0}; \pi). \tag{17}$$

In view of the variance adjustment for the distribution of $\hat{\lambda}_i|\lambda_i$ induced by taking expectations of the kernel estimator, we replace the scale factor σ^2/T in the Tweedie correction term in (12) by $\hat{\sigma}^2/T + B_N^2$. Moreover, we truncate the absolute value of the posterior mean approximation from above. For $C > 0$ and for any $x \in \mathbb{R}$, define $[x]^C = \mathrm{sgn}(x)\min\{|x|, C\}$. The resulting predictor is

$$\widehat{Y}_{iT+1} = \left[\hat{\lambda}_i(\rho) + \left(\frac{\hat{\sigma}^2}{T} + B_N^2\right)\frac{\partial}{\partial\hat{\lambda}_i(\rho)}\ln\hat{p}^{-i}(\hat{\lambda}_i(\rho), Y_{i0})\right]^{C_N} + \hat{\rho}Y_{iT}, \tag{18}$$

where $C_N \longrightarrow \infty$ slowly.

3.2 Main Theorem

Let Π be a collection of joint densities $\pi(\lambda, y)$. The theoretical analysis relies heavily on slowly diverging sequences. To state the assumptions and prove the main theorem, the following definitions will be convenient:

Definition 3.1

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

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

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

Our first assumption controls the tails of the marginal distributions of λ and the initial condition Y_0.[6] We essentially assume that λ and Y_0 are subexponential random variables with finite fourth moments. The assumed tail probability and moment bounds are uniform for $\pi \in \Pi$.

Assumption 3.2 (Correlated Random Effects Distribution, Part 1) *There exist positive constants $M_1 < \infty$, $M_2 < \infty$, $M_3 < \infty$, and $M_4 < \infty$ such that for every $\pi \in \Pi$:*

(i) $\int_{|\lambda| \geq C} \pi(\lambda) d\lambda \leq M_1 \exp(-M_2(C - M_3))$ *and* $\int \lambda^4 \pi(\lambda) d\lambda \leq M_4$.

(ii) $\int_{|y_0| \geq C} \pi(y_0) dy_0 \leq M_1 \exp(-M_2(C - M_3))$ *and* $\int y_0^4 \pi(y_0) dy_0 \leq M_4$.

The proof of the main theorem relies on truncations. Assumption 3.3 imposes a lower bound and an upper bound on the diverging truncation sequences C_N and C'_N. These bounds on the truncation sequences constrain the rate at which the bandwidth B_N of the kernel density estimator has to shrink to zero. For technical reasons, the bandwidth cannot shrink as quickly as in typical density estimation problems.[7]

Assumption 3.3 (Trimming and Bandwidth)

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

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

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

We also need to impose a restriction on the conditional distribution of Y_0 given λ. First, we assume that the conditional density is bounded. Second, we regularize the shape of the conditional density $\pi(y_0|\lambda)$ by assuming that the density of as a function of y_0 should be uniformly "smooth" such that its convolution with the Kernel density remains close (in relative terms) to the original density. This assumption is required because we use a single bandwidth (independent of y_0) when estimating the density $p(\hat{\lambda}(\rho), y_0)$ for the Tweedie correction term. Thus, we rule out, for instance, that $\pi(y_0|\lambda)$ is a point mass. However, it is important to note that we do allow the marginal distribution $\pi(\lambda)$ to be a point mass (or discrete).

[6] To simplify the notation we write $\pi(\lambda)$ and $\pi(y_0)$ to denote the marginals of $\pi(\lambda, y_0)$.

[7] In a nutshell, we need to control the behavior of $\hat{p}(\hat{\lambda}_i, Y_{i0})$ and its derivative uniformly, which, in certain steps of the proof, requires us to consider bounds of the form M/B_N^2, where M is a generic constant. If the bandwidth shrinks too fast, the bounds diverge too quickly to ensure that it suffices to standardize the regret in Theorem 3.7 by N^ϵ if the λ_i coefficients are identical for each cross-sectional unit.

Assumption 3.4 (Correlated Random Effects Distribution, Part 2)

(i) There exists an $M < \infty$ such that $\pi(y_0|\lambda) \leq M$ for all $\pi \in \Pi$.

(ii) For given sequences C_N, C_N', and B_N satisfying Assumption 3.3, the conditional distribution $\pi(y_0|\lambda)$ satisfies

$$\sup_{|y_0| \leq C_N', |\lambda| \leq C_N, \pi \in \Pi} \left| \frac{\frac{1}{B_N} \int \phi\left(\frac{\tilde{y}_0 - y_0}{B_N}\right) \pi(\tilde{y}_0|\lambda) d\tilde{y}_0}{\pi(y_0|\lambda)} - 1 \right| = o(1).$$

Next, we impose some restrictions on the sampling distributions of the posterior means. To do so, define the posterior mean function as

$$m(\hat{\lambda}, y_0; \pi) = \hat{\lambda} + \left(\sigma^2/T\right) \frac{\partial \ln p(\hat{\lambda}, y_0; \pi)}{\partial \hat{\lambda}}, \tag{19}$$

where the joint sampling distribution of the sufficient statistic and the initial condition is given by

$$p(\hat{\lambda}, y_0; \pi) = \int \frac{1}{\sqrt{\sigma^2/T}} \phi\left(\frac{\hat{\lambda} - \lambda}{\sqrt{\sigma^2/T}}\right) \pi(y_0, \lambda) d\lambda.$$

In order to be precise about the uniformity requirements for $\pi \in \Pi$, we now included the prior density π as a conditioning argument in the functions $m(\cdot)$ and $p(\cdot)$. Moreover, we denote the posterior mean function under the $*$-distribution defined in (16) as

$$m_*(\hat{\lambda}, y_0; \pi, B_N) = \hat{\lambda} + \left(\sigma^2/T + B_N^2\right) \frac{\partial \ln p_*(\hat{\lambda}, y_0; \pi, B_N)}{\partial \hat{\lambda}}. \tag{20}$$

While the sampling distribution of $(\hat{\lambda}_i, Y_{i0})$ is tightly linked to the prior $\pi(\lambda, y_0)$ and the distribution of the sufficient statistic $\hat{\lambda}|\lambda \sim N(\lambda, \sigma^2/T)$, we find it convenient to postulate some high-level conditions, that we will verify for collections of finite mixtures of multivariate Normals (FNMN) in Section 3.3.

Assumption 3.5 (Posterior Mean Functions) *For a given sequence C_N satisfying Assumption 3.3, the posterior mean functions satisfy:*

(i) $N \int \int m(\hat{\lambda}, y_0; \pi)^2 \, \mathbb{I}\left\{ |m(\hat{\lambda}, y_0; \pi)| \geq C_N \right\} p(\hat{\lambda}, y_0; \pi) \, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+),$

(ii) $N \int \int m_*(\hat{\lambda}, y_0; \pi, B_N)^2 \, \mathbb{I}\left\{ |m_*(\hat{\lambda}, y_0; \pi, B_N)| \geq C_N \right\} p(\hat{\lambda}, y_0; \pi) \, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+),$

(iii) $N\int\int m(\hat{\lambda}, y_0; \pi)^2\, \mathbb{I}\left\{|m(\hat{\lambda}, y_0; \pi)| \geq C_N\right\} p_*(\hat{\lambda}, y_0; \pi^*, B_N)\, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+)$.

Finally, we impose moment restrictions on the sampling distribution of the estimators of the homogeneous parameters $\hat{\rho}$, and $\hat{\sigma}^2$.

Assumption 3.6 (Estimators of ρ and σ^2) *The estimators $\hat{\rho}$ and $\hat{\sigma}^2$ have the following properties: (i) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[|\sqrt{N}(\hat{\rho} - \rho)|^4\right] = o_{u,\pi}(N^+)$, (ii) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\hat{\sigma}^4\right] = o_{u.\pi}(N^+)$, and (iii) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[|\sqrt{N}(\hat{\sigma}^2 - \sigma^2)|^2\right] = o_{u.\pi}(N^+)$.*

An example of an estimator $\hat{\rho}$ that satisfies Assumption (3.6)(i) is the truncated instrumental variable (IV) estimator

$$\hat{\rho}_{IV} = \left[\left(\sum_{i=1}^{N}\sum_{t=2}^{T} Y_{it-2}\Delta Y_{it-1},\right)^{-1}\right]^{M_N} \left(\sum_{i=1}^{N}\sum_{t=2}^{T} Y_{it-2}\Delta Y_{it}\right),$$

where M_N is a sequence that slowly diverges to infinity. Define the residuals $\hat{U}_{it} = Y_{it} - -\hat{\lambda}(\hat{\rho}_{IV}) - \hat{\rho}_{IV}\hat{Y}_{it-1}$. Then, the sample variance of the $\hat{U}_{it}$'s is an estimator of σ^2 that satisfies Assumption (3.6).

We are now in a position to present the main result, which states that the regret associated with the vector of predictors $\widehat{Y}^N_{T+1}$, standardized by the posterior variance of the heterogeneous parameters λ_i, converges to zero as the cross-sectional dimension N of the sample becomes large.

Theorem 3.7 *Suppose that Assumptions 3.2 to 3.6 are satisfied. Then, in the basic dynamic panel model (2), the predictor $\widehat{Y}_{iT+1}$ defined in (18) achieves ϵ_0-ratio optimality uniformly in $\pi \in \Pi$, that is, for every $\epsilon_0 > 0$*

$$\limsup_{N\to\infty}\ \sup_{\pi\in\Pi} \frac{R_N(\widehat{Y}^N_{T+1};\pi) - R_N^{opt}(\pi)}{N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\mathbb{V}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]\right] + N^{\epsilon_0}} \leq 0. \tag{21}$$

A detailed proof of Theorem 3.7 can be found in the Online Appendix. We will provide a brief outline of the main steps subsequently. The statement in the theorem follows if we show

$$\limsup_{N\longrightarrow\infty}\ \sup_{\pi\in\Pi}\ \frac{N\mathbb{E}^{\mathcal{Y}^N,\lambda_i}_{\theta,\pi}\left[\left(\widehat{Y}_{iT+1} - (\lambda_i + \rho Y_{iT})\right)^2\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}^i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 1. \tag{22}$$

The numerator involves integration over the entire sample $\mathcal{Y}^N$ because the estimation of the Tweedie correction relies on cross-sectional information. The discrepancy between the predictor $\widehat{Y}_{iT+1}$ and $\lambda_i + \rho Y_{iT}$ can be decomposed into three terms:

$$\begin{aligned}
\widehat{Y}_{iT+1} - \lambda_i - \rho Y_{iT} &= \left[\hat{\lambda}_i(\hat{\rho}) + \left(\frac{\hat{\sigma}^2}{T} + B_N^2 \right) \frac{\partial}{\partial \hat{\lambda}_i(\hat{\rho})} \ln \hat{p}^{-i}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \right]^{C_N} - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) \\
&\quad + \big[m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i \big] + (\hat{\rho} - \rho) Y_{iT} \\
&= A_{1i} + A_{2i} + A_{3i}, \text{ say.}
\end{aligned}$$

We use the insight of Brown and Greenshtein (2009), that it is convenient to center the predictor $\widehat{Y}_{iT+1}$ at the posterior mean of Y_{iT+1} under the $*$-distribution – see (16) and (20) – because the leave-i-out cross-sectional averaging of the Tweedie correction generates an approximation of $m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)$; see (17).

In the above decomposition, the term A_{1i} captures the difference between the posterior mean of λ_i obtained from the estimated Tweedie correction (with the truncation) and the posterior mean of λ_i under the $*$-distribution. The term A_{2i} is the difference between the posterior mean $m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)$ of λ_i under the $*$-distribution defined in (16), and the "true" parameter λ_i. Finally, A_{3i} captures the direct effect of replacing ρ by $\hat{\rho}$. To prove the sufficient condition (22) of Theorem 3.7, we show that

$$\text{(i)} \quad N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\big[A_{1i}^2\big] = o_{u.\pi}(N^{\epsilon_0}), \quad \text{(ii)} \ \limsup_{N\longrightarrow\infty} \sup_{\pi\in\Pi} \frac{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\big[A_{2i}^2\big]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\big[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}^i}[\lambda_i])^2\big] + N^{\epsilon_0}} \le 1,$$

$$\text{(iii)} \quad N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\big[A_{3i}^2\big] = o_{u.\pi}(N^{+})$$

Result (iii) is relatively straightforward under Assumption 3.6. Result (ii) is based on the inequalities

$$\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{\theta,\pi}[(\lambda_i - m_i)^2] \le \mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{\theta,\pi}[(\lambda_i - m_{*i})^2], \quad \mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}[(\lambda_i - m_{*i})^2] \le \mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}[(\lambda_i - m_i)^2],$$

where $m_i = m(\hat{\lambda}_i, y_{i0}; \pi, B_N)$ and $m_{*i} = m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)$, and $\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}[\cdot]$ is the expectation operator associated with the $*-$ distribution defined in (16). The inequalities exploit the insight that the posterior means m_{*i} and m_i minimize the integrated risk under the $*$-distribution and the baseline distribution, respectively. Moreover, because $B_N \longrightarrow 0$ the discrepancies between expectations under the $*$-distribution and the benchmark distribution will vanish asymptotically.

Establishing Result (i) is the most complicated part of the proof. One can obtain the following inequality:

$$\begin{aligned}|A_{1i}| &\leq \left(\frac{\hat{\sigma}^2}{T} + B_N^2\right)\left|\frac{d\widetilde{p}_i^{(-i)} - dp_{*i}}{\widetilde{p}_i^{(-i)} - p_{*i} + p_{*i}} - \frac{dp_{*i}}{p_{*i}}\left(\frac{\widetilde{p}_i^{(-i)} - p_{*i}}{\widetilde{p}_i^{(-i)} - p_{*i} + p_{*i}}\right)\right| \\ &\quad + \left|\hat{\rho} - \rho\right|\left|\bar{Y}_{i,-1}\right| + \left|\frac{\hat{\sigma}^2}{T} - \frac{\sigma^2}{T}\right|\left|\frac{dp_{*i}}{p_{*i}}\right|,\end{aligned}$$

where

$$\begin{aligned}\widetilde{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}), \quad d\widetilde{p}_i^{(-i)} = \frac{1}{\partial\hat{\lambda}_i(\hat{\rho})}\partial\hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \\ p_{*i} &= p_*(\hat{\lambda}_i(\rho), Y_{i0}), \quad dp_{*i} = \frac{1}{\partial\hat{\lambda}_i(\rho)}\partial p_*(\hat{\lambda}_i(\rho), Y_{i0}).\end{aligned}$$

While the second and third term can be handled in a relatively straightforward manner, the first term is delicate. The $\hat{\lambda}_i(\hat{\rho})$ argument needs to be replaced by $\hat{\lambda}_i(\rho)$. Moreover, the bandwidth B_N of the kernel estimator has to be chosen to ensure that the numerators converge while the denominators are bounded away from zero using Bernstein's inequality.

3.3 Two Examples of Π

We now provide two specific examples of classes of distributions Π that satisfy Assumptions 3.2, 3.4, and 3.5.

Multivariate Normal Distributions. To simplify the notation we write y instead of y_0. We define the class Π of correlated random effects densities as

$$\Pi = \left\{\pi(\lambda, y) = \pi(\lambda)\pi(y|\lambda) \;\middle|\; \pi(\lambda) \in \Pi_\lambda,\ \pi(y|\lambda) \in \Pi_{y|\lambda}\right\} \tag{23}$$

where

$$\begin{aligned}\Pi_\lambda &: \left\{N(\mu_\lambda, \sigma_\lambda^2) \;\middle|\; |\mu_\lambda| \leq M_{\mu_\lambda},\ 0 \leq \sigma_\lambda^2 \leq M_{\sigma_\lambda^2}\right\} \\ \Pi_{y|\lambda} &: \left\{N(\alpha_0 + \alpha_1\lambda, \sigma_{y|\lambda}^2) \;\middle|\; |\alpha_0| \leq M_{\alpha_0},\ |\alpha_1| \leq M_{\alpha_1},\ 0 < \delta_{\sigma_{y|\lambda}^2} \leq \sigma_{y|\lambda}^2 \leq M_{\sigma_{y|\lambda}^2}\right\}\end{aligned}$$

We interpret $\underline{\sigma}_\lambda^2 = 0$ as a point mass and impose upper bounds on the mean and the variance of λ. To obtain joint normality of (λ, y) the conditional mean function for $y|\lambda$ is linear in λ and the conditional variance is constant. We bound the absolute values of the conditional mean parameters α_0 and α_1 as well as the conditional variance from above. The lower bound $\delta_{\sigma^2_{y|\lambda}}$ rules out a point-mass prior for $y|\lambda$. Let $\Pi = \Pi_\lambda \otimes \Pi_{y|\lambda}$.

Finite Mixtures of Multivariate Normals. This class of distribution is able to approximate a wide variety of distributions with exponential tails as the number of mixture components, K, increases. Formal approximation results are provided, for instance, in Norets and Pelenis (2012). Let $K < \infty$ be the maximum number of mixture components and define:

$$\Pi^{(K)}_{mix} = \left\{ \pi_{mix}(\lambda, y) = \sum_{k=1}^{K} \omega_k \pi_k(\lambda, y) \;\middle|\; \pi_k \in \Pi \,\forall k,\ 0 \leq \omega_k \leq 1,\ \sum_{k=1}^{K} \omega_k = 1 \right\}, \tag{24}$$

where the class Π is defined in (23).

Theorem 3.8 *Assumptions 3.2, 3.4, and 3.5 are satisfied by (i)* Π *in (23) and (ii)* $\Pi^{(K)}_{mix}$ *in (24).*

4 Monte Carlo Simulations

We now conduct three Monte Carlo experiments based on the basic dynamic panel data model in (2) to illustrate the performance of the empirical Bayes predictor and compare it to two alternative predictors.[8]

Empirical Bayes Predictors. The empirical Bayes predictors used in this section are based on the QMLE of θ. The QMLE is derived from the possibly misspecified distribution $\lambda_i|(Y_{i0}, \xi) \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$, where $\xi = (\phi_0, \phi_1, \underline{\Omega})$, with density $\pi_Q(\lambda_i|Y_{i0}, \xi)$:

$$\big(\hat{\theta}_{QMLE}, \hat{\xi}_{QMLE}\big) = \text{argmax}_{\theta,\xi} \prod_{i=1}^{N} \int p(y_i^{1:T}|y_{i0}, \lambda_i, \theta) \pi_Q(\lambda_i|Y_{i0}, \xi) d\lambda_i. \tag{25}$$

We show in the Online Appendix that the QMLE in our Monte Carlo designs is consistent under misspecification of $\pi_Q(\cdot)$.

The kernel estimator (14) of the Tweedie correction for which we developed the asymptotic theory in Section 3 is implemented as follows. Prior to the density estimation we standardize the sequences $\hat{\lambda}_i(\hat{\rho})$ and Y_{i0}, $i = 1, \ldots, N$. For the standardized series, we set

[8] In the working paper versions Liu, Moon, and Schorfheide (2016, 2018b) we report results for additional predictors, none of which dominate the ones considered here.

the bandwidth to $B_N = bB_N^*$, where $b \in \mathcal{B}$ and $\mathcal{B}$ is a finite grid for the scaling constant. The baseline value for the bandwidth as well as lower and upper bounds for b are given by

$$B_N^* = \left(\frac{4}{d+2}\right)^{1/(d+4)} \max\left\{\frac{1}{N^{1/(d+4)}}, \frac{1}{(\ln N)^{1.01}}\right\}, \quad \underline{b} \le b \le \bar{b}. \tag{26}$$

Here d corresponds to the dimension of the density that is being estimated and is either one or two. The scaling constant in front of the max operator as well as the first rate inside the max operator correspond to Silverman (1986)'s rule-of-thumb bandwidth choice. The second rate in the max operator is consistent with Assumption 3.3. We report forecast evaluation statistics for various choices of b as well as a $\hat{b}$ that is generated by minimizing the average forecast error loss for $b \in \mathcal{B}$ when predicting the last set of observations in the estimation sample, Y_{iT}, based on $Y_{i0}, \dots, Y_{iT-1}$.[9] By setting $\underline{B}_N = \underline{b}B_N^*$ and $\bar{B}_N = \bar{b}B_N^*$ we can deduce that the kernel-estimator with data-driven bandwidth scaling still satisfies Assumption 3.3.

QMLE Plug-In Predictor. This predictor takes the form $\widehat{Y}_{iT+1} = \hat{\lambda}_i(\hat{\rho}_{QMLE}) + \hat{\rho}_{QMLE} Y_{iT}$ and does not use the Tweedie correction.

Pooled-OLS Predictor. Ignoring the heterogeneity in the λ_is and imposing that $\lambda_i = \lambda$ for all i, we can define $(\hat{\rho}_P, \hat{\lambda}_P) = \operatorname{argmin}_{\rho,\lambda} \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \big(Y_{it} - \rho Y_{it-1} - \lambda\big)^2$. The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_P + \hat{\rho}_P Y_{iT}$.

4.1 Gamma-Distributed Random Effects

The design of the first experiment is summarized in Table 1. We assume that the λ_is follow a Gamma$(2, b)$ distribution and are uncorrelated with the initial condition Y_{i0} (random effects). The Gamma distribution has exponential tails and satisfies the tail bound condition in Assumption 3.2. We set $\rho = 0.8$, $\sigma^2 = 1$, and choose the parameter b to generate various values for $\mathbb{V}[\lambda_i]$. The number of time periods is $T = 4$. The subsequent results are based on $N_{sim} = 1,000$ Monte Carlo repetitions.

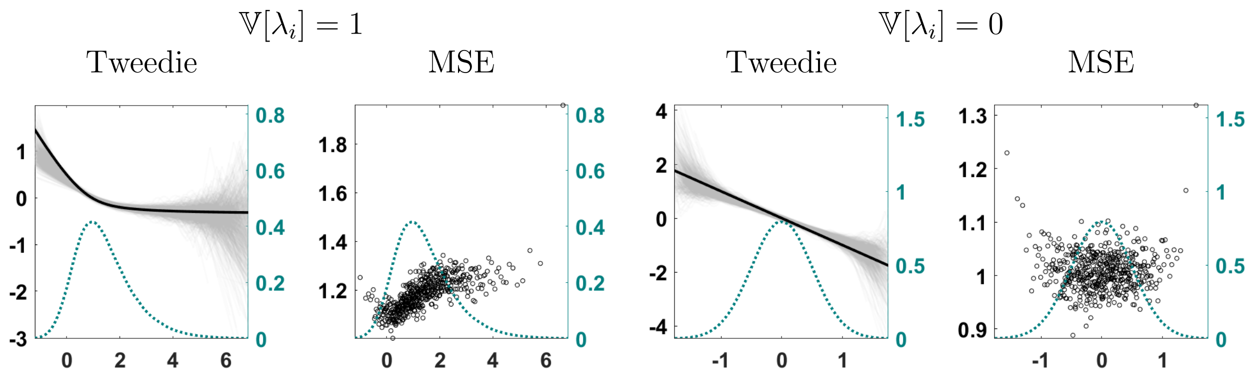
Kernel-Based Tweedie Corrections. We begin with a visual inspection of the Tweedie corrections for $N = 1,000$, comparing $\mathbb{V}[\lambda_i] = 1$ and $\mathbb{V}[\lambda_i] = 0$. In the first and third panels of Figure 1 we plot $(\sigma^2/T)\partial \ln p(\hat{\lambda}_i|\theta)/\partial\hat{\lambda}_i$ for the oracle forecast in (7) and estimates for the kernel-based empirical Bayes predictors with $b = \hat{b}$. Each hairline corresponds to an estimate from a particular Monte Carlo repetition. We overlay the density of $\hat{\lambda}_i$ to indicate

[9] To do so, we also re-estimate the homogeneous parameter θ based on the reduced sample $Y_{i0}, \dots, Y_{iT-1}$.

Table 1: Monte Carlo Design 1

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0, \gamma^2)$. $\rho = 0.8$, $\gamma = 1$
Initial Observations: $Y_{i0} \sim N(0, 1)$
Random effects: $\lambda_i \vert Y_{i0} \sim \text{Gamma}(2, b)$, various choices of b
Sample Size: $N = 1,000$, $T = 4$
Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

Figure 1: Tweedie Correction and Squared Forecast Error Losses as a Function of $\hat{\lambda}_i$



Notes: The Monte Carlo design is described in Table 1, $N = 1,000$. Tweedie: Solid (black) lines depict oracle Tweedie correction based on $p(\hat{\lambda}_i|y_{i0}, \theta)$ (scale on the left). Grey "hairs" depict estimates from the Monte Carlo repetitions. MSE: Each dot corresponds to a $\hat{\lambda}_i$ bin and the average squared-forecast error for observations assigned to that bin (scale on the left). The panels also show density estimates of the empirical distribution of $\hat{\lambda}_i(\rho)$ based on $N_{sim} \cdot N = 10^6$ simulations of $\hat{\lambda}_i(\rho)$ (scale on the right).

the likelihood of the various $\hat{\lambda}_i$ values on the x-axis. For $\mathbb{V}[\lambda_i] = 1$ the oracle correction is L-shaped. In the left tail of the $\hat{\lambda}_i$ distribution there is a lot of shrinkage to the prior mean and the correction is approximately linear with a large slope. In the right tail, the correction is essentially flat, meaning that for large values of $\hat{\lambda}_i$ (outliers) the optimal shrinkage is small in relative terms. For $\mathbb{V}[\lambda_i] = 0$ the $p(\hat{\lambda}_i|\theta)$ density is Normal and the oracle Tweedie correction is linear.

The second and fourth panels of Figure 1 depict forecast errors as a function of $\hat{\lambda}_i$ and overlays the density of $\hat{\lambda}_i$. The plots are generated as follows. For $N = 1,000$ we have a total of $N_{sim} \cdot N = 10^6$ forecasts across the 1,000 repetitions of the Monte Carlo experiment. We group the $\hat{\lambda}_i$'s into 500 bins such that each bin contains 2,000 observations. For each bin, we compute the mean squared forecast error, which leads to 500 pairs of bin location and forecast performance which are plotted in the figure. For $\mathbb{V}[\lambda_i] = 1$, the further $\hat{\lambda}$ is in

Table 2: Relative Regrets for Monte Carlo Design 1

		$\epsilon_0 = 0.7$				$\epsilon_0 = 0.2$			
	N	$\mathbb{V}[\lambda_i] = 1$	0.1	0.01	0	$\mathbb{V}[\lambda_i] = 1$	0.1	0.01	0
Empirical Bayes									
Kernel $b = \hat{b}$	100	.067	.097	.120	.129	.143	.333	.899	1.286
	1,000	.027	.043	.067	.073	.044	.117	.664	2.313
	10,000	.011	.021	.038	.045	.015	.040	.268	4.526
	100,000	.005	.009	.005	.029	.005	.013	.088	9.087
Kernel $b = 1.0$	100	.080	.154	.177	.189	.169	.529	1.326	1.889
	1,000	.056	.136	.173	.187	.093	.367	1.725	5.920
	10,000	.045	.128	.177	.205	.060	.245	1.261	2.507
	100,000	.029	.099	.044	.228	.034	.145	.714	72.214
Kernel $b = 1.5$	100	.062	.074	.093	.100	.130	.256	.702	1.004
	1,000	.028	.042	.059	.067	.047	.112	.594	2.108
	10,000	.014	.030	.045	.053	.019	.057	.320	5.328
	100,000	.007	.021	.010	.052	.008	.031	.163	16.536
Kernel $b = 2.0$	100	.077	.105	.144	.156	.163	.359	1.083	1.558
	1,000	.039	.051	.083	.094	.065	.136	.826	2.961
	10,000	.016	.021	.039	.047	.021	.040	.274	4.671
	100,000	.005	.010	.005	.028	.006	.014	.087	8.929
Plug-in Predictor	100	.194	.583	.856	.915	.411	1.999	6.431	9.155
	1,000	.249	.983	1.765	1.954	.417	2.652	17.622	61.785
	10,000	.304	1.410	3.302	3.929	.411	2.696	23.483	392.877
	100,000	.350	1.835	1.500	7.890	.412	2.686	24.171	2495.149
Pooled OLS	100	.804	.062	.024	.024	1.702	.212	.183	.238
	1,000	1.090	.060	.005	.004	1.823	.161	.047	.135
	10,000	1.364	.085	.002	.001	1.843	.163	.017	.098
	100,000	1.564	.111	.001	.000	1.842	.163	.011	.056

Notes: The Monte Carlo design is summarized in Table 1. The regret is standardized by the average posterior variance of λ_i and we consider ϵ_0 equal to 0.7 and 0.2; see Theorem 3.7.

the tails of its distribution, the larger the MSEs. This is the result of two effects. First, the Tweedie correction is less precisely estimated in the tails, because there are fewer realizations of $\hat{\lambda}_i$. Second, the population Tweedie correction in the right tail of the $\hat{\lambda}_i$ distribution is essentially flat. Thus, there is less shrinkage and the posterior mean has a higher sampling variance.

Table 2 summarizes the regret associated with each predictor relative to the posterior variance of λ_i, averaged over all trajectories $\mathcal{Y}^N$, as specified in Theorem 3.7. We report results for two choices of the constant ϵ_0 (0.7 and 0.2), four choices of cross-sectional heterogeneity $\mathbb{V}[\lambda_i]$ (1.0, 0.1, .01, and 0), and four different values for the number of cross-sectional units N (100, 1,000, 10,000 and 100,000).

The empirical Bayes estimators in Table 2 differ in the bandwidth used for the density estimation required to calculate the Tweedie correction. If the squared forecast error losses are averaged across all units, then a scaling of B_N^* between 1.5 and 2.0 minimizes the relative regret *ex post*. The *ex ante* selection of this scaling constant based on the pseudo-out-of-sample forecast of Y_{iT} works very well. We determine $\hat{b}$ by minimizing the regret over the interval $[1, 3]$ using a grid-spacing of 0.1. The regret obtained by setting $b = \hat{b}$ is very close to the regret with the *ex-post* optimal b. Thus, much of the subsequent discussion will focus on the results for $b = \hat{b}$.

Holding N fixed, the relative regret increases as $\mathbb{V}[\lambda_i]$ decreases to zero because the numerator of the relative regret formula in Theorem 3.7 is decreasing in $\mathbb{V}[\lambda_i]$. Our asymptotic theory predicts that the regret associated with the empirical Bayes predictors should converge to zero as $N \longrightarrow \infty$. Holding $\mathbb{V}[\lambda_i]$ fixed, the table confirms that except for the case of $\epsilon_0 = 0.2$ and $\mathbb{V}[\lambda_i] = 0$, the relative regret decreases with the sample size N. Our theory also predicts that this convergence is uniform in $\mathbb{V}[\lambda_i]$, meaning that for sequences of $N \longrightarrow \infty$ and $\mathbb{V}[\lambda_i] \longrightarrow 0$ we should also observe that the relative regret vanishes. Consider $\epsilon_0 = 0.7$ and the sequence of $(N, \mathbb{V}[\lambda_i])$ given by

$$(100,\ 1), \quad (1000,\ 0.1), \quad (10000,\ 0.01), \quad (100000,\ 0). \tag{27}$$

Along this sequence, the relative regret is clearly decreasing for $b = \hat{b}$ and $b = 2.0$. According to the asymptotic theory of Section 3, the bandwidth of the kernel estimator needs to be larger than in standard density estimation problems to achieve ratio optimality. Setting $b = 1$ corresponds to the density-estimation bandwidth choice because for the sample sizes considered $N^{-1/5} > (\ln N)^{-1.01}$. Thus a choice of $b = 2$ can be interpreted as oversmoothing, which improves the forecasting performance.

For the suboptimal choices of $b = 1.5$ the relative regret first falls and then stays approximately constant; and for $b = 1.0$ the regret is increasing along the sequence (27). The large sample behavior of the regret is linked to the divergence rates of the sequences $\ln N$ and N^{ϵ_0}. Asymptotically, $\ln N / N^{\epsilon_0} \longrightarrow 0$ for any $\epsilon_0 > 0$. However, if ϵ_0 is small, an N much larger than the sample sizes in the Monte Carlo simulation is required for $N^{\epsilon_0} > \ln N$. For this reason, we do not observe a monotonically decreasing regret along all $(N, \mathbb{V}[\lambda_i])$ sequences. This problem becomes more pronounced for $\epsilon_0 = 0.2$, where regrets tend to increase along the sequence (27).[10]

[10]Note that $\ln(100,000) \approx 11.5$ whereas $100,000^{0.2} = 10$.

The last eight rows of Table 2 contain relative regrets for the plug-in predictor and the pooled-OLS predictor. The plug-in predictor is clearly dominated by the empirical Bayes predictors and its regret increases with sample size N, holding $\mathbb{V}[\lambda_i]$ fixed, for both choices of ϵ_0. It also increases along the sequence (27). The relative performance of the pooled-OLS predictor depends on the level of heterogeneity captured by $\mathbb{V}[\lambda_i]$. For $\mathbb{V}[\lambda_i] = 1$, the pooled-OLS predictor is dominated by the empirical Bayes predictors and exhibits a regret that is increasing in N. On the other hand, if the population is near-homogeneous, i.e., $\mathbb{V}[\lambda_i]$ is 0.01 or 0, then its regret is smaller than the regret associated with the empirical Bayes predictor.

We conclude that the predictions of the theory are confirmed by the Monte Carlo experiment, with the caveat that small values of ϵ_0 require very large sample sizes for the relative regret to vanish as $\mathbb{V}[\lambda_i] \longrightarrow 0$. The empirical Bayes predictor clearly dominates the plug-in predictor. If there is no heterogeneity in the population, then, not surprisingly, imposing coefficient homogeneity through pooled OLS estimation, leads, in finite samples, to sharper forecasts. However, even for a small level of heterogeneity, e.g., $\mathbb{V}[\lambda_i] = 0.1$, the empirical Bayes estimator is associated with lower regrets for samples sizes of $N = 1,000$ or larger.

Alternative Estimates of the Tweedie Correction. We include two additional estimators of the Tweedie correction in the Monte Carlo. First, Gu and Koenker (2017a) proposed to estimate the density of the sufficient statistic by nonparametric maximum likelihood estimation (NP MLE), which can be implemented using the GLmix function in their REBayes package. The estimator is constructed by specifying bounds for the domain of λ_i and partition it into $K = 300$ (default setting) bins. Let λ_k be the right endpoint of bin k and Δ_k its width. Moreover, let $\tilde{\omega}_k$ be the probability associated with the k'th bin and define $\omega_k = \tilde{\omega}_k \Delta_k$. Then

$$p_{NP}\big(\hat{\lambda}_i \big| \{\omega_k\}_{k=1}^K\big) = \sum_{k=1}^K \omega_k p_N(\hat{\lambda}_i|\lambda_k, \sigma^2/T), \quad \omega_k \geq 0, \quad \sum_{k=1}^K \omega_k = 1.$$

The bin probabilities are estimated by maximizing the log likelihood function and the Tweedie corrections is computed conditional on the estimates $\hat{\omega}_k$.

Second, we consider a finite mixture of normal distributions to approximate $p(\hat{\lambda}_i(\hat{\rho}), Y_{i0})$:

$$p_{mix}\big(\hat{\lambda}_i, Y_{i0} \big| \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^K\big) = \sum_{k=1}^K \omega_k p_N(\hat{\lambda}_i, Y_{i0}|\mu_k, \Sigma_k), \quad \omega_k \geq 0, \quad \sum_{k=1}^K \omega_K = 1,$$

where $p_N(\cdot)$ is the density of a $N(\mu_k, \Sigma_k)$ random variables. The mixture probabilities as

Table 3: Relative Regrets for Alternative Tweedie Corrections for Monte Carlo Design 1

	$\mathbb{V}[\lambda_i] = 1$	0.1	0.01	0
Kernel $b = \hat{b}$	.027	.043	.067	.073
Mixture $K = \hat{K}$	.022	.019	.010	.010
NP MLE	.018	.013	.011	.011

Notes: The design of the experiment is summarized in Table 1. We report results for $N = 1,000$. The regret is standardized by the average posterior variance of λ_i and we set $\epsilon_0 = 0.7$; see Theorem 3.7.

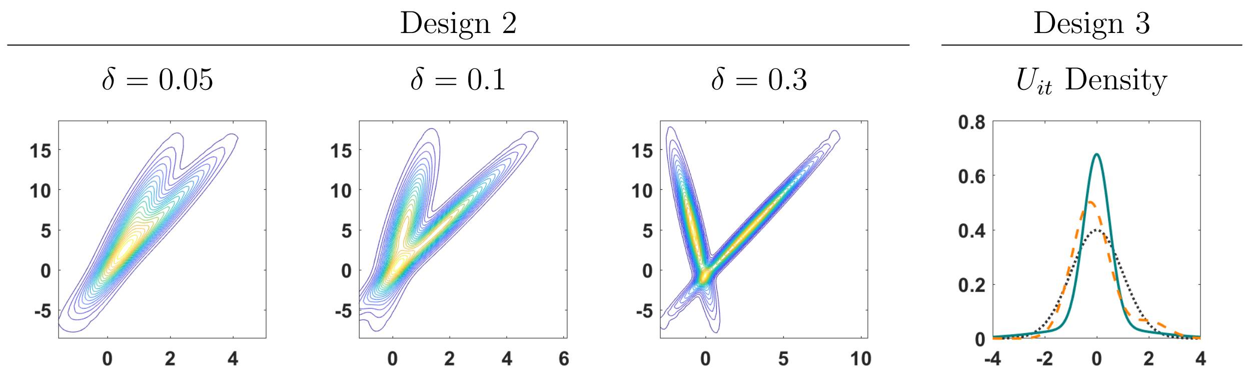
well as the means and covariance matrices are estimated by maximizing the log likelihood function using an EM algorithm. The Tweedie corrections are then computed conditional on the estimates $(\hat{\omega}_k, \hat{\mu}_k, \hat{\Sigma}_k)$. We report forecast evaluation statistics for the $\hat{K}$ that is obtained by minimizing the average forecast error loss for $1 \leq K \leq \bar{K}$ when predicting the last set of observations in the estimation sample, Y_{iT}, based on $Y_{i0}, \ldots, Y_{iT-1}$.

Table 3 provides relative regrets for the three different estimates of the Tweedie correction. The mixture approximation of $p(\hat{\lambda}_i)$ leads to a Tweedie correction that tends to generate more accurate forecasts than the kernel-based correction, in particular for small $\mathbb{V}[\lambda_i]$. The selection of the number of mixture components, K, based on pseudo-out-of-sample forecast errors for period T also works well. For large values of $\mathbb{V}[\lambda_i]$ when the distribution of $\hat{\lambda}_i$ has a long right tail, we select a large value of K, whereas for small values of $\mathbb{V}[\lambda_i]$ when the distribution of $\hat{\lambda}_i$ is approximately normal, we select a small value of K.

The NP MLE of $p(\hat{\lambda}_i)$ proposed by Gu and Koenker (2017a) is similar to the mixture estimator in this experiment. While the regret differentials, say, between the kernel-based correction versus the mixture-based or NP-MLE correction appear to be large, overall, in terms of MSE, the forecast performance of the various estimators is very similar. For instance, for $\mathbb{V}[\lambda_i] = 1$ the average MSEs for kernel $b = \hat{b}$, mixture $K = \hat{K}$, and NP MLE are 1.185, 1.184, and 1.183, respectively.

To summarize, the kernel estimators of the Tweedie correction lead to a predictor that has ratio optimality in the sense of Theorem 3.7 and performs well in the Monte Carlo experiment. In particular, the empirical Bayes predictor dominates the naive plug-in predictor and pooled OLS (unless coefficients are homogeneous). Our simulations also indicate that in finite samples the performance of the empirical Bayes predictor can be improved by replacing the kernel estimator with an NP MLE or a mixture estimator. The large sample analysis of these two alternative predictors is beyond the scope of this paper. Some results for NP MLE

Figure 2: Monte Carlo Designs 2 and 3



Notes: Design 2: The plots depict contours of the density $\pi(\lambda_i, Y_{i0}|\delta)$ for different settings for the hyperparameter δ; x-axis is λ_i and y-axis is y_{i0}. Design 3: The plot overlays a $N(0,1)$ density (black, dotted), the scale mixture (teal, solid), and the location mixture (orange, dashed). All three densities have $\mathbb{E}[U_{it}] = 0$, $\mathbb{V}[U_{it}] = 1$. Formulas are provided in the Online Appendix.

in the context of the basic vector-of-means (or Gaussian denoising) problem are available in Jiang and Zhang (2009) and Saha and Guntuboyina (2019).

4.2 Alternative Monte Carlo Designs

We consider two additional Monte Carlo designs and describe in which dimensions they differ from the experiment described in Table 1. One of them, which we refer to as Design 2, is based on a correlated random effects distribution for λ_i and the initial condition Y_{i0}. The first three panels of Figure 2 show contours of the joint density of $\pi(\lambda_i, y_{i0})$ for three choices of a hyperparameter δ. The precise formula for the densities is provided in the Online Appendix. The contours have the shape of a pair of scissors that opens up as δ increases from 0.05 to 0.3. For large values of δ the marginal densities of λ_i and y_{i0} become multi-modal.

Design 3 features non-Gaussian innovations U_{it}, which means that the Gaussian likelihood function, that is the basis for the posterior mean predictor and the Tweedie correction, is misspecified. We consider a scale mixture that generates excess kurtosis and a location mixture that generates skewness. The innovation distributions are normalized such that $\mathbb{E}[U_{it}] = 0$ and $\mathbb{V}[U_{it}] = 1$ and their densities are plotted in the rightmost panel of Figure 2. For (λ_i, Y_{i0}) we continue to use the correlated random effects distribution of Design 2 with $\delta = 0.1$.

Table 4: Relative Regrets for Monte Carlo Designs 2 and 3

	Design 2			Design 3	
U_{it} Distribution	Normal	Normal	Normal	Scale Mix.	Loc. Mix.
CRE Distribution	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.1$
Empirical Bayes					
Kernel $b = \hat{b}$	.106	.112	.188	.411	.259
Mixture $K = \hat{K}$	.027	.027	.025	.306	.154
Plug-in Predictor	.611	.515	.709	.970	.744
Pooled OLS	.096	.392	2.17	.786	.583

Notes: The Monte Carlo designs are illustrated in Figure 2. We report results for $N = 1,000$. The regret is standardized by the average posterior variance of λ_i and we set $\epsilon_0 = 0.7$; see Theorem 3.7.

The results are similar to Design 1 and summarized in Table 4.[11] The plug-in predictor is clearly dominated by the empirical Bayes predictors. The empirical Bayes predictors also beat the pooled OLS predictor by a significant margin for $\delta = 0.1$ and $\delta = 0.3$. At first glance the good performance of pooled OLS for $\delta = 0.05$ is surprising in view of the heterogeneity in λ_i implied by the Monte Carlo design. It turns out that under pooled OLS $\hat{\lambda}$ is close to zero and $\hat{\rho}$ is approximately one. Thus, the predictor essentially generates no-change forecasts that perform quite well.

For $\delta = 0.05$ and $\delta = 0.1$ the *ex post* optimal bandwidth scaling b is close to one, which is picked up by our bandwidth selection procedure based on pseudo-out-of-sample forecasts. For $\delta = 0.3$ the optimal scaling is approximately 0.5. This is qualitatively plausible, because the scissor-shape density of $(\hat{\lambda}_i, Y_{i0})$ has a relatively large overall variance, which translates into a large baseline bandwidth, but at the same time it has sharp peaks in the modal regions which require a small bandwidth. As in Experiment 1, the performance of the empirical Bayes estimator can be improved by replacing the kernel estimate of the correction with a mixture estimate, in part because the latter is more stable in the tails of the $(\hat{\lambda}_i, Y_{i0})$ distribution and in part because the density of $p(\hat{\lambda}_i, Y_{i0})$ is in fact a mixture of normal distributions.

The last two columns of Table 4 provide results under a misspecified likelihood function and can be compared to the results reported in the second column which are based on normally-distributed error terms U_{it}. The QMLE estimator of θ remains consistent under the likelihood misspecification. However, the (non-parametric) Tweedie correction no longer delivers a valid approximation of the posterior mean. Accordingly, the regrets under mixture

[11] The bandwidth scaling constant b is selected over the range $0.1, 0.2, \ldots, 1.9$. We dropped the NP MLE, because its performance under Design 1 was similar to the mixture estimator and Gu and Koenker (2017a)'s software is not written for a correlated random-effects specification.

innovations are generally higher than under the normal innovations. However, in comparison to the plug-in and pooled-OLS predictors the empirical Bayes predictors continue to perform very well and attain relative regrets that are more than 50% smaller than the regrets associated with the best competitors.

5 Extensions

Multi-Step Forecasting. While this paper focuses on single-step forecasting, we briefly discuss in the context of the basic dynamic panel data model how the framework can be extended to multi-step forecasts. Under the assumption that the oracle knows ρ and $\pi(\lambda_i, Y_{i0})$ we can express the oracle forecast as

$$\widehat{Y}^{opt}_{iT+h} = \left(\sum_{s=0}^{h-1} \rho^s\right) \mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i] + \rho^h Y_{iT}.$$

As in the case of the one-step-ahead forecasts, the posterior mean $\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i]$ can be replaced by an approximation based on Tweedie's formula and the ρ's can be replaced by consistent estimates. A model with additional covariates would require external multi-step forecasts of the covariates, or the specification in (1) would have to be modified such that all exogenous regressors appear with an h-period lag.

Tweedie's Formula (Generalization). The general model (1) distinguishes three types of regressors. First, the vector W_{it} interacts with the heterogeneous coefficients λ_i. In addition to a constant, we allow W_{it} to also include deterministic time effects such as seasonality, time trends and/or strictly exogenous variables observed at time t. To distinguish deterministic time effects $w_{1,t+1}$ from cross-sectionally varying and strictly exogenous variables $W_{2,it}$, we partition the vector into $W_{it} = (w_{1,t+1}, W_{2,it})$.[12] Second, X_{it} is a vector of predetermined predictors with homogeneous coefficients. Because the predictors X_{it} may include lags of Y_{it+1}, we collect all the predetermined variables other than the lagged dependent variable into the subvector $X_{2,it}$. Third, Z_{it} is a vector of strictly exogenous regressors, also with common coefficients. Let $H_i = (X_{i0}, W^{0:T}_{2,i}, Z^{0:T}_i)$.

Heteroskedasticity can be introduced as follows:

$$U_{it} = \sigma_t V_{it} = \varsigma(H_i, \gamma_t) V_{it}, \quad V_{it} \mid (Y^{1:t-1}_i, X^{1:t-1}_i, H_i, \lambda_i) \sim N(0,1), \tag{28}$$

[12]Because W_{it} is a predictor for Y_{it+1} we use a $t+1$ subscript for the deterministic trend component w_1.

where $\varsigma(\cdot)$ is a parametric function indexed by the (time-varying) finite-dimensional parameter γ_t. We allow $\varsigma(\cdot)$ to be dependent on the initial condition of the predetermined predictors, X_{i0}, and other exogenous variables. Because the time dimension T is assumed to be small, the dependence through X_{i0} can generate a persistent ARCH effect. Note that even in the homoskedastic case $\sigma_t = \sigma$, the distribution of Y_{it} given the regressors is non-normal because the distribution of the λ_i parameters is fully flexible.

Let $\gamma = [\gamma_1', \ldots, \gamma_T']$, $\theta = [\alpha', \rho', \gamma']'$, $\tilde{y}_t(\theta) = y_t - \rho' x_{t-1} - \alpha' z_{t-1}$, and $\Sigma(\theta, h) = \mathrm{diag}\big(\sigma_1^2, \ldots, \sigma_T^2\big)$ (the i subscripts are dropped to simplify the notation). Moreover, let $\tilde{y}(\theta)$ and w be the matrices with rows $\tilde{y}_t(\theta)$ and w_{t-1}', $t = 1, ..., T$. To generalize Tweedie's formula, we re-define the sufficient statistic $\hat{\lambda}$ as follows:

$$\hat{\lambda}(\theta) = \big(w'\Sigma^{-1}(\theta, h)w\big)^{-1} w'\Sigma^{-1}(\theta, h)\tilde{y}(\theta). \tag{29}$$

Using the same calculations as in Section 2.3, it can be shown that the posterior mean of λ_i has the representation

$$\mathbb{E}^{\lambda}_{\theta,\pi,\mathcal{Y}}[\lambda] = \hat{\lambda}(\theta) + \bigg(W^{0:T-1\prime}\Sigma^{-1}(\theta, H)W^{0:T-1}\bigg)^{-1} \frac{\partial}{\partial\hat{\lambda}(\cdot)} \ln p\big(\hat{\lambda}(\cdot), H|\theta\big). \tag{30}$$

The existing panel data model literature has developed numerous estimators for the homogeneous parameters that could be used to generate a $\hat{\theta}$ for the general model (1). In principle one can proceed with the estimation of the Tweedie correction as for the basic model. However, the larger the set of conditioning variables H_i the more difficult it becomes to estimate $p(\hat{\lambda}_i(\theta, H_i), H_i|\theta)$ precisely.

6 Empirical Application

We will now use the previously-developed predictors to forecast pre-provision net revenues of bank holding companies (BHC). The stress tests that have become mandatory under the 2010 Dodd-Frank Act require banks to establish how PPNRs vary in stressed macroeconomic and financial scenarios. A first step toward building and estimating models that provide trustworthy projections of PPNRs and other bank-balance-sheet variables under hypothetical stress scenarios, is to develop models that generate reliable forecasts under the observed macroeconomic and financial conditions. Because of changes in the regulatory environment in the aftermath of the financial crisis as well as frequent mergers in the banking industry our large N small T panel-data-forecasting framework is particularly attractive for stress-test applications.

We generate a collection of panel data sets in which PPNR as a fraction of consolidated assets (the ratio is scaled by 400 to obtain annualized percentages) is the dependent variable. The quarterly data sets are based on the FR Y-9C consolidated financial statements for bank holding companies for the years 2002 to 2014. We construct rolling samples that consist of $T+2$ observations, where T is the size of the estimation sample and varies between $T=4$ and $T=10$ quarters. The additional two observations are used, respectively, to initialize the lag in the first period of the estimation sample and to compute the error of the one-step-ahead forecast. For instance, with data from 2002:Q1 to 2014:Q4 we can construct $M=45$ samples of size $T=6$ with forecast origins running from $\tau=$ 2003:Q3 to $\tau=$ 2014:Q3. Each rolling sample is indexed by the pair (τ, T). The cross-sectional dimension N varies from sample to sample and ranges from 613 to 920.

We discuss the accuracy of baseline forecasts for various model specifications and predictors in Section 6.1 and compare the baseline predictions to predictions under stressed macroeconomic and financial conditions in Section 6.2. The Online Appendix provides further details about the data set.

6.1 Baseline Forecast Results

The forecast evaluation criterion is the mean-squared error (MSE) computed across institutions:

$$\text{MSE}(\widehat{Y}^N_{\tau+1}) = \frac{\frac{1}{N_\tau}\sum_{i=1}^{N_\tau}\left(Y_{i\tau+1} - \widehat{Y}_{i\tau+1}\right)^2}{\frac{1}{N_\tau}\sum_{i=1}^{N_\tau}}. \tag{31}$$

We consider four predictors. The first two predictors are empirical Bayes predictors based on $\hat{\theta}_{QMLE}$. The Tweedie corrections are generated either using a kernel estimator with $b=\hat{b}$ or a mixture estimator with $K=\hat{K}$. As in Section 4, we also report results for the plug-in predictor and the pooled-OLS predictor. We estimate three model specifications. The first model is the basic dynamic panel data model in (2). The other two specifications include additional covariates that reflect aggregate macroeconomic and financial conditions, which we will use subsequently to generate counterfactual forecasts under stress scenarios. We assume that the banks' exposure to the aggregate condition is heterogeneous and include these predictors into the vector W_{it-1}, using the notation in (1). The second model includes the unemployment rate as an additional predictor. The third model includes the unemployment rate, the federal funds rate, and an interest rate spread.[13] When analyzing stress scenarios, one is typically interested in the effect of stressed economic conditions on the current performance of the banking sector. For this reason, we are changing the timing convention slightly and include the time t macroeconomic and financial variables into the vector W_{it-1}.

Figure 3 depicts MSE differentials relative to the MSE of the plug-in predictor:

$$\Delta(\widehat{Y}^N_{\tau+1}) = \frac{\text{MSE}(\widehat{Y}^N_{\tau+1}) - \text{MSE}(\text{plug-in})}{\text{MSE}(\text{plug-in})},$$

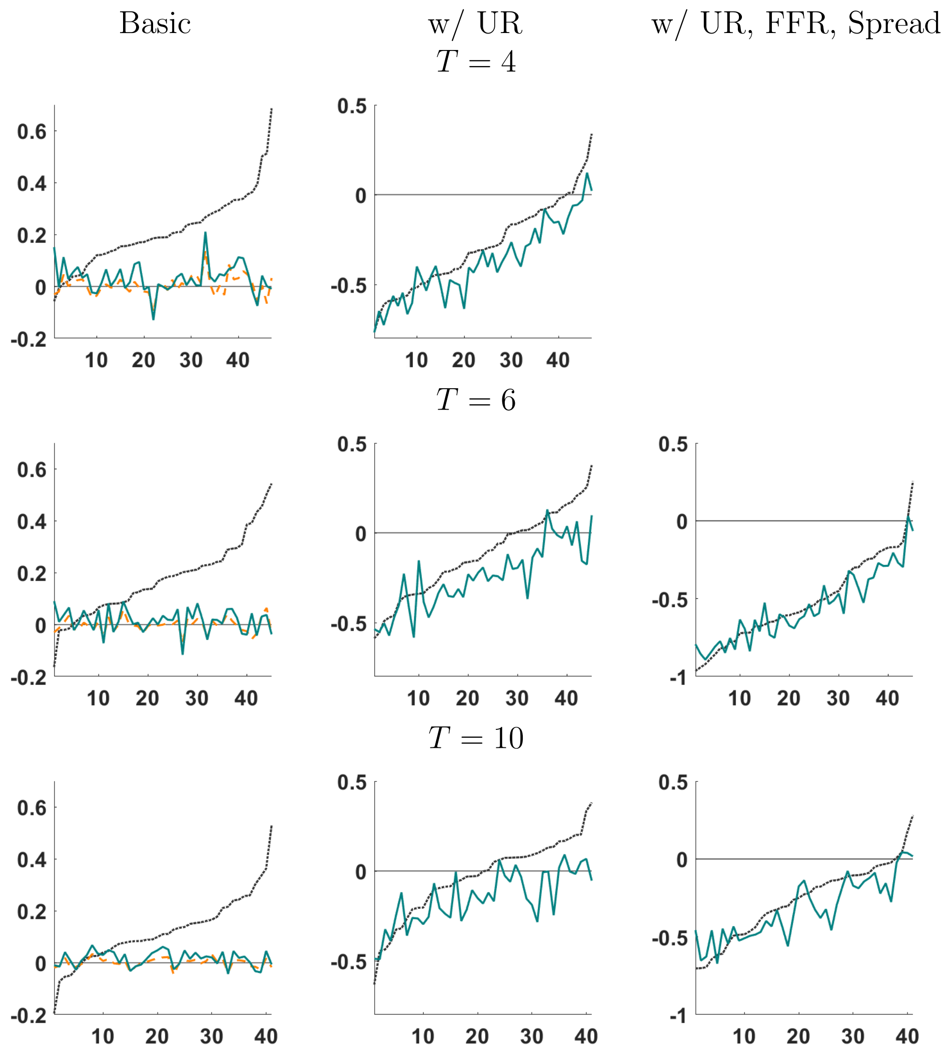
where $\text{MSE}(\widehat{Y}^N_{\tau+1})$ is defined in (31) and for now we set the selection operator $D_i(\mathcal{Y}_{i\tau}) = 1$, meaning we are averaging over all banks. If $\Delta(\widehat{Y}^N_{\tau+1}) < 0$, then the predictor $\widehat{Y}^N_{\tau+1}$ is more accurate than the plug-in predictor. The three columns correspond to the three different forecast models under consideration and the three rows correspond to the sample sizes $T = 4$, $T = 6$, and $T = 10$, respectively. In the x-dimension, the MSE differentials are not arranged in chronological order by τ. Instead, we sort the samples based on the magnitude of the MSE differential for the pooled OLS predictor.

The plug-in predictor (the zero lines in Figure 3) and the pooled-OLS predictor (black dotted lines) provide natural benchmarks for the assessment of the empirical Bayes predictors. The former is optimal if the heterogeneous coefficients are essentially "uniformly" distributed in $\mathbb{R}^{k_w}$, whereas the latter is optimal in the absence of heterogeneity. The plug-in predictor dominates the pooled-OLS predictor whenever the number of heterogeneous coefficients is small relative to the time series dimension, which is the case for the basic model. For the unemployment model with $T = 4$ and the model with three covariates the pooled OLS predictor is more accurate than the plug-in predictor. For the model with unemployment only, the ranking is sample dependent.

The empirical Bayes procedure works generally well, in that it is adaptive: for most samples the empirical Bayes predictor is at least as accurate as the better of the plug-in and the pooled-OLS predictor. The unemployment-rate model provides a nice illustration of this adaptivity. In panels (2,2) and (3,2) the fraction of samples in which the plug-in predictor dominates the pooled-OLS predictor ranges from 1/3 to 1/2. In all of these samples the MSE differential for the empirical Bayes predictor is close to zero or below zero. In the remaining

[13] All three series are obtained from the FRED database maintained by the Federal Reserve Bank of St. Louis: Unemployment is UNRATE, the effective federal funds rate is EFFR, and the spread between the federal funds rate and the 10-year treasury bill is T10YFF. We use temporal averaging to convert high-frequency observations into quarterly observations.

Figure 3: Percentage Change in MSE Relative to Plug-in Predictor, All Banks

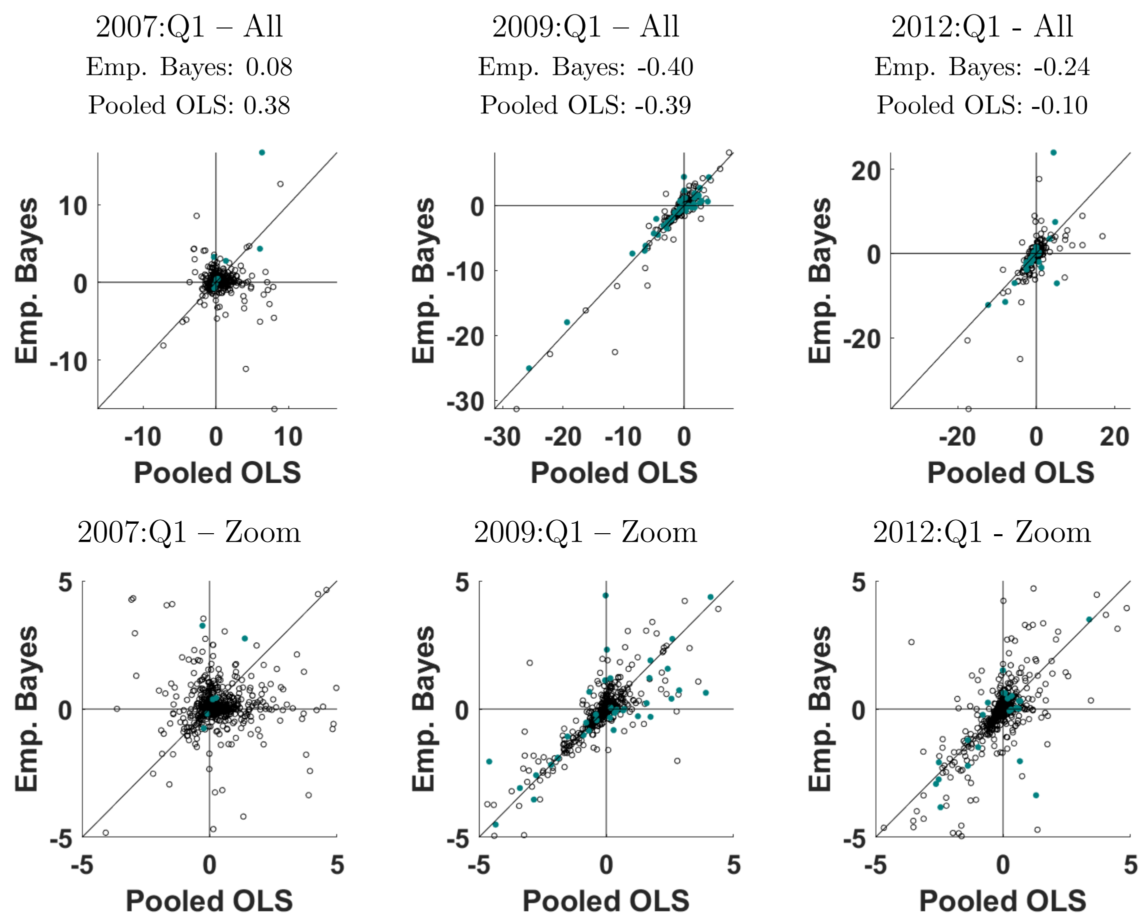


Notes: Benchmark is the plug-in predictor. y-axis shows percentage changes in MSE, whereby a negative value is an improvement compared to the plug-in predictor. Time periods are sorted such that the relative MSE of pooled OLS is monotonically increasing. Comparison to: (i) empirical Bayes predictor with kernel estimator ($b = \hat{b}$), dashed orange; (ii) empirical Bayes predictor with mixture estimator ($K = \hat{K}$), solid teal; (iii) pooled OLS, dotted black.

samples the MSE differential of the empirical Bayes predictor tends to be smaller than the MSE differential associated with pooled OLS, highlighting that the shrinkage induced by the estimated correlated random effects distribution improves on the two benchmark procedures.

For the basic panel data model, we report results for two versions of the empirical Bayes predictor: one is based on the kernel estimation of the Tweedie correction term with $b = \hat{b}$ and the other one is based on the mixture estimation with $K = \hat{K}$. Here the dimension

Figure 4: Squared Forecast Error Differentials Relative To Plug-in, Model w/ UR, $T = 6$



Notes: Figure depicts scatter plots of scaled squared forecast error differentials for pooled OLS and empirical Bayes with mixture estimator ($K = \hat{K}$) relative to the plug-in predictor. The differentials are divided by the average MSE (across all units) of the plug-in predictor. Cross-sectional averaging of the dots yields the values (%) that are listed above the plots and depicted in panel (2,2) of Figure 3 for the corresponding time periods. Negative values are improvements compared to the plug-in predictor. Teal dots indicate banks for which $y_{i\tau} \leq 0$. The thin solid lines correspond to zero lines and the 45-degree line, respectively.

of the density that needs to be estimated to construct the correction is equal to two and the kernel and mixture estimation perform approximately equally well. For the models with covariates, a higher-dimensional density needs to be estimated, and the mixture estimation approach works generally better. In fact, for some of the samples the kernel estimates were quite erratic, which is why in columns 2 and 3 of Figure 3 we only report results for the mixture-based predictor.

Figure 4 examines the bank-specific squared forecast error loss differentials (subtracting the squared forecast error associated with the plug-in predictor) for the unemployment model. Each dot corresponds to a bank. We standardize the squared forecast error loss differentials by the MSE of the plug-in predictor, i.e., we are plotting

$$\frac{1}{\text{MSE(plug-in)}}\left[\left(Y_{i\tau+1}-\widehat{Y}_{i\tau+1}\right)^2-\left(Y_{i\tau+1}-\widehat{Y}_{i\tau+1}(\text{plug-in})\right)^2\right].$$

Thus, averaging the dots leads to the MSE differentials in Figure 3. The two zero lines and the 45-degree line partition each panel into six segments. Dots to the left of the vertical zero line correspond to banks for which the pooled OLS forecast is more accurate (lower squared forecast error) than the plug-in predictor. Dots below the horizontal zero line are associated with banks for which the empirical Bayes forecast is more accurate than the forecast from the plug-in predictor. Finally, dots below the 45-degree line correspond to institutions for which the empirical Bayes forecast is more accurate than the pooled OLS forecast.

We focus on three different time periods. In 2007:Q1 the MSE of the pooled OLS predictor is 38% larger than that of the plug-in predictor, whereas the empirical Bayes predictor is only slightly worse, an 8% MSE increase, than the plug-in predictor. In 2009:Q1, the pooled OLS and empirical Bayes predictors perform equally well, and generate a 40% MSE reduction relative to the plug-in predictor. Finally, in 2012:Q1, the empirical Bayes predictor performs better than both the pooled OLS and the plug-in predictor. The top row of Figure 4 shows squared forecast error differentials for all banks, whereas the bottom figure zooms in on differentials between -5 and 5.

The visual impression from the panels is consistent with the MSE ranking of the predictors. For instance, in the left panels there are more banks above the horizontal zero line (410 vs. 317) and to the right of the vertical zero line (470 vs. 257). Moreover, there are more banks below the 45-degree line than above (440 vs. 287). The panels in the center column of the figure indicate that the good performance of the empirical Bayes and pooled OLS predictors is driven in part by some banks for which the plug-in predictor performs very poorly. The corresponding squared forecast error differentials line up along the 45-degree line. It is important to note that the empirical Bayes predictor and the pooled-OLS predictor, despite the similarity in average performance, are not based on the same prediction function. The estimated autoregressive coefficient for the pooled-OLS predictor is much larger than the QMLE estimate of ρ that is used for the empirical Bayes predictor.

6.2 Forecasts Under Stressed Macroeconomic Conditions

We proceed by comparing the baseline forecasts from the previous subsection to predictions under a stressed scenario, in which we use hypothetical values for the predictors. Our subsequent analysis assumes that in the short run there is no feedback from disaggregate BCH revenues to aggregate conditions. While this assumption is inconsistent with the notion that the performance of the banking sector affects macroeconomic outcomes, elements of the Comprehensive Capital Analysis and Review (CCAR) conducted by the Federal Reserve Board of Governors have this partial equilibrium flavor.

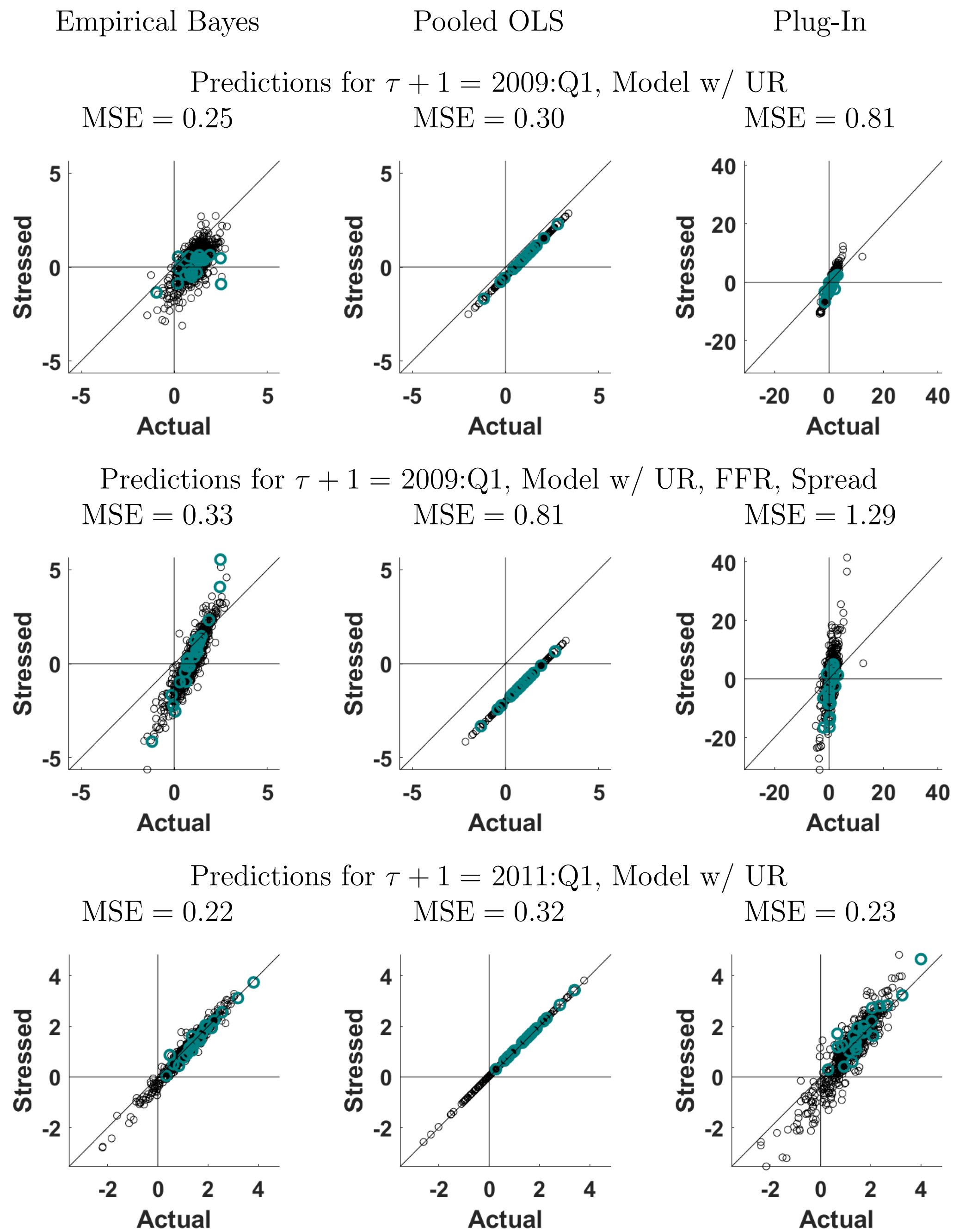
Each circle in Figure 5 corresponds to a one-quarter-ahead point prediction for a particular BHC. We indicate institutions with assets greater than 50 billion dollars[14] by teal circles, while the other BHCs appear as black circles. The x-dimension is the forecast under actual macroeconomic conditions and the y-dimension indicates the forecast under the stressed scenario. For the model with unemployment as covariate we impose stress by increasing the unemployment rate by 5%. This is a similar magnitude to the unemployment increase in the *severely adverse* macroeconomic scenario in the Federal Reserve's CCAR. For the model with three covariates the stressed scenario consists of an increase in the unemployment rate by 5% (as before) and an increase in nominal interest rates and spreads by 5%. This scenario could be interpreted as an aggressive monetary tightening that induced a sharp drop in macroeconomic activity. In each panel of Figure 5 we also report the MSE associated with the various forecasts conditional on the actual macroeconomic conditions.

The graphs in the top two rows of Figure 5 depict forecasts for 2009:Q1 made in the midst of the Great Recession. For the majority of banks – 90% or more based on the empirical Bayes and pooled OLS predictors, and between 80% and 85% under the plug-in predictor – the predicted PPNRs under the actual macroeconomic conditions are positive. The MSEs reported in the figure imply that the predictions from the model with one covariate are more accurate than the prediction for the model with three covariates. This result is not surprising, because in our sample we only have 10 time periods to disentangle the marginal effects of unemployment, federal funds rate, and spreads on bank revenues. For each of the two model specifications, empirical Bayes predictor dominates the pooled-OLS predictor, which in turn attains a lower MSE than the plug-in predictor.[15] Thus, overall, the lowest

[14] These are the BHCs that are subject to the CCAR requirements.

[15] The computation of the empirical Bayes predictor in this section is slightly different. After fitting mixture models to $p(\hat{\lambda}|y_0)$ we discovered that our data driven selection typically generates $\hat{K} = 1$, which means that $\hat{\lambda}|y_0$ is multivariate normal. Rather than directly estimating a normal distribution with an unrestricted

Figure 5: Predictions under Actual and Stressed Scenarios, $T = 10$



Notes: Forecast origins are $\tau =$ 2008:Q4 (panels in rows 1 and 2) and $\tau =$ 2010:Q4 (panels in row 3). Each dot corresponds to a BHC in our dataset. We indicate institutions with assets greater than 50 billion dollars by teal circles. We plot point predictions of PPNR under the actual macroeconomic conditions and a stressed scenario. Model w/ UR: the unemployment rate is 5% higher than its actual level. Model w/ UR, FFR, Spread: the unemployment rate, the federal funds rate, and spread are 5% higher than their actual level. We also report actual MSEs.

---

variance-covariance matrix, we parameterize $p(\hat{\lambda}|y_0)$ in terms of the coefficients of a Gaussian prior $\pi(\lambda|y_0)$, imposing that the prior covariance matrix is diagonal. While in most periods the two approaches lead to the same results, there are some periods in which the latter approach is numerically more stable.

MSE among the six predictors depicted in the top two rows of the figure is attained by the empirical Bayes predictor based on the model with unemployment.

Now consider the predictions under stressed macroeconomic conditions. Based on the model with one covariate, pooled OLS essentially predicts no effects on bank revenues in 2009:Q1 because the dots line up along the 45-degree line. The plug-in predictor, on the other hand, predicts a response that is very heterogeneous across institutions: 33% of the banks are predicted to be able to raise their revenues, whereas for 67% of the institutions the revenues are expected to fall relative to the baseline scenario. Predicted losses are as large as 10% of the bank assets. According to the preferred (based on the MSE under the baseline scenario) empirical Bayes predictor, 93% of the institutions are expected to experience a drop in PPNRs by 1 to 2 percent of their assets. In the model with three covariates, the drop in revenues is generally more pronounced, in particular for BHCs with negative predicted revenues under the actual macroeconomic conditions.

The last row of Figure 5 shows predictions for 2011:Q1 made during the recovery, based on the one-variable model. Unlike in the earlier sample, now the plug-in predictor generates more accurate forecasts (lower MSE) than the pooled-OLS predictor. As before, the empirical Bayes predictor beats both alternatives, albeit the plug-in predictor only by a small margin. The actual-versus-stressed predictions from the empirical Bayes procedure line up along the 45-degree line. For 68% of the institutions predicted profits are lower under the stressed scenario than under the benchmark scenario, but the drop in revenues is very small. Under the plug-in predictor, there is more heterogeneity in the response of banks' PPNRs, with some banks revenues dropping by 1.5 percentage points, whereas other banks are predicted to observe a modest increase in revenues. However, the baseline forecasts of this predictor are less accurate than those from the empirical Bayes predictor, lending more credibility to the latter.

We view this analysis as a first step toward applying state-of-the-art panel data forecasting techniques to stress tests. First, it is important to ensure that the empirical model is able to accurately predict bank revenues and balance sheet characteristics under observed macroeconomic conditions. Our analysis suggests that there are substantial performance differences among various plausible estimators and predictors. Second, a key challenge is to cope with the complexity of models that allow for heterogeneous coefficients in view of the limited information in the sample. There is a strong temptation to over-parameterize models that are used for stress tests. We use prior information to discipline the inference. In our empirical Bayes procedure, this prior information is essentially extracted from the cross-sectional variation in the data set. Third, our empirical results indicate that relative to the cross-sectional dispersion of PPNRs, the effect of severely adverse scenarios on revenue point predictions are very small. We leave it future research to explore richer empirical models that focus on specific revenue and accounting components, consider a broader set of covariates, and possibly allow for feedback from the performance of the banking sector into the aggregate conditions.

7 Conclusion

The literature on panel data forecasting in settings in which the cross-sectional dimension is large and the time-series dimension is small is very sparse. Our paper contributes to this literature by developing an empirical Bayes predictor that uses the cross-sectional information in the panel to construct a prior distribution that can be used to form a posterior mean predictor for each cross-sectional unit. The shorter the time-series dimension and the smaller the parameter heterogeneity, the more important this prior becomes for forecasting and the larger the gains from using the posterior mean predictor instead of a plug-in predictor. We consider a particular implementation of this idea for linear models with Gaussian innovations that is based on Tweedie's posterior mean formula. It can be implemented by estimating the cross-sectional distribution of sufficient statistics for the heterogeneous coefficients in the forecast model. We provide a theorem that establishes a ratio-optimality property for a nonparametric kernel estimator of the Tweedie correction and consider implementations based on the estimation of mixtures of normals and nonparametric MLE in Monte Carlo simulations. We illustrate in an application that our forecasting techniques work well in practice and may be useful to execute bank stress tests.

References

Alvarez, J., and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71(4), 1121–1159.

Anderson, T. W., and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American statistical Association*, 76(375), 598–606.

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

Arellano, M., and O. Bover (1995): "Another Look at the Instrumental Variable Estimation of Error-components Models," *Journal of Econometrics*, 68(1), 29–51.

Baltagi, B. H. (2008): "Forecasting with Panel Data," *Journal of Forecasting*, 27(2), 153–173.

Blundell, R., and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87(1), 115–143.

Brown, L. D., and E. Greenshtein (2009): "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-dimensional Vector of Normal Means," *The Annals of Statistics*, pp. 1685–1704.

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

Covas, F. B., B. Rump, and E. Zakrajsek (2014): "Stress-Testing U.S. Bank Holding Companies: A Dynamic Panel Quantile Regression Approach," *International Journal of Forecasting*, 30(3), 691–713.

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

Goldberger, A. S. (1962): "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57(298), 369–375.

Gu, J., and R. Koenker (2017a): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Economics*, 32(3), 575–599.

——— (2017b): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Business & Economic Statistics*, 35(1), 1–16.

Hirano, K. (2002): "Semiparametric Bayesian inference in autoregressive panel data models," *Econometrica*, 70(2), 781–799.

Hirtle, B., A. Kovner, J. Vickery, and M. Bhanot (2016): "Assessing Financial Stability: The Capital and Loss Assessment under Stress Scenarios (CLASS) Model," *Journal of Banking & Finance*, 69, S35–S55.

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

Kapinos, P., and O. A. Mitnik (2016): "A Top-down Approach to Stress-testing Banks," *Journal of Financial Services Research*, 49(2), 229–264.

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

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *The Review of Economic Studies*, 69(3), 647–666.

Liu, L. (2018): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *arXiv preprint arXiv:1805.04178.*

Liu, L., H. R. Moon, and F. Schorfheide (2016): "Forecasting with Dynamic Panel Data Models," *PIER Working Paper*, 2016-022.

——— (2018a): "Forecasting with a Panel Tobit Model," *Manuscript, University of Pennsylvania.*

——— (2018b): "Forecasting with Dynamic Panel Data Models," *NBER Working Paper*, 2016-022.

Norets, A., and J. Pelenis (2012): "Bayesian Modeling of Joint and Conditional Distributions," *Journal of Econometrics*, 168, 332–346.

Robbins, H. (1951): "Asymptocially Subminimax Solutions of Compound Decision Problems," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press, Berkeley and Los Angeles.

——— (1956): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, Berkeley and Los Angeles.

——— (1964): "The Empirical Bayes Approach to Statistical Decision Problems," *The Annals of Mathematical Statistics*, pp. 1–20.

Robert, C. (1994): *The Bayesian Choice*. Springer Verlag, New York.

Robinson, G. K. (1991): "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, pp. 15–32.

Saha, S., and A. Guntuboyina (2019): "On the Nonparametric Maximum Likelihood Estimator for Gaussian Location Mixture Densities with Application to Gaussian Denoising," *Annals of Statistics*, forthcoming.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Supplemental Appendix to "Forecasting with Dynamic Panel Data Models"

Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide

A Theoretical Derivations and Proofs

A.1 Proofs for Section 3.2

A.1.1 Preliminaries

Throughout the proofs, we use the notation ϵ for a small positive constant such that

$$0 < \epsilon < \epsilon_0.$$

In addition, we will make use of the following two lemmas.

Lemma A.1 *If $A_N(\pi) = o_{u.\pi}(N^+)$ and $B_N(\pi) = o_{u.\pi}(N^+)$, then $C_N(\pi) = A_N(\pi) + B_N(\pi) = o_{u.\pi}(N^+)$.*

Proof of Lemma A.1. Take an arbitrary $\epsilon > 0$. We need to show that there exists a sequence $\eta_N^c(\epsilon)$ such that

$$N^{-\epsilon} C_N(\pi) \leq \eta_N^c(\epsilon).$$

Write

$$N^{-\epsilon} C_N(\pi) = N^{-\epsilon}(A_N(\pi) + B_N(\pi)).$$

Because A_N and B_N are subpolynomial, there exist sequences $\eta_N^a(\epsilon)$ and $\eta_N^b(\epsilon)$ such that

$$N^{-\epsilon}(A_N(\pi) + B_N(\pi)) \leq \eta_N^a(\epsilon) + \eta_N^b(\epsilon).$$

Thus, we can choose $\eta_N^c(\epsilon) = \eta_N^a(\epsilon) + \eta_N^b(\epsilon) \longrightarrow 0$ to establish the claim. ■

Lemma A.2 *If $A_N(\pi) = o_{u.\pi}(N^+)$ and $B_N(\pi) = o_{u.\pi}(N^+)$, then $C_N(\pi) = A_N(\pi) B_N(\pi) = o_{u.\pi}(N^+)$.*

Proof of Lemma A.2. Take an arbitrary $\epsilon > 0$. We need to show that there exists a sequence $\eta_N^c(\epsilon)$ such that

$$N^{-\epsilon} C_N(\pi) \leq \eta_N^c(\epsilon).$$

Write

$$N^{-\epsilon} C_N(\pi) = \big(N^{-\epsilon/2} A_N(\pi)\big)\big(N^{-\epsilon/2} B_N(\pi)\big).$$

Because A_N and B_N are subpolynomial, there exist sequences $\eta_N^a(\epsilon/2)$ and $\eta_N^b(\epsilon/2)$ such that

$$\big(N^{-\epsilon/2} A_N(\pi)\big)\big(N^{-\epsilon/2} B_N(\pi)\big) \leq \eta_N^a(\epsilon/2)\eta_N^b(\epsilon/2).$$

Thus, we can choose $\eta_N^c(\epsilon) = \eta_N^a(\epsilon/2)\eta_N^b(\epsilon/2) \longrightarrow 0$ to establish the claim. ■

A.1.2 Main Theorem

Proof of Theorem 3.7. The goal is to prove that for any given $\epsilon_0 > 0$

$$\limsup_{N\longrightarrow\infty} \sup_{\pi\in\Pi} \frac{R_N(\widehat{Y}_{T+1}^N; \pi) - R_N^{\text{opt}}(\pi)}{N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i,\lambda_i}\big[(\lambda_i - \mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\big] + N^{\epsilon_0}} \leq 0, \tag{A.1}$$

where

$$\begin{aligned} R_N(\widehat{Y}_{T+1}^N; \pi) &= N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N,\lambda_i}\left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right] + N\sigma^2 \\ R_N^{\text{opt}}(\pi) &= N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}_i,\lambda_i}\left[\left(\lambda_i - \mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\lambda_i}[\lambda_i]\right)^2\right] + N\sigma^2. \end{aligned}$$

Here we used the fact that there is cross-sectional independence and symmetry in terms of i. The desired statement follows if we show

$$\limsup_{N\longrightarrow\infty} \sup_{\pi\in\Pi} \frac{N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N,\lambda_i}\left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right]}{N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i,\lambda_i}\big[(\lambda_i - \mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\big] + N^{\epsilon_0}} \leq 1. \tag{A.2}$$

Here we made the dependence on π of the risks and the posterior moments explicit. In the calculations below, we often drop the π argument to simplify the notation.

In the main text we asserted that

$$p_*(\hat{\lambda}_i, y_{i0}) = \mathbb{E}_{\theta,\pi,\mathcal{Y}_i}^{\mathcal{Y}^{(-i)}}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})]. \tag{A.3}$$

This assertion can be verified as follows. Taking expectations with respect to $(\hat{\lambda}_j, y_{j,0})$ for $j \neq i$ yields

$$\begin{aligned}
&\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i;\pi}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] \\
&= \sum_{j\neq i} \int\int \frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_i - \hat{\lambda}_j}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{y_{i0}-y_{j0}}{B_N}\right)p(\hat{\lambda}_j, y_{j0})d\hat{\lambda}_j dy_{j0} \\
&= \int\int \frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_i - \hat{\lambda}_j}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{y_{i0}-y_{j0}}{B_N}\right)p(\hat{\lambda}_j, y_{j0})d\hat{\lambda}_j dy_{j0}.
\end{aligned}$$

The second equality follows from the symmetry with respect to j and the fact that we integrate out $(\hat{\lambda}_j, y_{j0})$. We now substitute in

$$p(\hat{\lambda}_j, y_{j0}) = \int p(\hat{\lambda}_j|\lambda_j)\pi(\lambda_j, y_{j0})d\lambda_j,$$

where

$$p(\hat{\lambda}_j|\lambda_j) = \frac{1}{\sigma/T}\phi\left(\frac{\hat{\lambda}_j-\lambda_j}{\sigma/T}\right),$$

and change the order of integration. This leads to:

$$\begin{aligned}
&\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] \\
&= \int\int\left[\int \frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_i - \hat{\lambda}_j}{B_N}\right)p(\hat{\lambda}_j|\lambda_j)d\hat{\lambda}_j\right]\frac{1}{B_N}\phi\left(\frac{y_{i0}-y_{j0}}{B_N}\right)\pi(\lambda_j, y_{j0})d\lambda_j dy_{j0} \\
&= \int\int \frac{1}{\sqrt{\sigma^2/T+B_N^2}}\phi\left(\frac{\hat{\lambda}_i-\lambda_j}{\sqrt{\sigma^2/T+B_N^2}}\right)\frac{1}{B_N}\phi\left(\frac{y_{i0}-y_{j0}}{B_N}\right)\pi(\lambda_j, y_{j0})d\lambda_j dy_{j0} \\
&= \int \frac{1}{\sqrt{\sigma^2/T+B_N^2}}\phi\left(\frac{\hat{\lambda}_i-\lambda_j}{\sqrt{\sigma^2/T+B_N^2}}\right)\left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0}-y_{j0}}{B_N}\right)\pi(y_{j0}|\lambda_j)dy_{j0}\right]\pi(\lambda_j)d\lambda_j.
\end{aligned}$$

Now re-label λ_j and λ_i and y_{j0} as $\tilde{y}_{i0}$ to obtain:

$$\begin{aligned}
&p_*(\hat{\lambda}_i, y_{i0}) \\
&= \int \frac{1}{\sqrt{\sigma^2/T+B_N^2}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T+B_N^2}}\right)\left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0}-\tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}\right]\pi(\lambda_i)d\lambda_i.
\end{aligned}$$

Risk Decomposition. We begin by decomposing the forecast error. Let

$$\mu\big(\lambda, \omega^2, p(\lambda, y_0)\big) = \lambda + \omega^2\frac{\partial \ln p(\lambda, y_0)}{\partial \lambda}. \tag{A.4}$$

Using the previously developed notation, we expand the prediction error due to parameter estimation as follows:

$$\begin{aligned}
&\widehat{Y}_{iT+1} - \lambda_i - \rho Y_{iT} \\
= \; & \Big[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\Big]^{C_N} - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) \\
& + m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i \\
& + (\hat{\rho} - \rho) Y_{iT} \\
= \; & A_{1i} + A_{2i} + A_{3i}, \text{ say.}
\end{aligned}$$

Now write

$$N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right] = N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[(A_{1i} + A_{2i} + A_{3i})^2\right].$$

We deduce from the C_r inequality that the statement of the theorem follows if we can show that

$$\begin{aligned}
&\text{(i)} \qquad N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[A_{1i}^2\right] = o_{u.\pi}(N^{\epsilon_0}), \\
&\text{(ii)} \qquad \limsup_{N\longrightarrow\infty} \sup_{\pi\in\Pi} \frac{N\mathbb{E}^{\mathcal{Y}^N,\lambda_i}_{\theta,\pi}\left[A_{2i}^2\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}^i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 1, \\
&\text{(iii)} \qquad N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[A_{3i}^2\right] = o_{u.\pi}(N^{+}).
\end{aligned}$$

The required bounds are provided in Lemmas A.3 (term A_{1i}), A.4 (term A_{2i}), A.5 (term A_{3i}). ■

A.1.3 Three Important Lemmas

Truncations. The remainder of the proof involves a number of truncations that we will apply when analyzing the risk terms. We take the sequence C_N as given from Assumption 3.3. Recall that

$$\frac{2}{M_2}\ln N \leq C_N < \frac{1}{B_N}.$$

We introduce a new sequence diverging sequence L_N with the properties

$$\liminf_N L_N B_N > 1 \quad \text{and} \quad L_N = o(N^{+}). \tag{A.5}$$

Even though we do not indicate this explicitly through our notation, we also restrict the domain of (λ, y_0) arguments that appear in numerous expressions throughout the proof to the support of the distribution of the random variables (λ_i, Y_{i0}), which is defined as $\text{Supp}_{\lambda,Y_0} = \{(\lambda, y_0) \in \mathbb{R}^2 \,|\, \pi(\lambda, y_0) > 0\}$.

1. Define the truncated region $\mathcal{T}_\lambda = \{|\lambda| \leq C_N\}$. From Assumption 3.2 we obtain for $C_N \geq M_3$ that

$$
\begin{aligned}
N^{1-\epsilon}\mathbb{P}(\mathcal{T}_\lambda^c) &\leq M_1 \exp\big((1-\epsilon)\ln N - M_2(C_N - M_3)\big) \\
&= \widetilde{M}_1 \exp\left(-M_2\left[C_N - \frac{1-\epsilon}{M_2}\ln N\right]\right) \\
&= o(1),
\end{aligned}
\tag{A.6}
$$

for all $0 < \epsilon$ because, according to Assumption 3.3, $C_N > 2(\ln N)/M_2$. Thus, we can deduce

$$N\mathbb{P}(\mathcal{T}_\lambda^c) = o_{u.\pi}(N^+).$$

2. Define the truncated region $\mathcal{T}_{Y_0} = \{\max_{1\leq i\leq N} |Y_{i0}| \leq L_N\}$. Then,

$$
\begin{aligned}
N^{1-\epsilon}\mathbb{P}(\mathcal{T}_{Y_0}^c) &= N^{1-\epsilon}\mathbb{P}\{\max_{1\leq i\leq N} |Y_{i0}| \geq L_N\} \\
&\leq N^{1-\epsilon}\sum_{i=1}^{N} \mathbb{P}\{|Y_{i0}| \geq L_N\} \\
&= N^{2-\epsilon}\int_{|y_0|\geq L_N} \pi(y_0)dy_0 \\
&\leq \widetilde{M}_1 \exp\left(-M_2\left[L_N - \frac{2-\epsilon}{M_2}\ln N\right]\right) \\
&= o(1),
\end{aligned}
\tag{A.7}
$$

for all $\epsilon > 0$ because according to (A.5) $L_N > (2/M_2)\ln N$. Thus, we deduce that

$$N\mathbb{P}(\mathcal{T}_{Y_0}^c) = o_{u.\pi}(N^+).$$

3. Define the truncated region $\mathcal{T}_{\hat{\rho}} = \{|\hat{\rho} - \rho| \leq 1/L_N^2\}$. By Chebyshev's inequality, Assumption 3.6, and (A.5), we can bound

$$N\mathbb{P}(\mathcal{T}_{\hat{\rho}}^c) = N\mathbb{P}\{|\hat{\rho} - \rho| > 1/L_N^2\} \leq L_N^4\mathbb{E}\left[N(\hat{\rho} - \rho)^2\right] = o_{u.\pi}(N^+). \tag{A.8}$$

4. Define the truncated region $\mathcal{T}_{\hat{\sigma}^2} = \{|\hat{\sigma}^2 - \sigma^2| \leq 1/L_N\}$. By Chebyshev's inequality,

Assumption 3.6, and (A.5), we can bound

$$N\mathbb{P}(\mathcal{T}^c_{\hat{\sigma}^2}) = N\mathbb{P}\{|\hat{\sigma}^2 - \sigma^2| > 1/L_N\} \leq L_N^2 \mathbb{E}[N(\hat{\sigma}^2 - \sigma^2)^2] = o_{u.\pi}(N^+). \tag{A.9}$$

5. Let $\bar{U}_{i,-1}(\rho) = \frac{1}{T}\sum_{t=2}^T U_{it-1}(\rho)$ and $U_{it}(\rho) = U_{it} + \rho U_{it-1} + \cdots + \rho^{t-1}U_{i1}$. Define the truncated region $\mathcal{T}_{\bar{U}} = \left\{\max_{1\leq i\leq N}|\bar{U}_{i,-1}(\rho)| \leq L_N\right\}$. Notice that $\bar{U}_{i,-1}(\rho) \sim iidN(0, \sigma^2_{\bar{U}})$ with $0 < \sigma^2_{\bar{U}} < \infty$. Thus, we have

$$\begin{aligned} N\mathbb{P}(\mathcal{T}^c_{\bar{U}}) &= N\mathbb{P}\{\max_{1\leq i\leq N}|\bar{U}_{i,-1}(\rho)| \geq L_N\} \qquad \text{(A.10)} \\ &\leq N\sum_{i=1}^N \mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} = N^2\mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} \\ &\leq 2N^2\exp\left(-\frac{L_N^2}{2\sigma^2_{\bar{U}}}\right) = 2\exp\left(-\frac{L_N^2}{2\sigma^2_{\bar{U}}} + 2\ln N\right) \\ &\leq 2\exp\left(-2\left(\frac{\ln N}{M_2^2\sigma^2_{\bar{U}}} - 1\right)\ln N\right) \\ &= o_{u.\pi}(N^+), \end{aligned}$$

where the last inequality holds by (A.5).

6. Let $\bar{Y}_{i,-1} = C_1(\rho)Y_{i0} + C_2(\rho)\lambda_i + \bar{U}_{i,-1}(\rho)$, where $C_1(\rho) = \frac{1}{T}\sum_{t=1}^T \rho^{t-1}$, $C_2(\rho) = \frac{1}{T}\sum_{t=2}^T(1 + \cdots + \rho^{t-2})$. Because T is finite and $|\rho|$ is bounded, there exists a finite constant, say M such that $|C_1(\rho)| \leq M$ and $|C_2(\rho)| \leq M$. Then, in the region $\mathcal{T}_\lambda \cap \mathcal{T}_{Y_0} \cap \mathcal{T}_{\bar{U}}$:

$$\begin{aligned} \max_{1\leq i\leq N}|\bar{Y}_{i,-1}| &\leq |C_1(\rho)|\max_{1\leq i\leq N}|\lambda_i| + |C_2(\rho)|\max_{1\leq i\leq N}|Y_{i0}| + \max_{1\leq i\leq N}|\bar{U}_{i,-1}(\rho)| \\ &\leq M(C_N + L_N + L_N), \end{aligned}$$

which leads to

$$\max_{1\leq i,j\leq N}|\bar{Y}_{j,-1} - \bar{Y}_{i,-1}| \leq 2\max_{1\leq i\leq N}|\bar{Y}_{i,-1}| \leq 2M(C_N + 2L_N) = o_{u.\pi}(N^+). \tag{A.11}$$

7. For the region $\mathcal{T}_\lambda \cap \mathcal{T}_{Y_0} \cap \mathcal{T}_{\hat{\rho}} \cap \mathcal{T}_{\bar{U}}$ and with some finite constant M, we obtain the bound

$$\max_{1\leq i,j\leq N}|(\hat{\rho} - \rho)(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})| \leq \frac{M(C_N + L_N)}{L_N^2} = o_{u.\pi}(N^+). \tag{A.12}$$

8. Define the regions $\mathcal{T}_m = \{|m(\hat{\lambda}_i, Y_{i0})| \leq C_N\}$ and $\mathcal{T}_{m*} = \{|m_*(\hat{\lambda}_i, Y_{i0})| \leq C_N\}$. By Chebyshev's inequality and Assumption 3.5, we deduce

$$\begin{aligned} N\mathbb{P}(\mathcal{T}_m^c) &\leq \frac{1}{C_N^2} N\mathbb{E}(m(\hat{\lambda}_i, Y_{i0})^2 \mathcal{T}_m^c) \leq o_{u.\pi}(N^+) \\ N\mathbb{P}(\mathcal{T}_{m_*}^c) &\leq \frac{1}{C_N^2} N\mathbb{E}(m_*(\hat{\lambda}_i, Y_{i0})^2 \mathcal{T}_{m_*}^c) \leq o_{u.\pi}(N^+). \end{aligned} \tag{A.13}$$

We will subsequently use indicator function notation, abbreviating, say, $\mathbb{I}\{\lambda \in \mathcal{T}_\lambda\}$ by $\mathbb{I}(\mathcal{T}_\lambda)$ and $\mathbb{I}(\mathcal{T}_\lambda)\mathbb{I}(\mathcal{T}_{Y_0})$ by $\mathbb{I}(\mathcal{T}_\lambda \mathcal{T}_{Y_0})$.

A.1.3.1 Term A_{1i}

Lemma A.3 *Suppose the assumptions in Theorem 3.7 hold. Then,*

$$N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\left(\left[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\right]^{C_N} - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)\right)^2\right] = o_{u.\pi}(N^{\epsilon_0}).$$

Proof of Lemma A.3. We begin with the following bound: since $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned} |A_{1i}|^2 &= \left(\left[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\right]^{C_N} - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)\right)^2 \\ &\leq 2C_N^2 + 2m_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N) \\ &= 2C_N^2 + 2m_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N)\mathbb{I}(\mathcal{T}_{m*}) + 2m_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N)\mathbb{I}(\mathcal{T}_{m*}^c) \\ &\leq 4C_N^2 + 2m_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N)\mathbb{I}(\mathcal{T}_{m*}^c). \end{aligned} \tag{A.14}$$

Then,

$$\begin{aligned} N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[A_{1i}^2] &\leq N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_\lambda) \\ &\quad + N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[A_{1i}^2\big(\mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}^c) + \mathbb{I}(\mathcal{T}_{\hat{\rho}}^c) + \mathbb{I}(\mathcal{T}_{\bar{U}}^c) + \mathbb{I}(\mathcal{T}_{Y_0}^c) + \mathbb{I}(\mathcal{T}_{m*}^c) + \mathbb{I}(\mathcal{T}_\lambda^c)\big)\right] \\ &\leq N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_\lambda)] \\ &\quad + 4C_N^2 N\big(\mathbb{P}(\mathcal{T}_{\hat{\sigma}^2}^c) + \mathbb{P}(\mathcal{T}_{\hat{\rho}}^c) + \mathbb{P}(\mathcal{T}_{\bar{U}}^c) + \mathbb{P}(\mathcal{T}_{Y_0}^c) + \mathbb{P}(\mathcal{T}_{m*}^c) + \mathbb{P}(\mathcal{T}_\lambda^c)\big) \\ &\quad + 12N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[m_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N)\mathbb{I}(\mathcal{T}_{m*}^c)] \\ &= N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_\lambda)] + o_{u.\pi}(N^+) + o_{u.\pi}(N^+). \end{aligned} \tag{A.15}$$

The first $o_{u.\pi}(N^+)$ follows from the properties of the truncation regions discussed above and the second $o_{u.\pi}(N^+)$ follows from Assumption 3.5. In the remainder of the proof we will construct the desired bound, $o_{u.\pi}(N^{\epsilon_0})$, for the first term on the right-hand side of (A.15). We proceed in two steps.

Step 1. We introduce two additional truncation regions, $\mathcal{T}_{\hat{\lambda}Y_0}$ and $\mathcal{T}_{p(\cdot)}$, which are defined as follows:

$$\begin{aligned}\mathcal{T}_{\hat{\lambda}Y_0} &= \left\{(\hat{\lambda}_i, Y_{i0}) \,\middle|\, -C'_N \le \hat{\lambda}_i \le C'_N,\ -C'_N \le Y_{i0} \le C'_N\right\}\\ \mathcal{T}_{p(\cdot)} &= \left\{(\hat{\lambda}_i, Y_{i0}) \,\middle|\, p(\hat{\lambda}_i, Y_{i0}) \ge \frac{N^{\epsilon}}{N}\right\},\end{aligned}$$

where it is assumed that $0 < \epsilon < \epsilon_0$.

Notice that since $C_N = o(N^+)$ and $\sqrt{\ln N} = o(N^+)$,

$$C'_N = o(N^+). \tag{A.16}$$

In the first truncation region both $\hat{\lambda}_i$ and Y_{i0} are bounded by C'_N. In the second truncation region the density $p(\hat{\lambda}_i, Y_{i0})$ is not too low. We will show that

$$N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[A_{1i}^2\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}^c_{p(\cdot)})] \le o_{u.\pi}(N^{\epsilon_0}) \tag{A.17}$$

$$N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[A_{1i}^2\mathbb{I}(\mathcal{T}_{\lambda}\mathcal{T}^c_{\hat{\lambda}Y_0})] \le o_{u.\pi}(N^+). \tag{A.18}$$

Step 1.1. First, we consider the case where $(\hat{\lambda}_i, Y_{i0})$ are bounded and the density $p(\hat{\lambda}_i, y_{i0})$ is "low" in (A.17). Using the bound for $|A_{1i}|$ in (A.14) we obtain:

$$\begin{aligned}&N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[A_{1i}^2\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0})\mathbb{I}(\mathcal{T}^c_{p(\cdot)})\right]\\ &\le 4NC_N^2\mathbb{P}(\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}^c_{p(\cdot)}) + 2N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[m_*^2(\hat{\lambda}_i, y_{i0};\pi, B_N)\mathbb{I}(\mathcal{T}^c_{m*})\right]\\ &= 4NC_N^2\int_{\hat{\lambda}_i=-C'_N}^{C'_N}\int_{y_{i0}=-C'_N}^{C'_N}\mathbb{I}\left\{p(\hat{\lambda}_i, y_{i0}) < \frac{N^{\epsilon}}{N}\right\}p(\hat{\lambda}_i, y_{i0})d(\hat{\lambda}_i, y_{i0}) + o_{u.\pi}(N^+)\\ &\le 4NC_N^2\int_{\hat{\lambda}_i=-C'_N}^{C'_N}\int_{y_{i0}=-C'_N}^{C'_N}\left(\frac{N^{\epsilon}}{N}\right)dy_{i0}d\hat{\lambda}_i + o_{u.\pi}(N^+)\\ &= 4C_N^2(2C'_N)^2N^{\epsilon} + o_{u.\pi}(N^+)\\ &\le o_{u.\pi}(N^{\epsilon_0}).\end{aligned}$$

The $o_{u.\pi}(N^+)$ term in the first equality follows from Assumption 3.5. The last equality holds because $C_N, C'_N = o_{u.\pi}(N^+)$ (Assumption 3.3 and (A.16)) and $0 < \epsilon < \epsilon_0$. This establishes (A.17).

Step 1.2. Next, we consider the case where $(\hat{\lambda}_i, y_{i0})$ exceed the C'_N bound and the density $p(\hat{\lambda}_i, y_{i0})$ is "high." We will immediately replace the contribution of $2Nm_*^2(\hat{\lambda}_i, y_{i0}; \pi, B_N)\mathbb{I}(\mathcal{T}_{m*}^c)$ to the expected value of A_{1i}^2 by $o_{u.\pi}(N^+)$.

$$
\begin{aligned}
& N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[A_{1i}^2 \mathbb{I}(\mathcal{T}_\lambda \mathcal{T}_{\hat{\lambda}Y_0}^c))\right] \\
\leq \;& 4NC_N^2 \mathbb{P}(\mathcal{T}_\lambda \cap \mathcal{T}_{\hat{\lambda}Y_0}^c) + o_{u.\pi}(N^+) \\
= \;& 4NC_N^2 \int_{\mathcal{T}_{\hat{\lambda}Y_0}^c} \left[\int_{\lambda_i} \frac{1}{\sigma/\sqrt{T}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sigma/\sqrt{T}}\right) \pi(\lambda_i, y_{i0}) d\lambda_i\right] d(\hat{\lambda}_i, y_{i0}) + o_{u.\pi}(N^+) \\
\leq \;& 4NC_N^2 \int_{\lambda_i} \int_{|\hat{\lambda}_i| > C'_N} \left[\int_{y_{i0}} \frac{1}{\sigma/\sqrt{T}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sigma/\sqrt{T}}\right) \pi(y_{i0}|\lambda_i) dy_{i0}\right] \pi(\lambda_i) d(\hat{\lambda}_i, \lambda_i) \\
& + 4NC_N^2 \int_{\lambda_i} \int_{|y_{i0}| > C'_N} \left[\int_{\hat{\lambda}_i} \frac{1}{\sigma/\sqrt{T}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sigma/\sqrt{T}}\right) d\hat{\lambda}_i\right] \pi(\lambda_i, y_{i0}) d(\lambda_i, y_{i0}) + o_{u.\pi}(N^+) \\
= \;& 4NC_N^2 \int_{|\lambda_i| \leq C_N} \left[\int_{|\hat{\lambda}_i| > C'_N} \frac{1}{\sigma/\sqrt{T}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sigma/\sqrt{T}}\right) d\hat{\lambda}_i\right] \pi(\lambda_i) d\lambda_i \\
& + 4NC_N^2 \int_{|y_{i0}| > C'_N} \left[\int_{\lambda_i} \pi(\lambda_i | y_{i0}) d\lambda_i\right] \pi(y_{i0}) dy_{i0} \\
& + o_{u.\pi}(N^+) \\
\leq \;& 4NC_N^2 \int_{|\lambda_i| \leq C_N} \left[\int_{|\hat{\lambda}_i| > C'_N} \frac{1}{\sigma/\sqrt{T}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sigma/\sqrt{T}}\right) d\hat{\lambda}_i\right] \pi(\lambda_i) d\lambda_i \\
& + 4NC_N^2 \int_{|y_{i0}| > C'_N} \pi(y_{i0}) dy_{i0} \\
& + o_{u.\pi}(N^+) \\
= \;& B_1 + o_{u.\pi}(N^+) + o_{u.\pi}(N^+), \quad \text{say.}
\end{aligned}
$$

The second equality is obtained by integrating out $\hat{\lambda}_i$, recognizing that the integrand is a properly scaled probability density function that integrates to one. The last line follows from the calculations in (A.6), Lemma A.2, and $C'_N > C_N$.

We will first analyze term B_1. Note by the change of variable that

$$\begin{aligned}
&\int_{|\hat{\lambda}_i|>C'_N} \frac{1}{\sigma/\sqrt{T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)d\hat{\lambda}_i \\
&= \int_{-\infty}^{-\sqrt{T}(C'_N+\lambda_i)/\sigma} \phi(\tilde{\lambda}_i)d\tilde{\lambda}_i + \int_{\sqrt{T}(C'_N-\lambda_i)/\sigma}^{\infty} \phi(\tilde{\lambda}_i)d\tilde{\lambda}_i \\
&\leq \int_{-\infty}^{-\sqrt{T}(C'_N-|\lambda_i|)/\sigma} \phi(\tilde{\lambda}_i)d\tilde{\lambda}_i + \int_{\sqrt{T}(C'_N-|\lambda_i|)/\sigma}^{\infty} \phi(\tilde{\lambda}_i)d\tilde{\lambda}_i \\
&\leq 2\int_{\sqrt{T}(C'_N-|\lambda_i|)/\sigma}^{\infty} \phi(\tilde{\lambda}_i)d\tilde{\lambda}_i \\
&\leq 2\frac{\phi\big(\sqrt{T}(C'_N-|\lambda_i|)/\sigma\big)}{\sqrt{T}(C'_N-|\lambda_i|)/\sigma},
\end{aligned}$$

where we used the inequality $\int_x^\infty \phi(\lambda)d\lambda \leq \phi(x)/x$. Using the definition of C'_N in Assumption 3.3 we obtain the bound (for $\sqrt{2\ln N} \geq 1$):

$$\begin{aligned}
B_1 &\leq 4NC_N^2 \int_{|\lambda_i|<C_N} \frac{\phi\big(\sqrt{T}(C'_N-|\lambda_i|)/\sigma\big)}{\sqrt{T}(C'_N-|\lambda_i|)/\sigma}\pi(\lambda_i)d\lambda_i \\
&\leq 4NC_N^2 \int_{|\lambda_i|<C_N} \phi\big(\sqrt{2\ln N}\big)\pi(\lambda_i)d\lambda_i \\
&\leq 4NC_N^2 \exp\big(-\ln N\big)\int_{|\lambda_i|<C_N} \pi(\lambda_i)d\lambda_i \\
&\leq 4C_N^2 \\
&= o_{u.\pi}(N^+).
\end{aligned}$$

This leads to the desired bound in (A.18).

Step 2. It remains to be shown that

$$N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}[A_{1i}^2\mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\ \mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)})] \leq o_{u.\pi}(N^+). \tag{A.19}$$

We introduce the following notation:

$$
\begin{aligned}
\widetilde{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \\
d\widetilde{p}_i^{(-i)} &= \frac{1}{\partial \hat{\lambda}_i(\hat{\rho})} \partial \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \\
\hat{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\rho), Y_{i0}) \\
d\hat{p}_i^{(-i)} &= \frac{1}{\partial \hat{\lambda}_i(\rho)} \partial \hat{p}^{-i}(\hat{\lambda}_i(\rho), Y_{i0}) \\
p_i &= p(\hat{\lambda}_i(\rho), Y_{i0}) \\
p_{*i} &= p_*(\hat{\lambda}_i(\rho), Y_{i0}) \\
dp_{*i} &= \frac{1}{\partial \hat{\lambda}_i(\rho)} \partial p_*(\hat{\lambda}_i(\rho), Y_{i0}).
\end{aligned} \tag{A.20}
$$

Moreover, we introduce another truncation

$$
\mathcal{T}_{\tilde{p}(\cdot)} = \left\{ (\hat{\lambda}_i, Y_{i0}) \,\middle|\, \widetilde{p}_i^{(-i)} > \frac{p_{*i}}{2} \right\}. \tag{A.21}
$$

On the set $\mathcal{T}_{m*}$, we have $|m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N)| \leq C_N$, and so

$$
|A_{1i}| \leq 2C_N. \tag{A.22}
$$

For the required result of Step 2 in (A.19), we show the following two inequalities; see Steps 2.1 and 2.2 below:

$$
N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\ \mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)})\ \mathcal{T}_{\tilde{p}(\cdot)}] \leq o_{u.\pi}(N^+) \tag{A.23}
$$

$$
N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\ \mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)})\ \mathcal{T}^c_{\tilde{p}(\cdot)}] \leq o_{u.\pi}(N^+). \tag{A.24}
$$

Step 2.1. Using the triangle inequality, we obtain

$$
\begin{aligned}
|A_{1i}| &= \left| \left[\mu\big(\hat{\lambda}_i(\hat{\rho}), Y_{i0}, \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\right]^{C_N} - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) \right| \\
&\leq \left| \mu\big(\hat{\lambda}_i(\hat{\rho}), Y_{i0}, \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big) - m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) \right| \\
&= \left| \hat{\lambda}_i(\hat{\rho}) - \lambda_i(\rho) + \left(\frac{\hat{\sigma}^2}{T} - \frac{\sigma^2}{T}\right)\frac{dp_{*i}}{p_{*i}} + \left(\frac{\hat{\sigma}^2}{T} + B_N^2\right)\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right) \right| \\
&\leq |\hat{\rho} - \rho|\big|\bar{Y}_{i,-1}\big| + \left|\frac{\hat{\sigma}^2}{T} - \frac{\sigma^2}{T}\right|\left|\frac{dp_{*i}}{p_{*i}}\right| + \left(\frac{\hat{\sigma}^2}{T} + B_N^2\right)\left|\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right|, \\
&= A_{11i} + A_{12i} + A_{13i}, \quad \text{say.}
\end{aligned}
$$

Recall that $\bar{Y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} Y_{it-1}$. Using the Cauchy-Schwarz inequality, it suffices to show that

$$
N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{1ji}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\ \mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)}\ \mathcal{T}_{\tilde{p}(\cdot)})\right] \leq o_{u.\pi}(N^+), \quad j = 1, 2, 3.
$$

For term $\mathbf{A_{11i}}$. First, using a slightly more general argument than the one used in the proof of Lemma A.5 below, we can show that

$$
N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{11i}^2\right] = \mathbb{E}_\theta^{\mathcal{Y}^N}\left[N(\hat{\rho} - \rho)^2 \bar{Y}_{i,-1}^2\right] = o_{u.\pi}(N^+).
$$

For term $\mathbf{A_{12i}}$. Second, in the region $\mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{m*}$ we can bound the Tweedie correction term under p_{*i} by

$$
\left(\frac{\sigma^2}{T} + B_N^2\right)\left|\frac{dp_{*i}}{p_{*i}}\right| = \left| m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \hat{\lambda}_i(\rho) \right| \leq C_N + C_N'. \tag{A.25}
$$

Using Assumption 3.3, Assumption 3.6, and (A.16), we obtain the bound

$$
N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[A_{12i}^2 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0})\mathbb{I}(\mathcal{T}_{m*})\right] \leq \frac{1}{(\sigma^2/T + B_N^2)^2}\mathbb{E}_\theta^{\mathcal{Y}^N}\left[N(\hat{\sigma}^2 - \sigma^2)^2\right](C_N' + C_N)^2 = o_{u.\pi}(N^+).
$$

For term $\mathbf{A_{13i}}$. Finally, note that

$$
A_{13i}^2 \mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}) \leq \left(\frac{\sigma^2}{T} + B_N^2 + \frac{1}{T}\frac{1}{L_N}\right)^2 \left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2.
$$

Thus, the desired result follows if we show

$$N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}}-\frac{dp_{*i}}{p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\ \mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)})\right]=o_{u.\pi}(N^+) \tag{A.26}$$

Decompose

$$\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}}-\frac{dp_{*i}}{p_{*i}}=\frac{d\widetilde{p}_i^{(-i)}-dp_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}-\frac{dp_{*i}}{p_{*i}}\left(\frac{\widetilde{p}_i^{(-i)}-p_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right).$$

Using the C_r inequality, we obtain

$$\begin{aligned}
&N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}}-\frac{dp_{*i}}{p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\\
\leq\ & 2N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{d\widetilde{p}_i^{(-i)}-dp_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\\
&+2N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{dp_{*i}}{p_{*i}}\right)^2\left(\frac{\widetilde{p}_i^{(-i)}-p_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\\
=\ & 2\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{\sqrt{N}(d\widetilde{p}_i^{(-i)}-dp_{*i})}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\\
&+2o_{u.\pi}(N^+)\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\left(\frac{\sqrt{N}(\widetilde{p}_i^{(-i)}-p_{*i})}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\\
=\ & 2B_{1i}+2o_{u.\pi}(N^+)B_{2i},
\end{aligned}$$

say. The $o_{u.\pi}(N^+)$ bound follows from (A.25). Using the mean-value theorem, we can express

$$\begin{aligned}
\sqrt{N}(d\widetilde{p}_i^{(-i)}-dp_{*i}) &= \sqrt{N}(d\hat{p}_i^{(-i)}-dp_{*i})+\sqrt{N}(\hat{\rho}-\rho)R_{1i}(\widetilde{\rho})\\
\sqrt{N}(\widetilde{p}_i^{(-i)}-p_{*i}) &= \sqrt{N}(\hat{p}_i^{(-i)}-p_{*i})+\sqrt{N}(\hat{\rho}-\rho)R_{2i}(\widetilde{\rho}),
\end{aligned}$$

where

$$R_{1i}(\rho) = -\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)^2(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right)$$
$$+\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N^2}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right),$$
$$R_{2i}(\rho) = \frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right),$$

and $\widetilde{\rho}$ is located between $\hat{\rho}$ and ρ.

We proceed with the analysis of B_{2i}. Over the region $\mathcal{T}_{\tilde{p}(\cdot)}$, $\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i} > p_{*i}/2$. Using this, the C_r inequality, and the law of iterated expectations, we obtain

$$\begin{aligned} B_{2i} &\leq 4\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^2}\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{p}_i^{(-i)}-p_{*i})^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\right] \\ &\quad +4\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^2}\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{\rho}-\rho)^2R_{2i}^2(\tilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\right] \\ &= 4\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}[B_{21i}+B_{22i}], \end{aligned}$$

say.

According to Lemma A.8(c) (see Section A.1.4),

$$\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{p}_i^{(-i)}-p_{*i})^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right] \leq \frac{M}{B_N^2}p_i\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}).$$

This leads to

$$\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}[B_{21i}] \leq \frac{M}{B_N^2}\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\frac{p_i}{p_{*i}^2}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] = \frac{M}{B_N^2}\int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\frac{p_i^2}{p_{*i}^2}d\hat{\lambda}_i dy_{i0}.$$

According to Lemma A.8(e) (see Section A.1.4),

$$\int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\frac{p_i^2}{p_{*i}^2}d\hat{\lambda}_i dy_{i0} = o_{u.\pi}(N^+).$$

Because $1/B_N^2 = o(N^+)$ according to Assumption 3.3, we can deduce that

$$\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}[B_{21i}] \leq o_{u.\pi}(N^+).$$

Using the Cauchy-Schwarz Inequality, we obtain

$$B_{22i} \leq \frac{1}{p_{*i}^2}\sqrt{\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N^2(\hat{\rho}-\rho)^4\right]}\sqrt{\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]}.$$

Using the inequality once more leads to

$$\begin{aligned}
\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}[B_{22i}] &\leq \sqrt{\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[N^2(\hat{\rho}-\rho)^4\right]}\sqrt{\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^4}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\right]} \\
&\leq o_{u.\pi}(N^+)\sqrt{\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^4}\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right]\right]}.
\end{aligned}$$

The second inequality follows from Assumption 3.6.

According to Lemma A.8(a) (see Section A.1.4),

$$\mathbb{E}_{\theta,\pi,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{m*}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right] \leq ML_N^4 p_i^4\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}),$$

where $L_N = o(N^+)$ was defined in (A.5). This leads to the bound

$$\begin{aligned}
\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}[B_{22i}] &\leq o_{u.\pi}(N^+)ML_N^2\sqrt{\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i}\left[\left(\frac{p_i}{p_{*i}}\right)^4\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right]} \\
&= o_{u.\pi}(N^+)ML_N^2\sqrt{\int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\left(\frac{p_i}{p_{*i}}\right)^4 p_i d\hat{\lambda}_i dy_{i0}} \\
&\leq o_{u.\pi}(N^+)M_*L_N^2\sqrt{\int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\left(\frac{p_i}{p_{*i}}\right)^4 d\hat{\lambda}_i dy_{i0}} \\
&\leq o_{u.\pi}(N^+).
\end{aligned}$$

The second inequality holds because the density p_i is bounded from above and M_* is a constant. The last inequality is proved in Lemma A.8(e) (see Section A.1.4).

We deduce that $B_{2i} = o_{u.\pi}(N^+)$. A similar argument can be used to establish that $B_{1i} = o_{u.\pi}(N^+)$.

Step 2.2. Recall from (A.22) that over $\mathcal{T}_{m^*}$,

$$|A_{1i}| \leq 2C_N = o_{u.\pi}(N^+).$$

Then,

$$
\begin{aligned}
& N\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[A^2_{1i}\mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\ \mathcal{T}_{\hat{\rho}}\ \mathcal{T}_{\bar{U}}\ \mathcal{T}_{Y_0}\ \mathcal{T}_{m*}\ \mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\ \mathcal{T}_{p(\cdot)})\ \mathcal{T}^c_{\tilde{p}(\cdot)}] \\
\leq\ & o_{u.\pi}(N^+)N\mathbb{P}^{\mathcal{Y}^N}_{\theta,\pi}(\mathcal{T}_{\hat{\sigma}^2}\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}^c_{\tilde{p}(\cdot)}).
\end{aligned}
$$

Notice that

$$
\begin{aligned}
\mathcal{T}^c_{\tilde{p}(\cdot)} &= \left\{\hat{p}_i^{(-i)} - p_{*i} + (\hat{\rho}-\rho)R_{2i}(\widetilde{\rho}) < -\frac{p_{*i}}{2}\right\} \\
&\subset \left\{\hat{p}_i^{(-i)} - p_{*i} - |\hat{\rho}-\rho||R_{2i}(\widetilde{\rho})| < -\frac{p_{*i}}{2}\right\} \\
&\subset \left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} \cup \left\{|\hat{\rho}-\rho||R_{2i}(\widetilde{\rho})| > \frac{p_{*i}}{4}\right\}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
& N\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}(\mathcal{T}_{\hat{\sigma}^2}\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}^c_{\tilde{p}(\cdot)}) \\
\leq\ & N\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} \\
& + N\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left[\left\{|\hat{\rho}-\rho||R_{2i}(\widetilde{\rho})| > \frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] \\
\leq\ & N\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} + \frac{16L_N^4}{p_{*i}^2}\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left[R_{2i}(\widetilde{\rho})^2\mathbb{I}(\mathcal{T}_{\hat{\sigma}^2}\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\lambda}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] \\
\leq\ & N\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} + \frac{ML_N^6}{p_{*i}^2}p_i^2\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}).
\end{aligned}
$$

The first inequality is based on the superset of $\mathcal{T}^c_{\tilde{p}(\cdot)}$ from above. The second inequality is based on Chebychev's inequality and truncation $\mathcal{T}_{\hat{\rho}}$. The third inequality uses a version of the result in Lemma A.8(a) in which the remainder is raised to the power of two instead of to the power of four. Assumption 3.4 implies that p_i is bounded from above:

$$
p_i = \int p(\hat{\lambda}|\lambda)\pi(Y_{i0}|\lambda)\pi(\lambda)d\lambda \leq \widetilde{M}\int \pi(\lambda)d\lambda = \widetilde{M} < \infty, \tag{A.27}
$$

because $p(\hat{\lambda}|\lambda)$ is the density of a $N(\lambda, \sigma^2/T)$ and $\pi(Y_{i0}|\lambda)$ is bounded for every $\pi \in \Pi$ according to Assumption 3.4. Thus, in the previous calculation we can absorb one of the p_i terms in the constant M.

In Lemma A.8(f) (see Section A.1.4) we apply Bernstein's inequality to bound the probability $\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}$ uniformly over $(\hat{\lambda}_i, Y_{i0})$ in the region $\mathcal{T}_{\hat{\lambda}Y_0}$, showing that

$$N\mathbb{E}^{\mathcal{Y}^i}_{\theta,\pi}\left[\mathbb{P}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] = o_{u.\pi}(N^+),$$

as desired. Moreover, according to Lemma A.8(e) (see Section A.1.4)

$$\mathbb{E}^{\mathcal{Y}^i}_{\theta,\pi}\left[\frac{p_i}{p_{*i}^2}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] = \int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\left(\frac{p_i}{p_{*i}}\right)^2 d\hat{\lambda}_i dy_{i0} = o_{u.\pi}(N^+),$$

which gives us the required result for Step 2.2. Combining the results from Steps 2.1 and 2.2 yields (A.19).

The bound in (A.15) now follows from (A.17), (A.18), and (A.19), which completes the proof of the lemma. ■

A.1.3.2 Term A_{2i}

Lemma A.4 *Suppose the assumptions in Theorem 3.7 hold. Then, for every $\epsilon_0 > 0$,*

$$\limsup_{N\to\infty}\sup_{\pi\in\Pi}\frac{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(m_*(\hat{\lambda}_i, y_{i0};\pi, B_N) - \lambda_i\right)^2\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}^i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 1.$$

Proof of Lemma A.4. Recall that $m_*(\hat{\lambda}_i, y_{i0};\pi, B_N)$ can be interpreted as the posterior mean of λ_i under the $p_*(\hat{\lambda}_i, y_{i0};\pi)$ defined in (16). We will use $\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}[\cdot]$ to denote the integral

$$\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}[\cdot] = \int [\cdot]p_*(\hat{\lambda}|\lambda)\pi_*(y_0|\lambda)\pi(\lambda)d(\hat{\lambda}, \lambda, y_0),$$

where

$$\begin{aligned} p_*(\hat{\lambda}|\lambda) &= \frac{1}{\sqrt{\sigma^2/T + B_N^2}}\phi\left(\frac{\hat{\lambda} - \lambda}{\sqrt{\sigma^2/T + B_N^2}}\right) \\ \pi_*(y_0|\lambda) &= \int \frac{1}{B_N}\phi\left(\frac{y_0 - \tilde{y}_0}{B_N}\right)\pi(\tilde{y}_0|\lambda)d\tilde{y}_0. \end{aligned}$$

The desired result follows if we can show that

$$\text{(i)} \qquad \limsup_{N\to\infty} \limsup_{\pi\in\Pi} \frac{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}\left[\left(m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i\right)^2\right] + N^{\epsilon_0}}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(\lambda_i - m(\hat{\lambda}_i, y_{i0}; \pi)\right)^2\right] + N^{\epsilon_0}} \leq 1$$

$$\text{(ii)} \qquad \limsup_{N\to\infty} \limsup_{\pi\in\Pi} \frac{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i\right)^2\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}\left[\left(m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i\right)^2\right] + N^{\epsilon_0}} \leq 1.$$

Part (i): Notice that the denominator is bounded below by N^{ϵ_0}. We will proceed by constructing an upper bound for the numerator. Using the fact that the posterior mean minimizes the integrated risk, we obtain

$$\begin{aligned}
& N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}\left[\left(m_*(\hat{\lambda}_i, y_{i0}; \pi, B_N) - \lambda_i\right)^2\right] \\
\leq\ & N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}\left[\left(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\right)^2\right] \\
\leq\ & N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}\left[\left(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\right)^2 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_m\mathcal{T}_\lambda)\right] \\
& + N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}\left[\left(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\right)^2 \left(\mathbb{I}(\mathcal{T}^c_{\hat{\lambda}Y_0}) + \mathbb{I}(\mathcal{T}^c_{p(\cdot)}\mathcal{T}_{\hat{\lambda}Y_0}) + \mathbb{I}(\mathcal{T}^c_m) + \mathbb{I}(\mathcal{T}^c_\lambda)\right)\right] \\
=\ & B_{1i} + B_{2i},
\end{aligned}$$

say.

A bound for B_{1i} can be obtained as follows:

$$\begin{aligned}
B_{1i} &= N\int\int \left(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\right)^2 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_m\mathcal{T}_\lambda) p_*(\hat{\lambda}_i|\lambda_i)\pi_*(y_{i0}|\lambda_i)\pi(\lambda_i) d(\hat{\lambda}_i, \lambda_i, y_{i0}) \\
&\leq (1+o(1))N\int\int \left(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\right)^2 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_m\mathcal{T}_\lambda) p(\hat{\lambda}_i|\lambda_i)\pi(y_{i0}|\lambda_i)\pi(\lambda_i) d(\hat{\lambda}_i, \lambda_i, y_{i0}) \\
&\leq (1+o(1))N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(\lambda_i - m(\hat{\lambda}_i, y_{i0}; \pi)\right)^2\right].
\end{aligned}$$

The first inequality is based on Assumption 3.4 and an argument similar to the one used in the analysis of term I in the proof of Lemma A.7. The $o(1)$ term does not depend on $\pi \in \Pi$.

To derive a bound for B_{2i} first consider the inequalities

$$\begin{aligned}\Big(m(\hat{\lambda}_i, y_{i0}; \pi) - \lambda_i\Big)^2 &\leq 2m^2(\hat{\lambda}_i, y_{i0}; \pi)\big(\mathbb{I}(\mathcal{T}_m) + \mathbb{I}(\mathcal{T}_m^c)\big) + 2\lambda_i^2\big(\mathbb{I}(\mathcal{T}_\lambda) + \mathbb{I}(\mathcal{T}_\lambda^c)\big) \\ &\leq 4C_N^2 + 2m(\hat{\lambda}_i, y_{i0}; \pi)^2\mathbb{I}(\mathcal{T}_m^c) + 2\lambda_i^2\mathbb{I}(\mathcal{T}_\lambda^c).\end{aligned}$$

Thus,

$$\begin{aligned}B_{2i} &\leq 4C_N^2 N\big(\mathbb{P}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}^c) + \mathbb{P}(\mathcal{T}_{\hat{\lambda}Y_0}^c) + \mathbb{P}(\mathcal{T}_m^c) + \mathbb{P}(\mathcal{T}_\lambda^c)\big) \\ &\quad + 8N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i,\lambda_i}\big[m(\hat{\lambda}_i, y_{i0}; \pi)^2\mathbb{I}(\mathcal{T}_m^c)\big] + 8N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i,\lambda_i}\big[\lambda_i^2\mathbb{I}(\mathcal{T}_\lambda^c)\big].\end{aligned}$$

Notice that $C_N^2 = o_{u.\pi}(N^+)$, $N\mathbb{P}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}^c) = o_{u.\pi}(N^{\epsilon_0})$ (see Step 1.1), $N\mathbb{P}(\mathcal{T}_{\hat{\lambda}Y_0}^c) = o_{u.\pi}(N^+)$ (see Step 1.2), and $N\mathbb{P}(\mathcal{T}_m^c) = o_{u.\pi}(N^+)$ (see Truncation 1). Also, notice that

$$\begin{aligned}N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^i,\lambda_i}\big[\lambda_i^2\mathbb{I}(\mathcal{T}_\lambda^c)\big] &= N\int_{|\lambda|>C_N}\lambda^2\pi(\lambda)d\lambda \\ &\leq \sqrt{\int_{|\lambda|>C_N}\lambda^4\pi(\lambda)d\lambda}\sqrt{N^2\int_{|\lambda|>C_N}\pi(\lambda)d\lambda} \\ &\leq M\sqrt{N^2\int_{|\lambda|>C_N}\pi(\lambda)d\lambda} \\ &= o_{u.\pi}(N^+).\end{aligned}$$

The first inequality is the Cauchy-Schwartz inequality, the second inequality holds by Assumption 3.2, and the last line follows from calculations similar to the ones in (A.6). Therefore,

$$B_{2i} \leq o_{u.\pi}(N^{\epsilon_0}).$$

Combining the bounds for B_{1i} and B_{2i}, we have

$$\begin{aligned}
&\limsup_{N\to\infty}\limsup_{\pi\in\Pi} \frac{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{*,\theta,\pi}\left[\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\right]+N^{\epsilon_0}}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(\lambda_i-m(\hat{\lambda}_i,y_{i0};\pi)\right)^2\right]+N^{\epsilon_0}} \\
&\le \limsup_{N\to\infty}\limsup_{\pi\in\Pi} \frac{(1+o(1))N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(m(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\right]+o_{u.\pi}(N^{\epsilon_0})+N^{\epsilon_0}}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(\lambda_i-m(\hat{\lambda}_i,y_{i0};\pi)\right)^2\right]+N^{\epsilon_0}} \\
&\le \limsup_{N\to\infty}\limsup_{\pi\in\Pi} \frac{(1+o(1))\left[N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(m(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\right]+N^{\epsilon_0}\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_{\theta,\pi}\left[\left(\lambda_i-m(\hat{\lambda}_i,y_{i0};\pi)\right)^2\right]+N^{\epsilon_0}} \\
&= 1,
\end{aligned}$$

where the term $o(1)$ holds uniformly in $\pi\in\Pi$. We have the required result for Part (i).

Part (ii): The proof of Part (ii) is similar to that of Part (i). We construct an upper bound for the numerator as follows

$$\begin{aligned}
&N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{\theta,\pi}\left[\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\right] \\
&\le\ N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{\theta,\pi}\left[\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{m*}\mathcal{T}_{\lambda})\right] \\
&\quad +N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{\theta}\left[\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\left(\mathbb{I}(\mathcal{T}^c_{\hat{\lambda}Y_0})+\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}^c_{p(\cdot)})+\mathbb{I}(\mathcal{T}^c_{m*})+\mathbb{I}(\mathcal{T}^c_{\lambda})\right)\right] \\
&=\ B_{1i}+B_{2i},
\end{aligned}$$

say. Now consider the term B_{1i}:

$$\begin{aligned}
B_{1i} &= N\int\int\int\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2 p_*(\hat{\lambda}_i|\lambda_i)\pi_*(y_{i0}|\lambda_i)\frac{p(\hat{\lambda}_i|\lambda_i)\pi(y_{i0}|\lambda_i)}{p_*(\hat{\lambda}_i|\lambda_i)\pi_*(y_{i0}|\lambda_i)}\pi(\lambda_i) \\
&\qquad \times\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{m*}\mathcal{T}_{\lambda})d(\hat{\lambda}_i,\lambda_i,y_{i0}) \\
&= (1+o(1))N\int\int\int\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2 p_*(\hat{\lambda}_i|\lambda_i)\pi_*(y_{i0}|\lambda_i)\pi(\lambda_i) \\
&\qquad \times\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{m*}\mathcal{T}_{\lambda})d(\hat{\lambda}_i,d\lambda_i,dy_{i0}) \\
&\le (1+o(1))N\mathbb{E}^{\mathcal{Y}_i,\lambda_i}_{*,\theta,\pi}\left[\left(m_*(\hat{\lambda}_i,y_{i0};\pi,B_N)-\lambda_i\right)^2\right],
\end{aligned}$$

where the $o(1)$ term is uniform in $\pi \in \Pi$. Using the a similar argument as in Part (i) we can establish that $B_{2i} = o_{u.\pi}(N^{\epsilon_0})$, which leads to the desired result. ■

A.1.3.3 Term A_{3i}

Lemma A.5 *Suppose the assumptions in Theorem 3.7 hold. Then,*

$$N\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\left(\hat{\rho}-\rho\right)^2 Y_{iT}^2\right] = o_{u.\pi}(N^+).$$

Proof of Lemma A.5. Using the Cauchy-Schwarz inequality, we can bound

$$\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\left(\sqrt{N}(\hat{\rho}-\rho)\right)^2 Y_{iT}^2\right] \;\leq\; \sqrt{\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\left(\sqrt{N}(\hat{\rho}-\rho)\right)^4\right]\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[Y_{iT}^4\right]}.$$

By Assumption 3.6, we have

$$\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\left(\sqrt{N}(\hat{\rho}-\rho)\right)^4\right] \leq o_{u.\pi}(N^+).$$

For the second term, write

$$Y_{iT} = \rho^T Y_{i0} + \sum_{\tau=0}^{T-1}\rho^\tau(\lambda_i + U_{iT-\tau}).$$

Using the C_r inequality and noting that T is finite and $U_{it} \sim iidN(0,\sigma^2)$, we deduce that there is a finite constant M that does not depend on $\pi \in \Pi$ such that

$$\begin{aligned}
\mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[Y_{iT}^4\right] &\leq M\left(\mathbb{E}_{\theta}^{\mathcal{Y}^N}\left[Y_{i0}^4\right] + \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[\lambda_i^4\right] + \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N}\left[U_{i1}^4\right]\right) \\
&= o_{u.\pi}(N^+).
\end{aligned}$$

The last line holds according to Assumption 3.2 and because U_{i1} is normally distributed and therefore all its moments are finite. ■

A.1.4 Further Details

We now provide more detailed derivations for some of the bounds used in Section A.1.3. Recall that

$$\begin{aligned} R_{1i}(\rho) &= -\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)^2(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right) \\ &\quad +\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N^2}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right) \\ R_{2i}(\rho) &= \frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right) \end{aligned}$$

For expositional purposes, our analysis focuses on the slightly simpler term $R_{2i}(\widetilde{\rho})$. The extension to $R_{1i}(\widetilde{\rho})$ is fairly straightforward. By definition,

$$\hat{\lambda}_j(\widetilde{\rho})-\hat{\lambda}_i(\widetilde{\rho})=\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)-(\widetilde{\rho}-\rho)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1}).$$

Therefore,

$$\begin{aligned} R_{2i}(\widetilde{\rho}) &= \frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-(\widetilde{\rho}-\rho)\left(\frac{\bar{Y}_{j,-1}-\bar{Y}_{i,-1}}{B_N}\right)\right) \\ &\quad \times\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-(\widetilde{\rho}-\rho)\left(\frac{\bar{Y}_{j,-1}-\bar{Y}_{i,-1}}{B_N}\right)\right) \\ &\quad \times(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right). \end{aligned}$$

Consider the region $\mathcal{T}_{\hat{\rho}}\cap\mathcal{T}_{\bar{U}}\cap\mathcal{T}_{Y_0}$. First, using (A.12) we can bound

$$\max_{1\leq i,j\leq N}\left|(\hat{\rho}-\rho)(\bar{Y}_{j,-1}-\bar{Y}_{i,-1})\right|\leq\frac{M}{L_N}.$$

Thus,

$$\begin{aligned}
&\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-(\widetilde{\rho}-\rho)\left(\frac{\bar{Y}_{j,-1}-\bar{Y}_{i,-1}}{B_N}\right)\right)\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}) \\
\leq\ & \phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}+\left(\frac{M}{L_N B_N}\right)\right)\mathbb{I}\left\{\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\leq-\frac{M}{L_N B_N}\right\} \\
& +\phi(0)\mathbb{I}\left\{\left|\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right|\leq\frac{M}{L_N B_N}\right\} \\
& +\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-\left(\frac{M}{L_N B_N}\right)\right)\mathbb{I}\left\{\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\geq\frac{M}{L_N B_N}\right\} \\
=\ & \bar{\phi}\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right),
\end{aligned}$$

say. The function $\bar{\phi}(x)$ is flat for $|x|<\frac{M}{L_N B_N}$ and is proportional to a Gaussian density outside of this region.

Second, we can use the bound

$$\left|\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-(\widetilde{\rho}-\rho)\left(\frac{\bar{Y}_{j,-1}-\bar{Y}_{i,-1}}{B_N}\right)\right|\leq\left|\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right|+\frac{M}{L_N B_N}.$$

Third, for the region $\mathcal{T}_{\bar{U}}\cap\mathcal{T}_{Y_0}$ we can deduce from (A.11) that

$$\max_{1\leq i,j\leq N}|\bar{Y}_{j,-1}-\bar{Y}_{i,-1}|\leq ML_N.$$

Therefore,

$$\left|\bar{Y}_{j,-1}-\bar{Y}_{i,-1}\right|\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right)\leq\frac{ML_N}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right).$$

Now, define the function

$$\bar{\phi}_*(x)=\bar{\phi}\left(x\right)\left(|x|+\frac{M}{L_N B_N}\right).$$

Because for random variables with bounded densities and Gaussian tails all moments exist and because $L_N B_N>1$ by definition of L_N in (A.5), the function $\bar{\phi}_*(x)$ has the property that for any finite positive integer m there is a finite constant M such that

$$\int\bar{\phi}_*(x)^m dx\leq M.$$

Combining the previous results we obtain the following bound for $R_{2i}(\widetilde{\rho})$:

$$\left|R_{2i}(\widetilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0})\right| \leq \frac{ML_N}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right). \tag{A.28}$$

For the subsequent analysis it is convenient define the function

$$f(\hat{\lambda}_j-\hat{\lambda}_i, Y_{j0}-Y_{i0}) = \frac{1}{B_N^2}\bar{\phi}_*\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right). \tag{A.29}$$

In the remainder of this section we will state and prove three technical lemmas that establish moment bounds for $R_{1i}(\widetilde{\rho})$ and $R_{2i}(\widetilde{\rho})$. The bounds are used in Section A.1.3. We will abbreviate $\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}^i}[\cdot] = \mathbb{E}_i[\cdot]$ and simply use $\mathbb{E}[\cdot]$ to denote $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}[\cdot]$.

Lemma A.6 *Suppose the assumptions in Theorem 3.7 hold. Then, for any finite positive integer $m \geq 1$, over the regions $\mathcal{T}_{\hat{\lambda}Y_0}$ and $\mathcal{T}_{p(\cdot)}$, there exists a finite constant M that does not depend on π such that*

$$\mathbb{E}_i\big[f^m(\hat{\lambda}_j-\hat{\lambda}_i, Y_{j0}-Y_{i0})\big] \leq \frac{M}{B_N^{2(m-1)}}p_i.$$

Proof of Lemma A.6. We have

$$\begin{aligned}
&\mathbb{E}_i\big[f^m(\hat{\lambda}_j-\hat{\lambda}_i, Y_{j0}-Y_{i0})\big] \\
&= \frac{1}{B_N^{2(m-1)}}\int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m p(\hat{\lambda},y_0)d(\hat{\lambda},y_0) \\
&= \frac{1}{B_N^{2(m-1)}}\int\left[\int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m p(\hat{\lambda},y_0|\lambda_i)d(\hat{\lambda},y_0)\right]\pi(\lambda_i)d\lambda_i \\
&= \frac{1}{B_N^{2(m-1)}}\int_{\mathcal{T}_\lambda}\left[\int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m p(\hat{\lambda},y_0|\lambda_i)d(\hat{\lambda},y_0)\right]\pi(\lambda_i)d\lambda_i \\
&\quad +\frac{1}{B_N^{2(m-1)}}\int_{\mathcal{T}_\lambda^c}\left[\int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m p(\hat{\lambda},y_0|\lambda_i)d(\hat{\lambda},y_0)\right]\pi(\lambda_i)d\lambda_i.
\end{aligned}$$

The required result of the lemma follows if we show

$$
\begin{aligned}
I = {} & \frac{1}{p(\hat{\lambda}_i, Y_{i0})} \int_{\mathcal{T}_\lambda} \left[\int \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right)^m \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right)^m p(\hat{\lambda}, y_0 | \lambda_i) d(\hat{\lambda}, y_0) \right] \pi(\lambda_i) d\lambda_i \\
& \leq M \qquad \text{(A.30)} \\
II = {} & \frac{1}{p(\hat{\lambda}_i, Y_{i0})} \int_{\mathcal{T}_\lambda^c} \left[\int \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right)^m \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right)^m p(\hat{\lambda}, y_0 | \lambda_i) d(\hat{\lambda}, y_0) \right] \pi(\lambda_i) d\lambda_i \\
& \leq M \qquad \text{(A.31)}
\end{aligned}
$$

over the regions $\mathcal{T}_{\hat{\lambda} Y_0}$ and $\mathcal{T}_{p(\cdot)}$ and uniformly in π.

For (A.30), notice that the inner integral of term I is

$$
\begin{aligned}
& \int \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right)^m \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right)^m p(\hat{\lambda}, y_0 | \lambda) d(\hat{\lambda}, y_0) \\
= {} & \int \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right)^m \frac{1}{\sigma / \sqrt{T}} \exp \left(-\frac{1}{2} \left(\frac{\hat{\lambda} - \lambda_i}{\sigma / \sqrt{T}} \right)^2 \right) d\hat{\lambda} \\
& \times \int \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right)^m \pi(y_0 | \lambda) dy_0 \\
= {} & I_1 \times I_2,
\end{aligned}
$$

say.

Notice that

$$
\begin{aligned}
I_1 &= \int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)d\hat{\lambda} \\
&= \int \bar{\phi}_*(\lambda^*)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i+B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^* \\
&= \int \bar{\phi}_*(\lambda^*)^m \exp\left(-\left((\hat{\lambda}_i-\lambda_i)B_N\lambda^*\right)\frac{1}{\sigma^2/T}\right)\exp\left(-\frac{1}{2}\left(\frac{B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^* \\
&\quad \times \left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&\leq \int \bar{\phi}_*(\lambda^*)^m \exp\left(-\left(\frac{(\hat{\lambda}_i-\lambda_i)B_N}{\sigma^2/T}\right)\lambda^*\right)d\lambda^* \\
&\quad \times \left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&\leq M\left(\int_0^\infty \bar{\phi}_*(\lambda^*)^m \exp\left(v_N\lambda^*\right)d\lambda^*\right)\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&\leq M\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&= Mp(\hat{\lambda}_i|\lambda_i),
\end{aligned}
$$

where $v_N = \frac{T}{\sigma^2}(C'_N + C_N)B_N$.

Here, for the second equality, we used the change-of-variable $\lambda_* = (\hat{\lambda}-\hat{\lambda}_i)/B_N$ to replace $\hat{\lambda}$. The first inequality holds because the exponential function $\exp\left(-\frac{1}{2}\left(\frac{B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)$ is bounded by one. Moreover, under truncations $\mathcal{T}_{\hat{\lambda}Y_0}$ and $\mathcal{T}_\lambda$, $|\hat{\lambda}_i| \leq C'_N$ and $|\lambda_i| \leq C_N$. According to Assumption 3.3, $v_N = \frac{T}{\sigma^2}(C'_N + C_N)B_N = o(1)$. Thus, the last inequality holds because $\int_0^\infty \bar{\phi}_*(\lambda^*)^m \exp\left(v_N\lambda^*\right)d\lambda^*$ is finite.

We now proceed with a bound for the second integral, I_2. Using the fact that the Gaussian pdf $\phi(x)$ is bounded and by Assumption 3.4, we can write

$$\begin{aligned} I_2 &= \int \frac{1}{B_N}\phi\left(\frac{y_0 - Y_{i0}}{B_N}\right)^m \pi(y_0|\lambda)dy_0 \\ &\leq M\int \frac{1}{B_N}\phi\left(\frac{y_0 - Y_{i0}}{B_N}\right)\pi(y_0|\lambda)dy_0 \\ &= M\big(1+o(1)\big)\pi(Y_{i0}|\lambda), \end{aligned}$$

uniformly in $|y_0| \leq C'_N$ and $|\lambda| \leq C_N$ and in $\pi \in \Pi$.

Combining the bounds for I_1 and I_2 and integrating over λ, we obtain

$$\begin{aligned} I &\leq M\frac{1}{p(\hat{\lambda}_i, Y_{i0})}\int_{\mathcal{T}_\lambda}\Big[p(\hat{\lambda}_i|\lambda_i)\pi(Y_{i0}|\lambda_i)\Big]\pi(\lambda_i)d\lambda_i \\ &\leq M\frac{1}{p(\hat{\lambda}_i, Y_{i0})}\int p(\hat{\lambda}_i|\lambda_i)\pi(Y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i = M, \end{aligned}$$

as required for (A.30).

Next, for (A.31), since $\bar{\phi}_*(x), \phi(x), p(\hat{\lambda}, y_0|\lambda_i)$ are bounded uniformly in π and $p(\hat{\lambda}_i, Y_{i0}) > N^{\epsilon-1}$ over $\mathcal{T}_{p(\cdot)}$, we have

$$\begin{aligned} II &\leq \frac{M}{p(\hat{\lambda}_i, Y_{i0})B_N^2}\int_{\mathcal{T}_\lambda^c}\pi(\lambda_i)d\lambda_i \\ &\leq MN^{-\epsilon}\left(\frac{1}{B_N^2}\right)\left(N\int_{\mathcal{T}_\lambda^c}\pi(\lambda_i)d\lambda_i\right) \\ &\leq MN^{-\epsilon}o_{u.\pi}(N^+)o_{u.\pi}(N^+) \\ &\leq M, \end{aligned}$$

where the second-to-last line holds because according to Assumption 3.3 $1/B_N^2 = o_{u.\pi}(N^+)$ and because of the tail bound in (A.6). This yields the required result for (A.31). ■

Lemma A.7 *Suppose the assumptions required for Theorem 3.7 are satisfied. Then,*

$$\begin{aligned} &(a)\ \sup_{\pi\in\Pi}\ \sup_{(\hat{\lambda}_i, Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\left|\frac{p_{*i}}{p_i} - 1\right| = o(1), \\ &(b)\ \sup_{\pi\in\Pi}\ \sup_{(\hat{\lambda}_i, Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}}\left|\frac{p_i}{p_{*i}} - 1\right| = o(1). \end{aligned}$$

Proof of Lemma A.7.

Part (a). Denote

$$
\begin{aligned}
p(\hat{\lambda}_i, y_{i0}|\lambda_i) &= \frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)\pi(y_{i0}|\lambda_i)\\
p_*(\hat{\lambda}_i, y_{i0}|\lambda_i) &= \frac{1}{\sqrt{B_N^2+\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{B_N^2+\sigma^2/T}}\right)\left[\int\frac{1}{B_N}\phi\left(\frac{y_{i0}-\tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}\right],
\end{aligned}
$$

so that

$$
p_i = \int p(\hat{\lambda}_i, y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i, \quad p_{*i} = \int p_*(\hat{\lambda}_i, y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i.
$$

Notice that

$$
\begin{aligned}
\left|\frac{p_{*i}}{p_i}-1\right| &= \left|\frac{p_{*i}-p_i}{p_i}\right|\\
&\leq \frac{1}{p_i}\int\left|p_*(\hat{\lambda}_i, y_{i0}|\lambda_i)-p(\hat{\lambda}_i, y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i\\
&= \frac{1}{p_i}\int_{\mathcal{T}_\lambda}\left|p_*(\hat{\lambda}_i, y_{i0}|\lambda_i)-p(\hat{\lambda}_i, y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i\\
&\quad+\frac{1}{p_i}\int_{\mathcal{T}_\lambda^c}\left|p_*(\hat{\lambda}_i, y_{i0}|\lambda_i)-p(\hat{\lambda}_i, y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i\\
&= I + II, \text{ say.}
\end{aligned}
$$

For term I, since $|\lambda_i| \leq C_N$ in the region $\mathcal{T}_\lambda$ and $|\hat{\lambda}_i| \leq C_N'$ in the region $\mathcal{T}_{\hat{\lambda}Y_0}$, we can choose a constant M that does not depend on π such that for N sufficiently large

$$
\begin{aligned}
\frac{1}{\sqrt{B_N^2+\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{B_N^2+\sigma^2/T}}\right) &= \frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)\\
&\quad\times\frac{\sqrt{\sigma^2/T}}{\sqrt{B_N^2+\sigma^2/T}}\exp\left\{\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{B_N^2+\sigma^2/T}}\right)^2\frac{B_N^2}{\sigma^2/T}\right\}\\
&\leq \frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)\exp(M(C_N'+C_N)^2B_N^2)\\
&= \frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)(1+o(1)),
\end{aligned}
$$

where the inequality holds by Assumption 3.3 which implies that $(C_N'+C_N)B_N = o(1)$, and the $o(1)$ term in the last line is uniform in $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}$ and in $\pi \in \Pi$.

According to Assumption 3.4,

$$\int \frac{1}{B_N}\phi\left(\frac{y_{i0}-\tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0} = (1+o(1))\pi(y_{i0}|\lambda_i)$$

uniformly in $|y_{i0}| \le C'_N$ and $|\lambda_i| \le C_N$ and in $\pi \in \Pi$.

Then,

$$\begin{aligned}
I &= \frac{1}{p_i}\int_{\mathcal{T}_\lambda}\left|p_*(\hat{\lambda}_i, y_{i0}|\lambda_i) - p(\hat{\lambda}_i, y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i \\
&= \frac{1}{p_i}\int_{\mathcal{T}_\lambda}\left|\frac{1}{\sqrt{B_N^2+\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{B_N^2+\sigma^2/T}}\right)\int\frac{1}{B_N}\phi\left(\frac{y_{i0}-\tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}\right. \\
&\qquad\left. -\frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)\pi(y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i \\
&\le |(1+o(1))^2-1|\frac{1}{p_i}\int_{\mathcal{T}_\lambda}\frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sqrt{\sigma^2/T}}\right)\pi(y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i \\
&\le |(1+o(1))^2-1| = o(1).
\end{aligned}$$

Note that the $o(1)$ term does not depend on $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}$ nor on $\pi \in \Pi$.

For term II, calculations similar to the one in (A.27) imply that the densities $p_*(\hat{\lambda}_i, y_{i0}|\lambda_i)$ and $p(\hat{\lambda}_i, y_{i0}|\lambda_i)$ are bounded, say, by M. Thus, we have

$$\begin{aligned}
II &= \frac{1}{p_i}\int_{\mathcal{T}_\lambda^c}\left|p_*(\hat{\lambda}_i, y_{i0}|\lambda_i) - p(\hat{\lambda}_i, y_{i0}|\lambda_i)\right|\pi(\lambda_i)d\lambda_i \\
&\le \frac{2M}{p_i}\int_{\mathcal{T}_\lambda^c}\pi(\lambda_i)d\lambda_i \\
&\le 2M\sup_{\pi\in\Pi} N^{1-\epsilon}\int_{\mathcal{T}_\lambda^c}\pi(\lambda_i)d\lambda_i \\
&= o(1),
\end{aligned}$$

where the second inequality holds since $p_i > \frac{N^\epsilon}{N}$ under the truncation $\mathcal{T}_{p(\cdot)}$ and the last line holds according to (A.6).

Combining the upper bounds of I and II yields the required result for Part (a).

Part (b). According to Part (a),

$$p_{*i} = p_i(1+o(1))$$

uniformly in $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}$ and in $\pi \in \Pi$. Then, for some finite constant M that does not depend on $(\hat{\lambda}_i, Y_{i0})$ and π,

$$
\begin{aligned}
\sup_{\pi\in\Pi} \sup_{(\hat{\lambda}_i,Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}} \frac{|p_i - p_{*i}|}{p_{*i}} &= \sup_{\pi\in\Pi} \sup_{(\hat{\lambda}_i,Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}} \frac{|p_i - p_{*i}|}{p_i}\frac{p_i}{p_{*i}} \\
&= \sup_{\pi\in\Pi} \sup_{(\hat{\lambda}_i,Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}} \frac{|p_i - p_{*i}|}{p_i}\frac{p_i}{p_i(1+o(1))} \\
&\le M \sup_{\pi\in\Pi} \sup_{(\hat{\lambda}_i,Y_{i0})\in\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}} \frac{|p_i - p_{*i}|}{p_i} \\
&= o(1),
\end{aligned}
$$

as required for Part (b). ■

Lemma A.8 *Under the assumptions required for Theorem 3.7, we obtain the following bounds:*

(a) $\mathbb{E}_i\big[R_{2i}^4(\widetilde{\rho})\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\big] \le M L_N^4 p_i^4 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})$,

(b) $\mathbb{E}_i\big[R_{1i}^4\mathbb{I}(\mathcal{T}_{\tilde{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\big] \le M \frac{L_N^4}{B_N^4} p_i^4 \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})$,

(c) $\mathbb{E}_i\left[N(\hat{p}_i^{(-i)} - p_{*i})^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right] \le \frac{M}{B_N^2} p_i \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})$,

(d) $\mathbb{E}_i\left[N(d\hat{p}_i^{(-i)} - dp_{*i})^2\mathbb{I}(\mathcal{T}_{\hat{\rho}}\mathcal{T}_{\bar{U}}\mathcal{T}_{Y_0}\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}\mathcal{T}_{\tilde{p}(\cdot)})\right] \le \frac{M}{B_N^2} p_i \mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})$, *where M is a finite constant that does not depend on $\pi \in \Pi$.*

(e) For any finite $m > 1$, $\int_{\mathcal{T}_{\hat{\lambda}Y_0}\cap\mathcal{T}_{p(\cdot)}} \left(\frac{p_i}{p_{*i}}\right)^m d\hat{\lambda}_i dy_{i0} = o_{u.\pi}(N^+)$.

(f) $N\mathbb{E}\big[\mathbb{P}_i\{\hat{p}_i^{(-i)} - p_{*i} < -p_{*i}/4\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\big] = o_{u.\pi}(N^+)$.

Proof of Lemma A.8. **Part (a).** Recall the following definitions

$$
\begin{aligned}
\bar{\phi}(x) &= \phi\left(x + \frac{M}{L_N B_N}\right)\mathbb{I}\left\{x \le -\frac{M}{L_N B_N}\right\} + \phi(0)\mathbb{I}\left\{|x| \le \frac{M}{L_N B_N}\right\} \\
&\quad + \phi\left(x - \frac{M}{L_N B_N}\right)\mathbb{I}\left\{x \ge \frac{M}{L_N B_N}\right\} \\
\bar{\phi}_*(x) &= \bar{\phi}(x)\left(|x| + \frac{M}{L_N B_N}\right).
\end{aligned}
$$

First, recall that according to (A.28) and (A.29), in the region $\mathcal{T}_{\hat{\rho}} \cap \mathcal{T}_{\bar{U}} \cap \mathcal{T}_{Y_0}$

$$|R_{2i}(\widetilde{\rho})| \leq \frac{ML_N}{N-1} \sum_{j \neq i}^{N} f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}).$$

Then,

$$\begin{aligned}
|R_{2i}(\widetilde{\rho})|^4 &\leq \left[\frac{ML_N}{N-1} \sum_{j \neq i}^{N} f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) \right]^4 \\
&= \left[\frac{ML_N}{N-1} \sum_{j \neq i}^{N} \Big\{ f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) - \mathbb{E}_i[f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})] \right. \\
&\quad \left. + \mathbb{E}_i[f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})] \Big\} \right]^4 \\
&\leq ML_N^4 \left[\frac{1}{N-1} \sum_{j \neq i}^{N} \Big(f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) - \mathbb{E}_i[f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})] \Big) \right]^4 \\
&\quad + ML_N^4 \Big[\mathbb{E}_i[f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})] \Big]^4 \\
&= ML_N^4 \big(A_1 + A_2 \big),
\end{aligned}$$

say. The first equality holds since $f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})$ are iid conditional on $(\hat{\lambda}_i, Y_{i0})$. The second inequality holds because $|x+y|^4 \leq 8(|x|^4 + |y|^4)$.

The term $(N-1)^4 A_1$ takes the form

$$\begin{aligned}
\left(\sum a_j \right)^4 &= \left(\sum a_j^2 + 2 \sum_j \sum_{i>j} a_j a_i \right)^2 \\
&= \left(\sum a_j^2 \right)^2 + 4 \left(\sum a_j^2 \right) \left(\sum_j \sum_{i>j} a_j a_i \right) + 4 \left(\sum_j \sum_{i>j} a_j a_i \right)^2 \\
&= \sum a_j^4 + 6 \sum_j \sum_{i>j} a_j^2 a_i^2 \\
&\quad + 4 \left(\sum a_j^2 \right) \left(\sum_j \sum_{i>j} a_j a_i \right) + 4 \sum_j \sum_{i>j} \sum_{l \neq j} \sum_{k>l} a_j a_i a_l a_k,
\end{aligned}$$

where

$$a_j = f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) - \mathbb{E}_i[f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})], \quad j \neq i.$$

Notice that conditional on $(\hat{\lambda}_i(\rho), Y_{i0})$, the random variables a_j have mean zero and are *iid* across $j \neq i$. This implies that

$$\mathbb{E}_i\left[\left(\sum a_j\right)^4\right] = \sum \mathbb{E}_i\left[a_j^4\right] + 6\sum_j \sum_{i>j} \mathbb{E}_i\left[a_j^2 a_i^2\right].$$

The remaining terms drop out because they involve at least one term a_j that is raised to the power of one and therefore has mean zero.

Using the C_r inequality, Jensen's inequality, the conditional independence of a_j^2 and a_i^2 and Lemma A.6, we can bound

$$\mathbb{E}_i[a_j^4] \leq \frac{M}{B_N^6} p_i, \quad \mathbb{E}_i[a_j^2 a_i^2] \leq \frac{M}{B_N^4} p_i^2.$$

Thus, in the region $\mathcal{T}_{\hat{\rho}} \cap \mathcal{T}_{\bar{U}} \cap \mathcal{T}_{Y_0} \cap \mathcal{T}_{\hat{\lambda} Y_0} \cap \mathcal{T}_{p(\cdot)} \cap \mathcal{T}_{\tilde{p}(\cdot)}$,

$$\mathbb{E}_i[A_1] \leq \frac{M p_i}{N^3 B_N^6} + \frac{M p_i^2}{N^2 B_N^4} \leq M p_i^4,$$

The second inequality holds because over $\mathcal{T}_{p(\cdot)}$, $p_i \geq \frac{N^\epsilon}{N} \geq \frac{M}{N B_N^2}$ and for N large, $N^3 B_N^6$ and $N^2 B_N^4$ are larger than one under Assumption 3.3. Here M is uniform in $\pi \in \Pi$.

Using a similar argument, we can also deduce that

$$\mathbb{E}_i[A_2] \leq M p_i^4,$$

which proves Part (a) of the lemma.

Part (b). Similar to proof of Part (a).

Part (c). Can be established using existing results for the variance of a kernel density estimator.

Part (d). Similar to proof of Part (c).

Part (e). We have the desired result because by Lemma A.7 we can choose a constant c that does not depend on π such that

$$p_i - p_{*i} \leq c p_{*i}$$

over the region $\mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}$. Thus,

$$\left(\frac{p_i}{p_{*i}}\right)^m = \left(1 + \frac{p_i - p_{*i}}{p_{*i}}\right)^m \leq (1+c)^m.$$

We deduce that

$$\int_{\mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}} \left(\frac{p_i}{p_{*i}}\right)^m d\hat{\lambda}_i dy_{i0} \leq (1+c)^m \int_{\mathcal{T}_{\hat{\lambda}Y_0} \cap \mathcal{T}_{p(\cdot)}} d\hat{\lambda}_i dy_{i0} \leq (1+c)^m \big(2C'_N\big)^2 = o_{u.\pi}(N^+),$$

as required.

Part (f). Define

$$\psi_i(\hat{\lambda}_j, Y_{j0}) = \phi\left(\frac{\hat{\lambda}_j - \hat{\lambda}_i}{B_N}\right)\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right)$$

and write

$$\begin{aligned} \hat{p}_i^{(-i)} - p_{*i} &= \frac{1}{N-1}\sum_{j\neq i}^{N}\left\{\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j - \hat{\lambda}_i}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right)\right. \\ &\quad \left. - \mathbb{E}_i\left[\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j - \hat{\lambda}_i}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right)\right]\right\} \\ &= \frac{1}{B_N^2(N-1)}\sum_{j\neq i}^{N}\Big(\psi_i(\hat{\lambda}_j, Y_{j0}) - \mathbb{E}_i[\psi_i(\hat{\lambda}_j, Y_{j0})]\Big). \end{aligned}$$

Notice that conditional on $(\hat{\lambda}_i, Y_{i0})$, $\psi_i(\lambda_j, Y_{j0}) \sim iid$ across $j \neq i$ with $|\psi_i(\hat{\lambda}_j, Y_{j0})| \leq M$ for some finite constant M. Then, by Bernstein's inequality [16] (e.g., Lemma 19.32 in van der Vaart (1998)),

$$\begin{aligned} & N\mathbb{P}_i\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}) \\ &= N\mathbb{P}_i\left\{\frac{1}{B_N^2(N-1)}\sum_{j\neq i}^{N}\Big(\psi_i(\hat{\lambda}_j, Y_{j0}) - \mathbb{E}_i[\psi_i(\hat{\lambda}_j, Y_{j0})]\Big) < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}) \\ &\leq 2N\exp\left(-\frac{1}{4}\frac{B_N^4(N-1)p_{*i}^2/16}{\mathbb{E}_i[\psi_i(\hat{\lambda}_j, Y_{j0})^2] + MB_N^2 p_{i*}/4}\right)\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}). \end{aligned}$$

[16] For a bounded function f and a sequence of iid random variables X_i,

$$\mathbb{P}\left\{\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(f(X_i) - \mathbb{E}[f(X_i)])\right| > x\right\} \leq 2\exp\left(-\frac{1}{4}\frac{x^2}{\mathbb{E}[f(X_i)^2] + \frac{1}{\sqrt{N}}x\sup_x|f(x)|}\right).$$

Using an argument similar to the proof of Lemma A.6 one can show that

$$\mathbb{E}_i[\psi_i(\lambda_j, Y_{j0})^2/B_N^4] \leq M p_i/B_N^2.$$

In turn

$$N\mathbb{P}_i\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}) \leq 2\exp\left(-MNB_N^2\frac{p_{*i}^2}{p_i+p_{*i}} + \ln N\right)\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)}).$$

From Lemma A.7 we can find a constant c such that $p_i \leq (1+c)p_{*i}$ and $p_{*i} \leq (1+c)p_i$. This leads to

$$\frac{p_{*i}^2}{p_i+p_{*i}} \geq \frac{p_i}{(2+c)(1+c)^2}.$$

Then, on the region $\mathcal{T}_{p(\cdot)}$

$$\begin{aligned}
& N\mathbb{E}\left[\mathbb{P}_i\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] \\
\leq\ & 2\mathbb{E}\left[\exp\left(-MNB_N^2\frac{p_{*i}^2}{p_i+p_{*i}} + \ln N\right)\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\right] \\
\leq\ & 2\mathbb{E}\big[\exp\big(-MNB_N^2 p_i + \ln N\big)\,\mathbb{I}(\mathcal{T}_{\hat{\lambda}Y_0}\mathcal{T}_{p(\cdot)})\big] \\
\leq\ & 2\exp\big(-MB_N^2 N^{\epsilon} + \ln N\big) \\
=\ & o(1),
\end{aligned}$$

where the last line holds by Assumption 3.3 and the $o(1)$ bound in the last line is uniform in $\pi \in \Pi$. Then, we have the required result for Part (f). ■

A.2 Proofs for Section 3.3

Proof of Theorem 3.8.

Part (i): we verify that our assumptions hold uniformly for the multivariate normal distributions $\pi \in \Pi$.

Assumption 3.2. Because λ is normally distributed, the uncentered fourth moment is finite for each $\pi(\lambda) \in \Pi_\lambda$ and can be bounded uniformly. Note that

$$\mathbb{P}\big(|\lambda| \geq C\big) \leq \mathbb{P}\Big(|\lambda - \mu_\lambda| \geq C - |\mu_\lambda|\Big) \leq 2\exp\left(-\frac{C-|\mu_\lambda|}{2\sigma_\lambda^2}\right) \tag{A.32}$$

for $C > |\mu_\lambda| + 1$. By plugging the bounds from (23) into (A.32) one can obtain constants M_1, M_2, and M_3 such that Assumption 3.2(i) is satisfied. The second part can be verified by noting that $Y_0 \sim N(\mu_y, \sigma_y^2)$, where $\mu_y = \alpha_0 + \alpha_1\mu_\lambda$ and $\sigma_y^2 = \sigma_{y|\lambda}^2 + \alpha_1^2\sigma_\lambda^2$.

Assumption 3.4 The boundedness of the conditional density follows from $0 < \delta_{\sigma^2_{y|\lambda}} \leq \sigma^2_{y|\lambda}$ in (23). To verify Part (ii) define $\mu_y(\lambda) = \alpha_0 + \alpha_1\lambda$ and notice that $\tilde{y}|\lambda \sim N\big(\mu_y(\lambda), \sigma^2_{y|\lambda} + B_N^2\big)$. Thus, we can write

$$\sup_{\pi\in\Pi} \sup_{|y|\leq C_N', |\lambda|<C_N} \left| \frac{\int \frac{1}{B_N}\phi\left(\frac{\tilde{y}-y}{B_N}\right)\pi(\tilde{y}|\lambda)d\tilde{y}}{\pi(y|\lambda)} - 1 \right| = \sup_{\pi\in\Pi} \sup_{|y|\leq C_N', |\lambda|<C_N} \big|\mathcal{R}_{1,N}\cdot\mathcal{R}_{2,N} - 1\big|,$$

where

$$\mathcal{R}_{1,N} = \sqrt{\frac{\sigma^2_{y|\lambda}}{\sigma^2_{y|\lambda} + B_N^2}} \leq 1, \quad \mathcal{R}_{2,N} = \exp\left\{-\frac{1}{2}(y-\mu_y(\lambda))^2\left(\frac{1}{\sigma^2_{y|\lambda}+B_N^2} - \frac{1}{\sigma^2_{y|\lambda}}\right)\right\} \geq 1.$$

$\mathcal{R}_{1,N}$ can be bounded from below by replacing $\sigma^2_{y|\lambda}$ with $\delta_{\sigma^2_{y|\lambda}}$. Because $B_N \longrightarrow 0$ as $N \longrightarrow \infty$, $\mathcal{R}_{1,N} \longrightarrow 1$ uniformly. For the term $\mathcal{R}_{2,N}$ notice that

$$\begin{aligned} \big(y-\mu_y(\lambda)\big)^2\left(\frac{1}{\sigma^2_{y|\lambda}} - \frac{1}{\sigma^2_{y|\lambda}+B_N^2}\right) &= \big(y-\alpha_0-\alpha_1\lambda\big)^2 \frac{B_N^2}{\sigma^2_{y|\lambda}\big(\sigma^2_{y|\lambda}+B_N^2\big)} \\ &\leq 3\big((C_N')^2 + M_{\alpha_0}^2 + M_{\alpha_1}^2 C_N^2\big)\frac{B_N^2}{\big(\delta_{\sigma^2_{y|\lambda}}\big)^2} \longrightarrow 0 \end{aligned}$$

as $N \longrightarrow \infty$ because $B_N C_N = o(1)$ and $B_N C_N' = o(1)$ according to Assumption 3.3. Thus, $\mathcal{R}_{2,N} \longrightarrow 1$ uniformly which delivers the desired result.

Assumption 3.5. The first step is to derive the conditional prior distribution $\pi(\lambda|y)$ which is of the form $\lambda|y \sim N(\mu_{\lambda|y}, \sigma^2_{y|\lambda})$. The prior mean function is of the form $\mu_{\lambda|y} = \gamma_0 + \gamma_1 y$. If the prior for λ is a point mass, i.e., $\sigma_\lambda^2 = 0$, then the distribution of $(\lambda|y)$ is also a point mass with $\mu_{\lambda|y} = \mu_\lambda$ and $\sigma^2_{\lambda|y} = 0$. It can be verified that the coefficients γ_0, γ_1, and the variance $\sigma^2_{y|\lambda}$ are bounded from above in absolute value.

The prior is combined with the Gaussian likelihood function $\hat{\lambda}|\lambda \sim N\big(\lambda, \sigma^2/T\big)$, which leads to a posterior mean function that is linear in $\hat{\lambda}$ and y:

$$m(\hat{\lambda}, y; \pi) = \left(\frac{1}{\sigma^2_{y|\lambda}} + \frac{1}{\sigma^2/T}\right)^{-1}\left(\frac{1}{\sigma^2_{y|\lambda}}(\gamma_0 + \gamma_1 y_0) + \frac{1}{\sigma^2/T}\hat{\lambda}\right) = \bar{\gamma}_0 + \bar{\gamma}_1 y_0 + \bar{\gamma}_2\hat{\lambda}. \quad \text{(A.33)}$$

The $\bar{\gamma}$ coefficients are also bounded in absolute value for $\pi \in \Pi$.

The sampling distribution of $(\hat{\lambda}, y_0)$ is jointly normal with mean and covariance matrix

$$\mu_{\hat{\lambda},y} = \begin{bmatrix} \mu_\lambda \\ \alpha_0 + \alpha_1 \mu_\lambda \end{bmatrix}, \quad \Sigma_{\hat{\lambda},y} = \begin{bmatrix} \sigma_\lambda^2 + \sigma^2/T & \gamma_1\big(\sigma_{y|\lambda}^2 + \alpha_1^2\sigma_\lambda^2\big) \\ \gamma_1\big(\sigma_{y|\lambda}^2 + \alpha_1^2\sigma_\lambda^2\big) & \sigma_{y|\lambda}^2 + \alpha_1^2\sigma_\lambda^2 \end{bmatrix}. \tag{A.34}$$

It can be verified that the covariance matrix is always positive definite. The variances of $\hat{\lambda}$ and y_0 are strictly greater than some $\delta > 0$ and the two random variables are never perfectly correlated because $\hat{\lambda} = \lambda + (\sum_{t=1}^T u_t)/T$. Moreover, the covariance matrix can be bounded from above. By combining (A.33) and (A.34) we can deduce that the posterior mean has a Gaussian sampling distribution and we can use standard moment and tail bounds to establish the validity of the assumption. Calculations under the $p_*(\cdot)$ distributions are very similar.

Part (ii): we verify that our assumptions hold uniformly for the finite mixtures of multivariate normal distributions $\pi_{mix} \in \Pi_{mix}^{(K)}$.

Assumption 3.2. Consider the marginal density of λ given by $\pi_{mix}(\lambda) = \sum_{k=1}^K \omega_k \pi_k(\lambda)$. Thus, for any non-negative integrable function $h(\cdot)$ we can use the crude bound

$$\int h(\lambda)\pi_{mix}(\lambda)d\lambda \leq \sum_{k=1}^K \int h(\lambda)\pi_k(\lambda)d\lambda.$$

In turn, uniform tail probability and moment bounds for the mixture components translate into uniform bounds for $\pi_{mix}(\cdot)$.

Assumption 3.4 The key insight is that we can express

$$\pi_{mix}(y|\lambda) = \frac{\sum_{k=1}^K \omega_k \pi_k(\lambda, y)}{\sum_{k=1}^K \omega_k \pi_k(\lambda)} = \sum_{k=1}^K \left(\frac{\omega_k \pi_k(\lambda)}{\sum_{k=1}^K \omega_k \pi_k(\lambda)} \right) \frac{\pi_k(\lambda, y)}{\pi_k(\lambda)} \leq \sum_{k=1}^K \pi_k(y|\lambda).$$

This allows us to directly translate bounds for the mixture components $\pi_k(y|\lambda) \in \Pi_{y|\lambda}$ into results for $\pi_{mix}(y|\lambda)$. Using a similar argument, we can also deduce that

$$\begin{aligned}
&\sup_{\pi_{mix} \in \Pi_{mix}^{(K)}} \; \sup_{|y| \leq C_N', |\lambda| < C_N} \left| \frac{\int \frac{1}{B_N}\phi\left(\frac{\tilde{y}-y}{B_N}\right)\big[\pi_{mix}(\tilde{y}|\lambda) - \pi_{mix}(y|\lambda)\big]d\tilde{y}}{\pi_{mix}(y|\lambda)} \right| \\
\leq \; & \sum_{k=1}^K \sup_{\pi_k \in \Pi_k} \; \sup_{|y| \leq C_N', |\lambda| < C_N} \left| \frac{\int \frac{1}{B_N}\phi\left(\frac{\tilde{y}-y}{B_N}\right)\big[\pi_k(\tilde{y}|\lambda) - \pi_k(y|\lambda)\big]d\tilde{y}}{\pi_k(y|\lambda)} \right| \\
= \; & o(1).
\end{aligned}$$

Assumption 3.5. The prior distribution of λ given y is a mixture of normals with weights that are a function of y:

$$\pi_{mix}(\lambda|y) = \sum_{k=1}^{K}\left(\frac{\omega_k\pi_k(y)}{\sum_{k=1}^{K}\omega_k\pi_k(y)}\right)\frac{\pi_k(\lambda,y)}{\pi_k(y)} = \sum_{k=1}^{K}\underline{\omega}_k(y)\pi_k(\lambda|y). \tag{A.35}$$

Because $\hat{\lambda}|\lambda \sim N(\lambda,\sigma^2/T)$, the posterior mean function is given by

$$\begin{aligned} & m(\hat{\lambda},y;\pi_{mix}) \\ = & \sum_{k=1}^{K}\left(\frac{\underline{\omega}_k(y)\int\pi_k(\lambda|y)\phi_N(\hat{\lambda};\lambda,\sigma^2/T)d\lambda}{\sum_{k=1}^{K}\underline{\omega}_k(y)\int\pi_k(\lambda|y)\phi_N(\hat{\lambda};\lambda,\sigma^2/T)d\lambda}\right)\frac{\int\lambda\pi_k(\lambda|y)\phi_N(\hat{\lambda};\lambda,\sigma^2/T)d\lambda}{\int\pi_k(\lambda|y)\phi_N(\hat{\lambda};\lambda,\sigma^2/T)d\lambda} \\ = & \sum_{k=1}^{K}\bar{\omega}_k(\hat{\lambda},y)m(\hat{\lambda},y;\pi_k). \end{aligned} \tag{A.36}$$

Thus, the posterior mean is a weighted average of the posterior means derived from the K mixture components. The $\bar{\omega}(\hat{\lambda},y)$ can be interpreted as posterior probabilities of the mixture components. We can bound the posterior mean as follows:

$$\left|m(\hat{\lambda},y;\pi_{mix})\right| \le \sum_{k=1}^{K}\left|m(\hat{\lambda},y;\pi_k)\right| = \sum_{k=1}^{K}\left|\bar{\gamma}_{0,k}+\bar{\gamma}_{1,k}y+\bar{\gamma}_{2,k}\hat{\lambda}\right| \le M_0+M_y|y|+M_{\hat{\lambda}}|\hat{\lambda}|, \tag{A.37}$$

where the $\bar{\gamma}$ coefficients were defined in (A.33) and are bounded for $\pi_k \in \Pi$. Thus, that the overall bound for $\left|m(\hat{\lambda},y;\pi_{mix})\right|$ is piecewise linear in y and $\hat{\lambda}$. The joint sampling distribution of $(\hat{\lambda},y)$ is given by the following mixture of normals:

$$p(\hat{\lambda},y;\pi_{mix}) = \int p(\hat{\lambda}|\lambda)\sum_{k=1}^{K}\omega_k\pi_k(\lambda,y)d\lambda = \sum_{k=1}^{K}\omega_k p(\hat{\lambda},y;\pi_k). \tag{A.38}$$

Based on (A.37) and (A.38) one can establish the uniform tailbounds in the assumption. Calculations under the $p_*(\cdot)$ distributions are very similar. ■

B Monte Carlo Experiments

B.1 Data Generating Processes

Monte Carlo Design 2:

- Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0, \gamma^2)$; $\rho = 0.8$, $\gamma = 1$.
- Initial Observation: $Y_{i0} \sim N\left(\frac{\underline{\mu}_\lambda}{1-\rho}, V_Y + \frac{\underline{V}_\lambda}{(1-\rho)^2}\right)$, $V_Y = \gamma^2/(1-\rho^2)$; $\underline{\mu}_\lambda = 1$, $\underline{V}_\lambda = 1$.
- Correlated Random Effects, $\delta \in \{0.05, 0.1, 0.3\}$:

$$\begin{aligned}
\lambda_i|Y_{i0} &\sim \begin{cases} N\big(\phi_+(Y_{i0}), \underline{\Omega}\big) & \text{with probability } p_\lambda \\ N\big(\phi_-(Y_{i0}), \underline{\Omega}\big) & \text{with probability } 1-p_\lambda \end{cases}, \quad \begin{array}{l} \phi_+(Y_{i0}) = \phi_0 + \delta + (\phi_1 + \delta)Y_{i0} \\ \phi_-(Y_{i0}) = \phi_0 - \delta + (\phi_1 - \delta)Y_{i0} \end{array} \\
\underline{\Omega} &= \left[\frac{1}{(1-\rho)^2} V_Y^{-1} + \underline{V}_\lambda^{-1}\right]^{-1}, \quad \phi_0 = \underline{\Omega}\,\underline{V}_\lambda^{-1}\underline{\mu}_\lambda, \quad \phi_1 = \frac{1}{1-\rho}\underline{\Omega} V_Y^{-1}, \quad p_\lambda = 1/2
\end{aligned}$$

- Sample Size: $N = 1,000$, $T = 4$.
- Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

Monte Carlo Design 3:

- Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$, $\rho = 0.8$, $\mathbb{E}[U_{it}] = 0$, $\mathbb{V}[U_{it}] = 1$
- Scale Mixture:

$$U_{it} \sim iid \begin{cases} N(0, \gamma_+^2) & \text{with probability } p_u \\ N(0, \gamma_-^2) & \text{with probability } 1-p_u \end{cases},$$

  where $\gamma_+^2 = 4$, $\gamma_-^2 = 1/4$, and $p_u = (1-\gamma_-^2)/(\gamma_+^2 - \gamma_-^2) = 1/5$.

- Location Mixture:

$$U_{it} \sim iid \begin{cases} N(\mu_+, \gamma^2) & \text{with probability } p_u \\ N(\mu_-, \gamma^2) & \text{with probability } 1-p_u \end{cases},$$

  where $\mu_- = -1/4$, $\mu_+ = 2$, $p_u = |\mu_-|\big/(|\mu_-| + \mu_u^+) = 1/9$, and $\gamma^2 = 1 - p_u\mu_+^2 - (1 - p_u)\mu_-^2 = 1/2$.

- Correlated Random Effects, Initial Observations: same as Design 2 with $\delta = 0.1$.
- Sample Size: $N = 1,000$, $T = 4$.
- Number of Monte Carlo Repetitions: $N_{sim} = 1,000$.

B.2 Consistency of QMLE in Monte Carlo Designs 2 and 3

We show for the basic dynamic panel data model that even if the Gaussian correlated random effects distribution is misspecified, the pseudo-true value of the QMLE estimator of θ corresponds to the "true" θ_0. We do so, by calculating

$$(\theta_*, \xi_*) = \text{argmax}_{\theta,\xi}\, \mathbb{E}^{\mathcal{Y}}_{\theta_0}\left[\ln p(Y, X_2|H, \theta, \xi)\right], \tag{A.39}$$

and verifying that $\theta_* = \theta_0$. Because the observations are conditionally independent across i and the likelihood function is symmetric with respect to i, we can drop the i subscripts.

We make some adjustment to the notation. The covariance matrix Σ only depends on γ, but not on (ρ, α). Moreover, we will split ξ into the parameters that characterize the conditional mean of λ, denoted by Φ, and ω, which are the non-redundant elements of the prior covariance matrix $\underline{\Omega}$. Finally, we define

$$\tilde{Y}(\theta_1) = Y - X\rho - Z\alpha$$

with the understanding that $\theta_1 = (\rho, \alpha)$ and excludes γ. Moreover, let $\phi = \text{vec}(\Phi')$ and $\tilde{h}' = I \otimes h'$, such that we can write $\Phi h = \tilde{h}'\phi$. Using this notation, we can write

$$\begin{aligned}
&\ln p(y, x_2|h, \theta_1, \gamma, \phi, \omega) \\
&= C - \frac{1}{2}\ln|\Sigma(\gamma)| - \frac{1}{2}\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\gamma)\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big) \\
&\quad -\frac{1}{2}\ln\left|\underline{\Omega}\right| + \frac{1}{2}\ln\left|\bar{\Omega}(\gamma,\omega)\right| \\
&\quad -\frac{1}{2}\Big(\hat{\lambda}(\theta)'w'\Sigma^{-1}(\gamma)w\hat{\lambda}(\theta) + \phi'\tilde{h}\underline{\Omega}^{-1}\tilde{h}'\phi - \bar{\lambda}'(\theta,\xi)\bar{\Omega}^{-1}(\gamma,\omega)\bar{\lambda}(\theta,\xi)\Big),
\end{aligned} \tag{A.40}$$

where

$$\begin{aligned}
\hat{\lambda}(\theta) &= (w'\Sigma^{-1}(\gamma)w)^{-1}w'\Sigma^{-1}(\gamma)\tilde{y}(\theta_1) \\
\bar{\Omega}^{-1}(\gamma,\omega) &= \underline{\Omega}^{-1} + w'\Sigma^{-1}(\gamma)w, \quad \bar{\lambda}(\theta,\xi) = \bar{\Omega}(\gamma,\omega)\big(\underline{\Omega}^{-1}\tilde{h}'\phi + w'\Sigma^{-1}(\gamma)w\hat{\lambda}(\theta)\big).
\end{aligned}$$

In the basic dynamic panel data model λ is scalar, $w = \iota$, $\Sigma(\gamma) = \gamma^2 I$, $x_2 = \emptyset$, $z = \emptyset$, $h = [1, y_0]'$, $\underline{\Omega} = \omega^2$. Thus, splitting the $(T-1)(\ln\gamma^2)/2$, we can write

$$\begin{aligned}
\ln p(y|h,\rho,\gamma,\phi,\omega) &= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\big(\tilde{y}(\rho) - \iota\hat{\lambda}(\rho)\big)'\big(\tilde{y}(\rho) - \iota\hat{\lambda}(\rho)\big) \\
&\quad -\frac{1}{2}\ln\left|\omega^2\right| - \frac{1}{2}\ln\left|\gamma^2/T\right| + \frac{1}{2}\ln(1/T) + \frac{1}{2}\ln\left|\bar{\Omega}(\gamma,\omega)\right| \\
&\quad -\frac{1}{2}\left(\frac{T}{\gamma^2}\hat{\lambda}^2(\rho) + \frac{1}{\omega^2}\phi'\tilde{h}\tilde{h}'\phi - \frac{1}{\bar{\Omega}(\gamma,\omega)}\bar{\lambda}^2(\theta,\xi)\right),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\lambda}(\rho) &= \frac{1}{T}\iota'\tilde{y}(\rho) \\
\bar{\Omega}^{-1}(\gamma,\omega) &= \frac{1}{\omega^2} + \frac{1}{\gamma^2/T}, \quad \bar{\lambda}(\theta,\xi) = \bar{\Omega}(\gamma,\omega)\left(\frac{1}{\omega^2}\tilde{h}'\phi + \frac{T}{\gamma^2}\hat{\lambda}(\rho)\right).
\end{aligned}$$

Note that

$$-\frac{1}{2}\ln\left|\omega^2\right| + \frac{1}{2}\ln\left|T/\gamma^2\right| + \frac{1}{2}\ln\left|\bar{\Omega}(\gamma,\omega)\right| = \frac{1}{2}\ln\left|\frac{\frac{1}{\omega^2}\frac{T}{\gamma^2}}{\frac{1}{\omega^2} + \frac{T}{\gamma^2}}\right| = -\frac{1}{2}\ln\left|\omega^2 + \gamma^2/T\right|.$$

In turn, we can write

$$\begin{aligned}
&\ln p(y|h,\rho,\gamma,\phi,\omega) \\
&= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\tilde{y}(\rho)'(I - \iota\iota'/T)\tilde{y}(\rho) - \frac{1}{2}\ln\left|\omega^2 + \gamma^2/T\right| \\
&\quad -\frac{1}{2}\left(\frac{T}{\gamma^2}\hat{\lambda}^2(\rho) + \frac{1}{\omega^2}\phi'\tilde{h}\tilde{h}'\phi - \frac{\omega^2\gamma^2/T}{\omega^2 + \gamma^2/T}\left(\frac{1}{\omega^2}\tilde{h}'\phi + \frac{T}{\gamma^2}\hat{\lambda}(\rho)\right)^2\right) \\
&= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\tilde{y}(\rho)'(I - \iota\iota'/T)\tilde{y}(\rho) - \frac{1}{2}\ln\left|\omega^2 + \gamma^2/T\right| \\
&\quad -\frac{1}{2(\omega^2 + \gamma^2/T)}\left(\phi'\tilde{h}\tilde{h}'\phi - 2\hat{\lambda}(\rho)\tilde{h}'\phi + \hat{\lambda}^2(\rho)\right).
\end{aligned}$$

Taking expectations (we omit the subscripts from the expectation operator), we can write

$$
\begin{aligned}
&\mathbb{E}\big[\ln p(Y|H,\rho,\gamma,\phi,\omega)\big] \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(A.41)}\\
&= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big] - \frac{1}{2}\ln\big|\omega^2+\gamma^2/T\big| \\
&\quad -\frac{1}{2(\omega^2+\gamma^2/T)}\bigg(\big(\phi-\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big)'\mathbb{E}[\tilde{H}\tilde{H}']\big(\phi-\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big) \\
&\quad -\mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]+\mathbb{E}[\hat{\lambda}^2(\rho)]\bigg).
\end{aligned}
$$

We deduce that

$$
\phi_*(\rho) = \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]. \tag{A.42}
$$

To evaluate $\phi_*(\rho_0)$, note that $\hat{\lambda}(\rho_0)=\lambda+\iota'u/T$. Using that fact that the initial observation Y_{i0} is uncorrelated with the shocks U_{it}, $t\geq 1$, we deduce that $\mathbb{E}[\tilde{H}\hat{\lambda}(\rho_0)]=\mathbb{E}[\tilde{H}\lambda]$. Thus,

$$
\phi_*(\rho_0) = \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.43}
$$

The pseudo-true value is obtained through a population regression of λ on H.

Plugging the pseudo-true value for ϕ into (A.41) yields the concentrated objective function

$$
\begin{aligned}
&\mathbb{E}\big[\ln p(Y|H,\rho,\gamma,\phi_*(\rho),\omega)\big] \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(A.44)}\\
&= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big] \\
&\quad -\frac{1}{2}\ln\big|\omega^2+\gamma^2/T\big| - \frac{1}{2(\omega^2+\gamma^2/T)}\big(\mathbb{E}[\hat{\lambda}^2(\rho)]-\mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big).
\end{aligned}
$$

Using well-known results for the maximum likelihood estimator of a variance parameter in a Gaussian regression model, we can immediately deduce that

$$
\begin{aligned}
\gamma_*^2(\rho) &= \frac{1}{T-1}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big] \qquad\qquad \text{(A.45)}\\
\omega_*^2(\rho)+\gamma_*^2(\rho)/T &= \big(\mathbb{E}[\hat{\lambda}^2(\rho)]-\mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big).
\end{aligned}
$$

At $\rho=\rho_0$ we obtain $\tilde{Y}(\rho_0)=\iota\lambda+u$. Thus, $\mathbb{E}[\hat{\lambda}^2(\rho_0)]=\gamma_0^2/T+\mathbb{E}[\lambda^2]$ and $\mathbb{E}[\tilde{H}\hat{\lambda}(\rho_0)]=\mathbb{E}[\tilde{H}\lambda]$. In turn,

$$
\gamma_*^2(\rho_0)=\gamma_0^2, \quad \omega_*^2(\rho_0)=\mathbb{E}[\lambda^2]-\mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.46}
$$

Given $\rho = \rho_0$ the pseudo-true value for γ^2 is the "true" γ_0^2 and the pseudo-true variance of the correlated random-effects distribution is given by the expected value of the squared residual from a projection of λ onto H.

Using (A.45), we can now concentrate out γ^2 and ω^2 from the objective function (A.44):

$$
\begin{aligned}
&\mathbb{E}\big[\ln p(Y|H,\rho,\gamma_*(\rho),\phi_*(\rho),\omega_*(\rho)\big] \\
&= C - \frac{T-1}{2}\ln\big|\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big]\big| \\
&\quad -\frac{1}{2}\ln\big|\mathbb{E}[\tilde{Y}'(\rho)\iota\iota'\tilde{Y}(\rho)] - \mathbb{E}[\tilde{Y}'(\rho)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho)]\big|.
\end{aligned}
\tag{A.47}
$$

To find the maximum of $\mathbb{E}\big[\ln p(Y|H,\rho,\gamma_*(\rho),\phi_*(\rho),\omega_*(\rho)\big]$ with respect to ρ we will calculate the first-order condition. Differentiating (A.47) with respect to ρ yields

$$
\begin{aligned}
\text{F.O.C.}(\rho) &= (T-1)\frac{\mathbb{E}\big[X'(I-\iota\iota'/T)\tilde{Y}(\rho)\big]}{\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big]} \\
&\quad + \frac{\mathbb{E}[X'\iota\iota'\tilde{Y}(\rho)] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho)]}{\mathbb{E}[\tilde{Y}'(\rho)\iota\iota'\tilde{Y}(\rho)] - \mathbb{E}[\tilde{Y}'(\rho)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho)]}.
\end{aligned}
$$

We will now verify that F.O.C.$(\rho_0) = 0$. Because both denominators are strictly positive, we can rewrite the condition as

$$
\begin{aligned}
\text{F.O.C.}(\rho_0) &= (T-1)\mathbb{E}\big[X'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \\
&\quad \times\Big(\mathbb{E}[\tilde{Y}'(\rho_0)\iota\iota'\tilde{Y}(\rho_0)] - \mathbb{E}[\tilde{Y}'(\rho_0)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\Big) \\
&\quad +\mathbb{E}\big[\tilde{Y}(\rho_0)'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \\
&\quad \times\Big(\mathbb{E}[X'\iota\iota'\tilde{Y}(\rho_0)] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\Big).
\end{aligned}
\tag{A.48}
$$

Using again the fact that $\tilde{Y}(\rho_0) = \iota\lambda + U$, we can rewrite the terms appearing in the first-order condition as follows:

$$
\begin{aligned}
\mathbb{E}\big[X'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] &= \mathbb{E}\big[X'(I-\iota\iota'/T)u\big] = \mathbb{E}[X'u] - \mathbb{E}[X'\iota\iota'u]/T = -\mathbb{E}[X'\iota\iota'u]/T \\
\mathbb{E}[\tilde{Y}'(\rho_0)\iota\iota'\tilde{Y}(\rho)] &= \mathbb{E}\big[(\lambda\iota' + u')\iota\iota'(\iota\lambda + u)\big] = T^2\mathbb{E}[\lambda^2] + \mathbb{E}[u'\iota\iota'u] = T^2\mathbb{E}[\lambda^2] + T\gamma_0^2 \\
\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)] &= \mathbb{E}[\tilde{H}\iota'(\iota\lambda + u)] = T\mathbb{E}[\tilde{H}\lambda] \\
\mathbb{E}\big[\tilde{Y}(\rho_0)'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] &= \mathbb{E}\big[u'(I-\iota\iota'/T)u\big] = (T-1)\gamma^2 \\
\mathbb{E}[X'\iota\iota'\tilde{Y}(\rho_0)] &= \mathbb{E}[X'\iota\iota'(\iota\lambda + u)] = T\mathbb{E}[X'\iota\lambda] + \mathbb{E}[X'\iota\iota'u].
\end{aligned}
$$

For the first equality we used the fact that $X_{it} = Y_{it-1}$ is uncorrelated with U_{it}. We can now re-state the first-order condition (A.48) as follows:

$$\begin{aligned} & \text{F.O.C.}(\rho_0) && \text{(A.49)} \\ = \quad & -(T-1)\big(\mathbb{E}[X'\iota\iota' u]\big)\Big(\gamma_0^2 + T\big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big) \\ & + \Big(\mathbb{E}[X'\iota\iota' u] + T\big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big)(T-1)\gamma_0^2 \\ = \quad & T(T-1)\Big[\gamma_0^2\Big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big) \\ & - \mathbb{E}[X'\iota\iota' u]\Big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big)\Big]. \end{aligned}$$

We now have to analyze the terms involving $X'\iota$. Note that we can express

$$Y_t = \rho_0^t Y_0 + \sum_{\tau=0}^{t-1} \rho_0^\tau (\lambda + U_{t-\tau}).$$

Define $a_t = \sum_{\tau=0}^{t-1} \rho_0^\tau$ and $b = \sum_{t=1}^{T-1} a_t$. Thus, we can write

$$Y_t = \rho_0^t Y_0 + \lambda a_t + \sum_{\tau=0}^{t-1} \rho_0^\tau U_{t-\tau}, \quad t > 0.$$

Consequently,

$$X'\iota = \sum_{t=0}^{T-1} Y_t = Y_0 \left(\sum_{t=0}^{T-1} \rho_0^t\right) + \lambda \left(\sum_{t=1}^{T-1} a_t\right) + \sum_{t=1}^{T-1}\sum_{\tau=0}^{t-1} \rho_0^\tau U_{t-\tau} = a_T y_0 + b\lambda + \sum_{t=1}^{T-1} a_t U_{T-t}.$$

Thus, we obtain

$$\begin{aligned} \mathbb{E}[X'\iota\iota' u] &= \mathbb{E}\left[\left(a_T Y_0 + b\lambda + \sum_{t=1}^{T-1} a_t U_{T-t}\right)\left(\sum_{t=1}^{T} U_t\right)\right] = b\gamma_0^2 \\ \mathbb{E}[X'\iota\lambda] &= \mathbb{E}\left[\left(a_T Y_0 + b\lambda + \sum_{t=1}^{T-1} a_t U_{T-t}\right)\lambda\right] = a_T\mathbb{E}[Y_0\lambda] + b\mathbb{E}[\lambda^2] \\ \mathbb{E}[X'\iota\tilde{H}'] &= \mathbb{E}\left[\left(a_T Y_0 + b\lambda + \sum_{t=1}^{T-1} a_t U_{T-t}\right)\tilde{H}'\right] = a_T\mathbb{E}[Y_0\tilde{H}'] + b\mathbb{E}[\lambda\tilde{H}']. \end{aligned}$$

Using these expressions, most terms that appear in (A.49) cancel out and the condition simplifies to

$$\text{F.O.C.}(\rho_0) = T(T-1)\gamma_0 a_T \left(\mathbb{E}[Y_0\lambda] - \mathbb{E}[Y_0\tilde{H}'] \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda] \right). \tag{A.50}$$

Now consider

$$\begin{aligned} & \mathbb{E}[Y_0\tilde{H}'] \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda] \\ = \; & \frac{1}{\mathbb{E}[Y_0^2] - (\mathbb{E}[Y_0])} \begin{bmatrix} \mathbb{E}[Y_0] & \mathbb{E}[Y_0^2] \end{bmatrix} \begin{bmatrix} \mathbb{E}[Y_0^2] & -\mathbb{E}[Y_0] \\ -\mathbb{E}[Y_0] & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}[Y_0] \\ \mathbb{E}[Y_0^2] \end{bmatrix} \\ = \; & \mathbb{E}[Y_0\lambda]. \end{aligned}$$

Thus, we obtain the desired result that $\text{F.O.C.}(\rho_0) = 0$. To summarize, the pseudo-true values are given by

$$\begin{aligned} \rho_* &= \rho_0, \quad \gamma_*^2 = \gamma_0, \quad \phi_* = \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda], \\ \omega_*^2 &= \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}'] \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \quad \blacksquare \end{aligned} \tag{A.51}$$

B.3 Computation of the Oracle Predictor in Design 3

We are using a Gibbs sampler to compute the oracle predictor under the mixture distributions for both λ_i and U_{it}.

Here we combine the scale mixture and the location mixture in a unified framework. Let $a_{it} = 1$ if U_{it} is generated from the mixture component with mean μ_+ and variance γ_+^2, and $a_{it} = 0$ if U_{it} is generated from the mixture component with mean μ_- variance γ_-^2. Then, $\mu_+ = \mu_- = 0$ for the scale mixture, and $\gamma_+^2 = \gamma_-^2 = \gamma^2$ for the location mixture. Also, let b_i be an indicator of the components in the correlated random effects distribution, such that

$$\phi(Y_{i0}, b_i) = \begin{cases} \phi_+(Y_{i0}), & \text{if } b_i = 1, \\ \phi_-(Y_{i0}), & \text{if } b_i = 0. \end{cases}$$

Omitting i subscripts from now on, define

$$\tilde{Y}_t = Y_t - \rho Y_{t-1} - (a_t\mu_+ + (1-a_t)\mu_-), \quad \gamma^2(a_t) = a_t\gamma_+^2 + (1-a_t)\gamma_-^2,$$

so we have

$$\tilde{Y}_t(a_t)|(\lambda, a_t) \sim N\big(\lambda, \gamma^2(a_t)\big).$$

Conditional on b, the prior distribution is

$$\lambda|(Y_0, b) \sim N(\phi(Y_0, b), \underline{\Omega}),$$

and we obtain a posterior distribution of the form

$$\lambda|(Y_{0:T}, a_{1:T}, b) \sim N\big(\bar{\lambda}(a_{1:T}, b), \bar{\Omega}(a_{1:T})\big), \tag{A.52}$$

where

$$\begin{aligned}
\bar{\Omega}(a_{1:T}) &= \Big(\underline{\Omega}^{-1} + \sum_{t=1}^{T} (\gamma^2(a_t))^{-1}\Big)^{-1} \\
\bar{\lambda}(a_{1:T}, b) &= \bar{\Omega}(a_{1:T})\Big(\underline{\Omega}^{-1}\phi(Y_0, b) + \sum_{t=1}^{T} (\gamma^2(a_t))^{-1}\tilde{Y}_t(a_t)\Big).
\end{aligned}$$

The posterior probability of $a_t = 1$ conditional on $(\lambda, Y_{0:T})$ is given by

$$\begin{aligned}
&\mathbb{P}\big(a_t = 1|\lambda, Y_{0:T}\big) \qquad\qquad (\text{A.53}) \\
&= \frac{p_u(\gamma_+)^{-1} \exp\left\{-\frac{1}{2\gamma_+^2}(\tilde{Y}_t(1) - \lambda)^2\right\}}{p_u(\gamma_+)^{-1} \exp\left\{-\frac{1}{2\gamma_+^2}(\tilde{Y}_t(1) - \lambda)^2\right\} + (1 - p_u)(\gamma_-)^{-1} \exp\left\{-\frac{1}{2\gamma_-^2}(\tilde{Y}_t(0) - \lambda)^2\right\}},
\end{aligned}$$

And the posterior probability of $b = 1$ conditional on $(\lambda, Y_{0:T}, a_{1:T})$ is given by

$$\begin{aligned}
&\mathbb{P}\big(b = 1|\lambda, Y_{0:T}, a_{1:T}\big) \qquad\qquad (\text{A.54}) \\
&= \frac{p_\lambda \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{(\tilde{Y}_t(a_t) - \phi_+(Y_0))^2}{\underline{\Omega} + \gamma^2(a_t)}\right\}}{p_\lambda \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{(\tilde{Y}_t(a_t) - \phi_+(Y_0))^2}{\underline{\Omega} + \gamma^2(a_t)}\right\} + (1 - p_\lambda) \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{(\tilde{Y}_t(a_t) - \phi_-(Y_0))^2}{\underline{\Omega} + \gamma^2(a_t)}\right\}}.
\end{aligned}$$

The posterior mean $\mathbb{E}[\lambda|Y_{0:T}]$ can be approximated with the following Gibbs sampler. Generate a sequence of draws $\{\lambda^s, a^s_{1:T}, b^s\}_{s=1}^{N_{sim}}$ by iterating over the conditional distributions given in (A.52), (A.53), and (A.54). Denote $\bar{p}(a_{1:T}, b) = \mathbb{P}\big(b|\lambda, Y_{0:T}, a_{1:T}\big)$, then,

$$\begin{aligned}
\widehat{\mathbb{E}}[\lambda|Y_{0:T}] &= \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \sum_{b=0}^{1} \bar{p}(a^s_{1:T}, b)\bar{\lambda}(a^s_{1:T}, b), \qquad \text{(A.55)} \\
\widehat{\mathbb{V}}[\lambda|Y_{0:T}] &= \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \left(\bar{\Omega}(a^s_{1:T}) + \sum_{b=0}^{1} \bar{p}(a^s_{1:T}, b)\bar{\lambda}^2(a^s_{1:T}, b) \right) - \left(\widehat{\mathbb{E}}[\lambda|Y_{0:T}]\right)^2.
\end{aligned}$$

C Data Set

The construction of our data is based on Covas, Rump, and Zakrajsek (2014). We downloaded FR Y-9C BHC financial statements for the quarters 2002Q1 to 2014Q4 using the web portal of the Federal Reserve Bank of Chicago. We define PPNR (relative to assets) as follows

$$\text{PPNR} = 400\big(\text{NII} + \text{ONII} - \text{ONIE}\big)/\text{ASSETS},$$

where

NII	= Net Interest Income	BHCK 4074
ONII	= Total Non-Interest Income	BHCK 4079
ONIE	= Total Non-Interest Expenses	BHCK 4093 - C216 - C232
ASSETS	= Consolidated Assets	BHCK 3368

Here net interest income is the difference between total interest income and expenses. It excludes provisions for loan and lease losses. Non-interest income includes various types of fees, trading revenue, as well as net gains on asset sales. Non-interest expenses include, for instance, salaries and employee benefits and expenses of premises and fixed assets. As in Covas, Rump, and Zakrajsek (2014), we exclude impairment losses (C216 and C232). We divide the net revenues by the amount of consolidated assets. This ratio is multiplied by 400 to annualize the flow variables and convert the ratio into percentages.

The raw data take the form of an unbalanced panel of BHCs. The appearance and disappearance of specific institutions in the data set is affected by entry and exit, mergers and acquisitions, as well as changes in reporting requirements for the FR Y-9C form. Note that NII, ONII, and ONIE are reported as year-to-date values. Thus, in order to obtain quarterly data, we take differences: $Q1 \mapsto Q1$, $(Q2 - Q1) \mapsto Q2$, $(Q3 - Q2) \mapsto Q3$, and $(Q4 - Q3) \mapsto Q4$. ASSETS is a stock variable and no further transformation is needed.

Our goal is to construct rolling samples that consist of $T + 2$ observations, where T is the size of the estimation sample and varies between $T = 3$ and $T = 11$. The additional two observations in each rolling sample are used, respectively, to initialize the lag in the first period of the estimation sample and to compute the error of the one-step-ahead forecast. We index each rolling sample by the forecast origin $t = \tau$. For instance, taking the time period t to be a quarter, with data from 2002Q1 to 2014Q4 we can construct $M = 45$ samples of size $T = 6$ with forecast origins running from $\tau = 2003Q3$ to $\tau = 2014Q3$. Each rolling sample is indexed by the pair (τ, T). The following adjustment procedure that eliminates BHCs with missing observations and outliers is applied to each rolling sample (τ, T) separately:

1. Eliminate BCHs for which total assets are missing for all time periods in the sample.

2. Compute average non-missing total assets and eliminate BCHs with average assets below 500 million dollars.

3. Eliminate BCHs for which one or more PPNR components are missing for at least one period of the sample.

4. Eliminate BCHs for which the absolute difference between the temporal mean and the temporal median exceeds 10.

5. Define deviations from temporal means as $\delta_{it} = y_{it} - \bar{y}_i$. Pooling the δ_{it}'s across institutions and time periods, compute the median $q_{0.5}$ and the 0.025 and 0.975 quantiles, $q_{0.025}$ and $q_{0.975}$. We delete institutions for which at least one δ_{it} falls outside of the range $q_{0.5} \pm (q_{0.975} - q_{0.025})$.

The effect of the sample-adjustment procedure on the size of the rolling samples is summarized in Table A-1. Here we are focusing on samples with $T = 6$ as in the main text. The column labeled N_0 provides the number of raw data for each sample. In columns N_j, $j = 1, \dots, 4$, we report the observations remaining after adjustment j. Finally, N is the number of observations after the fifth adjustment. This is the relevant sample size for the subsequent empirical analysis. For many BCHs we do not have information on the consolidated assets, which leads to reduction of the sample size by 60% to 80%. Once we restrict average consolidated assets to be above 500 million dollars, the sample size shrinks to approximately 700 to 1,200 institutions. Roughly 10% to 25% of these institutions have missing observations for PPNR components, which leads to N_3. The outlier elimination in Steps 4. and 5. have a relatively small effect on the sample size.

Descriptive statistics for the $T = 6$ rolling samples are reported in Table A-2. For each rolling sample we pool observations across institutions and time periods. We do not weight the observations by the size of the institution. Notice that the mean PPNR falls from about 2% for the 2003 samples to 1.24% for the 2010Q2 sample, which includes observations starting in 2008Q4. Then, the mean slightly increases and levels off at around 1.3%. The means are close to the medians, suggesting that the samples are not very skewed, which is confirmed by the skewness measures reported in the second to last column. The samples also exhibit fat tails. The kurtosis statistics range from 4 to 190.

Table A-1: Size of Adjusted Rolling Samples ($T = 6$)

Sample	Adjustment Step					
τ	N_0	N_1	N_2	N_3	N_4	N
2003Q3	6176	2258	710	653	653	614
2003Q4	6177	2289	730	658	658	618
2004Q1	6142	2351	744	660	660	622
2004Q2	6089	2375	754	657	657	613
2004Q3	6093	2416	778	669	668	624
2004Q4	6090	2448	787	668	667	621
2005Q1	6101	2486	797	680	679	629
2005Q2	6077	2489	809	695	694	644
2005Q3	6083	2473	826	718	717	660
2005Q4	6050	2451	828	728	727	658
2006Q1	6054	2425	834	715	715	664
2006Q2	6024	2403	849	734	734	685
2006Q3	6053	2376	858	747	747	697
2006Q4	6038	2367	880	757	757	711
2007Q1	6075	2355	905	772	772	727
2007Q2	6044	2337	929	777	777	732
2007Q3	6054	1101	941	773	773	712
2007Q4	6038	1061	919	769	769	710
2008Q1	6014	1081	945	770	770	713
2008Q2	5997	1070	942	775	775	722
2008Q3	5953	1062	949	784	784	731
2008Q4	5947	1058	949	792	792	741
2009Q1	5904	1113	1006	795	795	744
2009Q2	5878	1104	996	795	795	745
2009Q3	5805	1087	986	799	799	749
2009Q4	5793	1081	977	809	808	754
2010Q1	5709	1124	1015	800	799	744
2010Q2	5700	1116	1005	800	799	738
2010Q3	5665	1105	997	795	794	727
2010Q4	5652	1105	996	844	843	780
2011Q1	5586	1131	1027	838	837	773
2011Q2	5566	1129	1027	836	836	777
2011Q3	5483	1119	1018	833	833	770
2011Q4	5636	1115	1011	864	864	797
2012Q1	5876	1259	1154	863	863	794
2012Q2	5847	1240	1140	858	858	792
2012Q3	5809	1226	1135	849	849	789
2012Q4	5793	1216	1124	878	878	811
2013Q1	5749	1246	1157	875	875	808
2013Q2	5739	1245	1153	874	874	806
2013Q3	5699	1230	1142	874	874	805
2013Q4	5695	1233	1143	997	995	920
2014Q1	5603	1253	1162	979	977	899
2014Q2	5572	1237	1143	973	972	897
2014Q3	5514	1231	1140	966	965	898

Table A-2: Descriptive Statistics for Rolling Samples ($T = 6$)

Sample			Statistics				
τ	Min	Mean	Median	Max	StdD	Skew	Kurt
2003Q3	2.04	-2.10	2.00	12.01	0.90	3.00	29.46
2003Q4	2.02	-1.43	1.98	11.18	0.87	2.75	25.03
2004Q1	1.99	-2.10	1.95	11.18	0.91	3.13	29.90
2004Q2	1.96	-0.98	1.92	11.18	0.83	2.76	27.24
2004Q3	1.92	-0.98	1.89	10.80	0.76	2.06	22.28
2004Q4	1.90	-0.83	1.88	6.06	0.69	0.52	4.85
2005Q1	1.89	-0.73	1.87	6.01	0.70	0.62	4.94
2005Q2	1.90	-0.73	1.87	5.76	0.70	0.61	4.74
2005Q3	1.91	-0.60	1.87	9.99	0.74	1.56	13.97
2005Q4	1.88	-0.60	1.85	5.30	0.70	0.46	4.13
2006Q1	1.87	-0.60	1.84	5.30	0.69	0.50	4.09
2006Q2	1.86	-0.89	1.82	5.30	0.71	0.50	4.09
2006Q3	1.83	-2.05	1.80	5.30	0.74	0.30	4.58
2006Q4	1.81	-2.05	1.77	5.30	0.75	0.32	4.45
2007Q1	1.78	-2.19	1.73	5.30	0.76	0.30	4.46
2007Q2	1.75	-2.36	1.70	5.68	0.77	0.32	4.97
2007Q3	1.71	-1.67	1.67	5.68	0.75	0.40	4.94
2007Q4	1.67	-1.67	1.63	6.00	0.75	0.50	5.33
2008Q1	1.64	-2.20	1.59	15.92	0.88	4.21	61.22
2008Q2	1.59	-2.20	1.56	15.92	0.88	4.23	63.45
2008Q3	1.52	-2.61	1.51	15.92	0.90	3.69	57.87
2008Q4	1.46	-3.56	1.47	15.70	0.90	3.12	50.67
2009Q1	1.39	-2.61	1.42	6.53	0.81	-0.13	6.22
2009Q2	1.33	-2.61	1.37	6.53	0.83	-0.23	6.33
2009Q3	1.29	-4.10	1.35	7.53	0.89	-0.46	7.09
2009Q4	1.27	-4.10	1.33	7.53	0.87	-0.45	6.93
2010Q1	1.26	-3.59	1.32	7.53	0.86	-0.41	6.92
2010Q2	1.24	-3.59	1.30	5.83	0.85	-0.68	5.97
2010Q3	1.26	-3.54	1.32	5.83	0.85	-0.56	5.70
2010Q4	1.27	-3.78	1.32	7.29	0.88	-0.26	6.51
2011Q1	1.29	-3.32	1.34	7.29	0.87	-0.27	6.58
2011Q2	1.31	-3.32	1.36	8.65	0.90	0.10	8.05
2011Q3	1.31	-2.83	1.36	8.65	0.91	0.38	9.20
2011Q4	1.32	-2.83	1.36	7.98	0.88	0.26	8.57
2012Q1	1.31	-2.80	1.36	7.98	0.87	0.22	8.48
2012Q2	1.30	-2.87	1.35	7.98	0.88	0.24	8.46
2012Q3	1.32	-3.03	1.35	7.98	0.90	0.47	9.09
2012Q4	1.32	-3.03	1.35	7.98	0.89	0.49	9.36
2013Q1	1.33	-3.03	1.35	7.98	0.86	0.51	9.31
2013Q2	1.36	-2.87	1.34	22.32	1.07	7.15	125.30
2013Q3	1.32	-2.78	1.32	6.89	0.82	0.71	9.54
2013Q4	1.32	-2.78	1.29	22.32	1.03	7.39	133.39
2014Q1	1.31	-2.78	1.28	22.32	1.01	8.43	160.34
2014Q2	1.29	-2.78	1.28	7.75	0.79	1.38	13.12
2014Q3	1.33	-2.78	1.28	24.49	1.08	9.82	191.05

Notes: The descriptive statistics are computed for samples in which we pool observations across institutions and time periods. We did not weight the statistics by size of the institution.