Forecasting with Dynamic Panel Data Models

Laura Liu

University of Pennsylvania

Hyungsik Roger Moon

University of Southern California

USC Dornsife INET, and Yonsei

Frank Schorfheide[*]

University of Pennsylvania

CEPR, NBER, and PIER

December 21, 2016

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 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, Tweedies 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:

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

Here, $(W_{it-1}, X_{it-1}, Z_{it-1})$ are predictors 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}, assuming that the time dimension T is short relative to the number of predictors (W_{iT}, X_{iT}, Z_{iT}). The forecasts are evaluated under a quadratic loss function. In this setting an accurate forecasts not only requires a precise estimate of the common parameters (α, ρ), but also of the parameters λ_i that are specific to the cross-sectional units i. The existing literature on dynamic panel data models almost exclusively studied the estimation of the common parameters, treating the unit-specific parameters as a nuisance. Our paper builds on the insights of the dynamic panel literature 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 so-called oracle forecast. The oracle is assumed to know the common coefficients (α, ρ) as well as the distribution of the heterogeneous coefficients λ_i, denoted by $\pi(\lambda_i|\cdot)$. Note that this distribution could be conditional on

some observable characteristics of unit i. Because we are interested in forecasts for the entire cross section of N units, 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.

The implementation of the oracle forecast is infeasible because in practice neither the common coefficients (ρ, α) nor the distribution of the unit-specific coefficients $\pi(\lambda_i|\cdot)$ is known. To obtain a feasible predictor, we extend the classical posterior mean formula attributed to separate works of Arthur Eddington and Maurice Tweedie to our dynamic panel data setup. According to this formula, the posterior mean of λ_i can be expressed as a function of the cross-sectional density of certain sufficient statistics. Conditional on the common parameters, this distribution can then be estimated either parametrically or non-parametrically from the panel data set. The unknown common parameters can be replaced by a generalized method of moments (GMM) estimator, a likelihood-based correlated random effects estimator, or a Bayes estimator.

Our paper makes three contributions. First, we show in the context of the linear dynamic panel data model that a feasible predictor based on a consistent estimator of (ρ, α) and a non-parametric estimator of the cross-sectional density of the relevant sufficient statistics can achieve the same compound risk as the oracle predictor asymptotically. Our main theorem extends a result from Brown and Greenshtein (2009) for a vector of means to a panel data model with estimated common coefficients. Importantly, this result also covers the case in which the distribution $\pi(\lambda_i|\cdot)$ degenerates to a point mass. 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 of a panel of banks.

If the time series dimension is small, our feasible predictor performs much better than a naive predictor of Y_{iT+1} that is based on within-group estimates of λ_i. A small T leads to a noisy estimate of λ_i. Moreover, from a compound risk perspective, there will be a selection bias. Consider the special case of $\alpha = \rho = 0$ and $W_{it} = 1$. Here, λ_i is simply a heterogeneous

intercept. Very large (small) realizations of Y_{it} will be attributed to large (small) values of λ_i, which means that the within-group mean will be upward (downward) biased for those units. The use of a prior distribution estimated from the cross-sectional information essentially corrects this bias, which facilitates the reduction of the prediction risk if it is averaged over the entire cross section. Alternatively, one could ignore the cross-sectional heterogeneity and estimate a (misspecified) model with a homogeneous coefficient λ. If the heterogeneity is small, this procedure is likely to perform well in a mean-squared-error sense. However, as the heterogeneity increases, the performance of a predictor that is based on a pooled estimation quickly deteriorates. We illustrate the performance of various implementations of the feasible predictor in a Monte Carlo study and provide comparisons with other predictors, including one that is based on quasi maximum likelihood estimation of the unit-specific coefficients and one that is constructed from a pooled OLS estimator that ignores parameter heterogeneity.

In an empirical application we forecast pre-provision net revenues 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 one-year-ahead forecast accuracy of the posterior mean 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. We find that our proposed posterior mean predictor is considerably more accurate than a predictor that does not utilize any prior distribution. The posterior mean predictor shrinks the estimates of the unit-specific coefficients toward a common prior mean, which reduces its sampling variability. According to our estimates, the effect of stressed macroeconomic conditions on bank revenues is very 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 problem of estimating a vector of means, which is a classic problem in the statistic literature. 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 also have a long history in the statistics literature going back to Robbins (1956) (see Robert (1994) for a textbook treatment).

We use compound decision theory as in Robbins (1964), Brown and Greenshtein (2009), Jiang, Zhang, et al. (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). Fully Bayesian approaches to the analysis of dynamic panel data models have been developed in Chamberlain and Hirano (1999), Hirano (2002), Lancaster (2002).

The papers that are most closely related to ours are Gu and Koenker (2016a,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 papers focus on the use of the Kiefer-Wolfowitz estimator for the cross-sectional distribution of the sufficient statistics, whereas our paper explores various plug-in estimators for the homogeneous coefficients in combination with both parametric and nonparametric estimates of the cross-sectional distribution. Moreover, our paper establishes the ratio-optimality of the forecast and presents a different application. Finally, Liu (2016) develops a fully Bayesian (as opposed to empirical Bayes) approach to construct density forecast. She uses a Dirichlet process mixture to construct a prior for the distribution of the heterogeneous coefficients, which then is updated in view of the observed panel data.

There is an earlier panel forecast literature (e.g., see the survey article by Baltagi (2008) and its references) that is based on the best linear unbiased prediction (BLUP) proposed by Goldberger (1962). 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 that constructs the optimal forecast without restricting the class of forecasts. 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.

There is a small academic literature on econometric techniques for stress test. Most

papers 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 conducted by the Federal Reserve Board of Governors. An important paper in this literature is Covas, Rump, and Zakrajsek (2014), which uses quantile autoregressive models to forecast bank balance sheet and revenue components. We work with a much larger panel of bank holding companies that comprises, depending on the sample period, between 460 and 725 institutions.

The remainder of the paper is organized as follows. Section 2 introduces the panel data model considered in this paper, derives the likelihood function, and provides an important identification result. Decision theoretic foundations for the proposed predictor and a derivation of the oracle forecast are provided in Section 3. Section 4 discusses feasible implementation strategies for the predictor and we show in Section 5 in the context of a basic dynamic panel data model that our proposed predictor asymptotically has the same risk as the oracle forecast. A simulation study is provided in Section 6. The empirical application is presented in Section 7 and Section 8 concludes. Technical derivations, proofs, the description of the data set used in the empirical analysis, and further empirical results are relegated to the Appendix.

2   A Dynamic Panel Forecasting Model

We consider a panel with observations for cross-sectional units $i = 1, \ldots, N$ in periods $t = 1, \ldots, T$. Observation Y_{it} is assumed to be generated by (1). We distinguish three types of regressors. First, the $k_w \times 1$ vector W_{it} interacts with the heterogeneous coefficients λ_i. In many panel data applications $W_{it} = 1$, meaning that λ_i is simply a heterogenous intercept. 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 $W_{it} = (w_{1,t+1}, W_{2,it})$.[1] The dimensions of the two components are k_{w_1} and k_{w_2}, respectively. Second, X_{it} is a $k_x \times 1$ vector of sequentially exogenous predictors with homogeneous coefficients. The predictors X_{it} may include lags of Y_{it+1} and we collect all the predetermined variables other than the lagged dependent variable into the subvector $X_{2,it}$. Third, Z_{it} is a k_z-vector of strictly exogenous regressors, also with common coefficients.

---

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

Our main goal is to construct optimal forecasts of $(Y_{1T+1}, ..., Y_{NT+1})$ conditional on the entire panel observations $\{(Y_{it}, W_{it-1}, X_{it-1}, Z_{it-1}), i = 1, ..., N$ and $t = 1, ..., T$ using the forecasting model (1). An important special case of model (1) is the basic dynamic panel data model

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \tag{2}$$

which is obtained by setting $W_{it} = 1$, $X_{it} = Y_{it}$ and $\alpha = 0$. The restricted model (2) has been widely studied in the literature. However, most studies focus on consistently estimating the common parameter ρ in the presence of an increasing (with the cross-sectional dimension N) number of λ_is. In forecasting applications, we also need to estimate the λ_is. In Section 2.1 we specify the likelihood function for model (1) and in Section 2.2 we establish the identifiability of the model parameters, including the distribution of the heterogeneous coefficients λ_i.

2.1 The Likelihood Function

Let $Y_i^{t_1:t_2} = (Y_{it_1}, ..., Y_{it_2})$ and use a similar notation to collect W_{it}s, X_{it}s, and Z_{it}s. We begin by making some assumptions on the joint distribution of $\{Y_i^{1:T+1}, X_i^{0:T}, W_{2,i}^{0:T}, Z_i^{0:T}, \lambda_i\}_{i=1}^N$ conditional on the regression coefficients ρ and α and the vector of volatility parameters γ (to be introduced below). We drop the deterministic trend regressors $w_{1,t}$ from the notation for now. We use $\mathbb{E}[\cdot]$ to denote expectations and $\mathbb{V}[\cdot]$ to denote variances.

Assumption 2.1

(i) $(Y_i^{1:T+1}, \lambda_i, X_i^{0:T}, W_{2i}^{0:T}, Z_i^{0:T})$ *are independent across i.*

(ii) $(\lambda_i, X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T})$ *are iid with joint density*

$$\pi(\lambda, x_0, w_2^{0:T}, z^{0:T}) = \pi(\lambda|x_0, w_2^{0:T}, z^{0:T})\pi(x_0, w_2^{0:T}, z^{0:T}).$$

(iii) *For $t = 1, ..., T$, the distribution of $X_{2,it}$ conditional on $(Y_i^{1:t}, X_i^{0:t-1}, W_{2,i}^{0:T}, Z_i^{0:T})$ does not depend on the heterogeneous parameters λ_i and parameters $(\rho, \alpha, \gamma_1, ...\gamma_T)$.*

(iv) *The distribution of $(W_{2,i}^{0:T}, Z_i^{0:T})$ does not depend on λ_i and $(\rho, \alpha, \gamma_1, ..., \gamma_T)$.*

(v) $U_{it} = \sigma_t(X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T}, \gamma_t)V_{it}$, *where V_{it} is iid across $i = 1, ..., N$ and independent over $t = 1, ..., T+1$ with $\mathbb{E}[V_{it}] = 0$ and $\mathbb{V}[V_{it}] = 1$ for $t = 1, ..., T+1$ and $(V_{i1}, ..., V_{iT})$ are independent of $X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T}$. We assume $\sigma_t(X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T}, \gamma_t)$ is a function that depends on the unknown finite-dimensional parameter vector γ_t.*

Assumption 2.1(i) states that conditionally on the predictors, the Y_{it+1}s are cross-sectionally independent. Thus, we assume that all the spatial correlation in the dependent variables is due to the observed predictors. Assumption 2.1(ii) formalizes the correlated random effects assumption. The subsequent Assumptions 2.1(iii) and (iv) imply that λ_i may affect X_{it} only indirectly through $Y_i^{1:t}$ – an assumption that is clearly satisfied in the dynamic panel data model (2) – and that the strictly exogenous predictors do not depend on λ_i. In Assumption 2.1(v), we allow the unpredictable shocks U_{it} to be conditionally heteroskedastic in both the cross section and over time. We allow $\sigma_t(\cdot)$ to be dependent on the initial condition of the sequentially exogenous predictors, X_{i0}, and other exogenous variables. Because throughout the paper we assume that the time dimension T is small, the dependence through X_{i0} can generate a persistent ARCH effect.

We now turn to the likelihood function. We use lower case $(y_{it}, w_{it}, x_{it}, z_{it})$ to denote the realizations of the random variables $(Y_{it}, X_{it}, W_{it}, Z_{it})$. The parameters that control the volatilities $\sigma_t(\cdot)$ are stacked into the vector $\gamma = [\gamma_1', ..., \gamma_T']'$ and we collect the homogeneous parameters into the vector $\theta = [\alpha', \rho', \gamma']'$. We use $H_i = (X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T})$ for the exogenous conditioning variables and $h_i = (x_{i0}, w_{2,i}^{0:T}, z_i^{0:T})$ for their realization. Finally, we denote the density of V_i by $\varphi(v)$. Recall that we used $x_{2,it}$ to denote predetermined predictors other than the lagged dependent variable. According to Assumption 2.1(iii) the density $q_t(x_{2,it}|y_i^{1:t}, x_i^{0:t-1}, w_{2i}, z_i)$ does not provide any information about λ_i and will subsequently be absorbed into a constant of proportionality. Combining the likelihood function for the observables with the conditional distribution of the heterogeneous coefficients leads to

$$p(y_i, x_{2,i}, \lambda_i | h_i, \theta) \propto \left(\prod_{t=1}^{T} \frac{1}{\sigma_t(h_i, \gamma_t)} \varphi \left(\frac{y_{it} - \lambda_i' w_{it-1} - \rho' x_{it-1} - \alpha' z_{it-1}}{\sigma_t(h_i, \gamma_t)} \right) \right) \pi(\lambda_i | h_i). \quad (3)$$

Because conditional on the predictors the observations are cross-sectionally independent, the joint densities for observations $i = 1, \ldots, N$ can be obtained by taking the product across i of (3).

2.2   Identification

We now provide conditions under which the forecasting model (1) is identifiable. While the identification of the finite-dimensional parameter vector θ is fairly straightforward, the empirical Bayes approach pursued in this paper also requires the identification of the correlated random effects distribution $\pi(\lambda_i | h_i)$ from the cross-sectional information in the panel.

Before presenting a general result which is formally proved in the Online Appendix, we sketch the identification argument in the context of the restricted dynamic model (2) with heterogeneous intercept and heteroskedastic innovations.

The identification can be established in three steps. First, the identification of the homogeneous regression coefficient ρ follows from a standard argument used in the instrumental variable (IV) estimation of dynamic panel data models. To eliminate the dependence on λ_i define $Y_{it}^* = Y_{it} - \frac{1}{T-t}\sum_{s=t+1}^{T} Y_{is}$ and $X_{it-1}^* = Y_{it-1} - \frac{1}{T-t}\sum_{s=t+1}^{T} Y_{is-1}$. Then, because $\mathbb{E}[U_{it}|Y_i^{0:t-1}, \lambda_i] = 0$, the orthogonality conditions $\mathbb{E}\left[(Y_{it}^* - \rho X_{it-1}^*)Y_{it-1}\right] = 0$ for $t = 1, \ldots, T-1$ in combination with a relevant rank condition can be used to identify ρ (see, e.g., Arellano and Bover (1995)). Second, to identify the variance parameters γ, let Y_i, X_i, and U_i denote the $T \times 1$ vectors that stack Y_{it}, Y_{it-1}, and U_{it}, respectively, for $t = 1, \ldots, T$. Moreover, let ι be a $T \times 1$ vector of ones and define $\Sigma_i^{1/2}(\tilde{\gamma}) = \text{diag}\left(\sigma_1(h_i, \tilde{\gamma}_1), \ldots, \sigma_T(h_i, \tilde{\gamma}_T)\right)$, $S_i(\tilde{\gamma}) = \Sigma_i^{-1/2}(\tilde{\gamma})\iota$, and $M_i(\tilde{\gamma}) = I - S_i(S_i'S_i)^{-1}S_i'$. Using this notation, we obtain

$$M_i(\tilde{\gamma})\Sigma_i^{-1/2}(\tilde{\gamma})\left(Y_i - X_i\rho\right) = M_i(\tilde{\gamma})S_i(\tilde{\gamma})\lambda_i + M_i(\tilde{\gamma})\Sigma_i^{-1/2}(\tilde{\gamma})U_i = M_i(\tilde{\gamma})V_i.$$

This leads to the conditional moment condition

$$\mathbb{E}\left[M_i(\tilde{\gamma})\Sigma_i^{-1/2}(\tilde{\gamma})\left(Y_i - X_i\rho\right)\left(Y_i - X_i\rho\right)'\Sigma_i^{-1/2}(\tilde{\gamma})M_i'(\tilde{\gamma}) - M_i(\tilde{\gamma})\Big|H_i\right] = 0 \tag{4}$$

if and only if $\tilde{\gamma} = \gamma$, which identifies γ. Third, let

$$\tilde{Y}_i = \Sigma_i^{-1/2}(\gamma)\left(Y_i - X_i\rho\right) = S_i(\gamma)\lambda_i + V_i. \tag{5}$$

The identification of $\pi(\lambda_i|h_i)$ can be established using a characteristic function argument similar to that in Arellano and Bonhomme (2012). For the general model (1) we make the following assumptions:

Assumption 2.2

(i) *The parameter vectors α and ρ are identifiable.*

(ii) *For each $t = 1, \ldots, T$ and almost all h_i $\sigma_t^2(h_i, \tilde{\gamma}_t) = \sigma_t^2(h_i, \gamma_t)$ implies $\tilde{\gamma}_t = \gamma_t$. Moreover, $\sigma_t^2(h_i, \gamma_t) > 0$.*

(iii) *The characteristic functions for $\lambda_i|(H_i = h_i)$ and V_i are non-vanishing almost everywhere.*

(iv) *$W_i = [W_{i0}, ..., W_{iT-1}]'$ has full rank k_w.*

Because the identification of α and ρ in panel data models with fixed or random effects is well established, we make the high-level Assumption 2.2(i) that the homogeneous parameters are identifiable.[2] We discuss in the appendix how the identification argument for ρ in the basic dynamic panel data model can be extended to a more general specification as in (1). Assumption 2.2(ii) enables us to identify the volatility parameters γ, and (iii) and (iv) deliver the identifiability of the distribution of heterogeneous coefficients. The following theorem summarizes the identification result and is proved in the Appendix.

Theorem 2.3 *Suppose that Assumptions 2.1 and 2.2 are satisfied. Then the parameters α, ρ, and γ as well as the correlated random effects distribution $\pi(\lambda_i|h_i)$ and the distribution of V_{it} in model (1) are identified.*

3 Decision-Theoretic Foundation

We adopt a decision-theoretic framework in which forecasts are evaluated based on cross-sectional sums of mean-squared error losses. Such losses are called compound loss functions. Section 3.1 provides a formal definition of the compound risk (expected loss). In Section 3.2 we derive the optimal forecasts under the assumption that the cross-sectional distribution of the λ_is is known (oracle forecast). While it is infeasible to implement this forecast in practice, the oracle forecast provides a natural benchmark for the evaluation of feasible predictors. Finally, in Section 3.3 we introduce the concept of ratio optimality, which describes forecasts that asymptotically (as $N \longrightarrow \infty$) attain the same risk as the oracle forecast.

3.1 Compound Risk

Let $L(\widehat{Y}_{iT+1}, Y_{iT+1})$ denote the loss associated with forecast $\widehat{Y}_{i,T+1}$ of individual $i's$ time $T+1$ observation, Y_{iT+1}. In this paper we consider the conventional quadratic loss function,

$$L(\widehat{Y}_{iT+1}, Y_{iT+1}) = (\widehat{Y}_{iT+1} - Y_{iT+1})^2.$$

The main goal of the paper is to construct optimal forecasts for groups of individuals selected by a known selection rule in terms of observed data. We express the selection rule as

$$D_i = D_i(\mathcal{Y}^N) \in \{0, 1\}, \quad i = 1, \dots, N, \tag{6}$$

---

[2]Textbook / handbook chapter treatments can be found in, for instance, Baltagi (1995), Arellano and Honoré (2001), Arellano (2003) and Hsiao (2014).

where $D_i(\mathcal{Y}^N)$ is a measurable function of the observations $\mathcal{Y}^N$, $\mathcal{Y}^N = (\mathcal{Y}_1, \ldots, \mathcal{Y}_N)$, and $\mathcal{Y}_i = (Y_i^{0:T}, X_i^{1:T}, H_i)$. For instance, suppose that $D_i(\mathcal{Y}^N) = \mathbb{I}\{Y_{iT} \in A\}$ for $A \subset \mathbb{R}$. In this case, the selection is homogeneous across i and, for individual i, depends only on its own sample. Alternatively, suppose that units are selected based on the ranking of an index, e.g., the empirical quantile of Y_{iT}. In this case, the selection dummy D_i depends on $(Y_{1T}, \ldots, Y_{NT})$ and thereby also on the data for the other $N-1$ individuals.

The compound loss of interest is the average of the individual losses weighted by the selection dummies:

$$L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N) = \sum_{i=1}^N D_i(\mathcal{Y}^N) L(\widehat{Y}_{iT+1}, Y_{iT+1}),$$

where $Y_{T+1}^N = (Y_{1T+1}, \ldots, Y_{NT+1})$. The compound risk is the expected compound loss

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_\theta^{\mathcal{Y}^N, \lambda^N, U_{T+1}^N} \left[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N) \right]. \tag{7}$$

We use the θ subscript for the expectation operator to indicate that the expectation is conditional on θ.[3] The superscript $(\mathcal{Y}^N, \lambda^N, U_{T+1}^N)$ indicates that we are integrating with respect to the observed data $\mathcal{Y}^N$ and the unobserved heterogeneous coefficients $\lambda^N = (\lambda_1, \ldots, \lambda_N)$ and $U_{T+1}^N = (U_{1T+1}, \ldots, U_{NT+1})$.

3.2 Optimal Forecast and Oracle Risk

We now derive the optimal forecast that minimizes the compound risk. The risk achieved by the optimal forecast will be called the oracle risk, which is the target risk to achieve. In the compound decision theory it is assumed that the oracle knows the vector θ as well as the distribution of the heterogeneous coefficients $\pi(\lambda_i, h_i)$ and observes $\mathcal{Y}^N$. However, the oracle does not know the specific λ_i for unit i. In order to find the optimal forecast, note that conditional on θ the compound risk takes the form of an integrated risk that can be expressed as

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_\theta^{\mathcal{Y}^N} \left[\mathbb{E}_{\theta, \mathcal{Y}^N}^{\lambda^N, U_{T+1}^N} [L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)] \right]. \tag{8}$$

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 $\mathcal{Y}^N$.

---

[3]Strictly speaking, the expectation also conditions on the deterministic trend terms W_1

It is well known that the integrated risk is minimized by choosing the forecast that minimizes the posterior risk for each realization $\mathcal{Y}^N$. Using the independence across i, the posterior risk can be written as follows:

$$\mathbb{E}_{\theta,\mathcal{Y}^N}^{\lambda^N,U_{T+1}^N}[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)] \tag{9}$$

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

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

$$\widehat{Y}_{iT+1}^{opt} = \mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i,U_{iT+1}}[Y_{iT+1}] = \mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i]'W_{iT} + \rho'X_{iT} + \alpha'Z_{iT}. \tag{10}$$

Note that the posterior expectation of λ_i only depends on observations for unit i, even if the selection rule $D_i(\mathcal{Y}^N)$ also depends on the data from other units $j \neq i$. The result is summarized in the following theorem:

Theorem 3.1 (Optimal Forecast) *Suppose Assumptions 2.1 are satisfied. The optimal forecast that minimizes the composite risk in (7) is given by $\widehat{Y}_{iT+1}^{opt}$ in (10). The compound risk of the optimal forecast is*

$$R_N^{opt} = \mathbb{E}_\theta^{\mathcal{Y}^N} \left[\sum_{i=1}^N D_i(\mathcal{Y}^N) \left(W_{iT}' \mathbb{V}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] W_{iT} + \sigma_{T+1}^2(H_i, \gamma_{T+1}) \right) \right]. \tag{11}$$

According to (11), 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 optimal 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.

3.3 Ratio Optimality

The identification result presented in Section 2.2 implies that as the cross-sectional dimension $N \longrightarrow \infty$, it might be possible to learn the unknown parameter θ and random-effects distribution $\pi(\cdot)$ and construct a feasible estimator that asymptotically attains the oracle

risk. Following Brown and Greenshtein (2009), we say that a predictor achieves ratio optimality if the regret $R_N(\widehat{Y}_{T+1}^N) - R_N^{\text{opt}}$ of the forecast $\widehat{Y}_{T+1}^N$ is negligible relative to the part of the optimal risk that is due to uncertainty about λ_i:

Definition 3.2 *For a given $\epsilon_0 > 0$, we say that forecast $\widehat{Y}_{T+1}^N$ achieves ϵ_0-ratio optimality, if*

$$\limsup_{N \to \infty} \frac{R_N(\widehat{Y}_{T+1}^N) - R_N^{opt}}{\mathbb{E}_\theta^{\mathcal{Y}^N} \left[\sum_{i=1}^N D_i(\mathcal{Y}^N) W_{iT}' \mathbb{V}_{\theta, \mathcal{Y}_i}^{\lambda_i}[\lambda_i] W_{iT} \right] + N^{\epsilon_0}} \leq 0. \tag{12}$$

Using (10), the risk differential in the numerator (called regret) can be written as

$$R_N(\widehat{Y}_{T+1}^N) - R_N^{\text{opt}} = \mathbb{E}_\theta^{\mathcal{Y}^N} \left[\sum_{i=1}^N D_i(\mathcal{Y}^N) \left(\widehat{Y}_{iT+1} - \mathbb{E}_{\theta, \mathcal{Y}_i}^{\lambda_i, U_{iT+1}}[Y_{iT+1}] \right)^2 \right]. \tag{13}$$

For illustrative purposes, Consider the basic dynamic panel data model (2). For this model $\mathbb{E}_{\theta, \mathcal{Y}_i}^{\lambda_i, U_{iT+1}}[Y_{iT+1}] = \mathbb{E}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i] + \rho Y_{iT}$. A natural class of predictors is given by $\widehat{Y}_{iT+1} = \widehat{\mathbb{E}}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i] + \hat{\rho} Y_{iT}$, where $\widehat{\mathbb{E}}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ is an approximation of the posterior mean of λ_i that replaces the unknown ρ and distribution $\pi(\cdot)$ by suitable estimates. The autoregressive coefficient in this 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 attains ratio optimality crucially depends on the rate at which the discrepancy between $\mathbb{E}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ and $\widehat{\mathbb{E}}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ vanishes.

The denominator of the ratio in Definition 3.2 is divergent. The rate of divergence depends on the posterior variance of λ_i. If the posterior variance is strictly greater than zero, then the denominator is of order $O(N)$. Note that for each unit i, the posterior variance is based on a finite number of observations T. Thus, for the posterior variance to be equal to zero, it must be the case that the prior density $\pi(\lambda)$ is a pointmass, meaning that there is a homogeneous intercept λ. 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}. Subsequently, we will pursue an empirical Bayes strategy to construct an approximation $\widehat{\mathbb{E}}_{\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ based on the cross-sectional information and show that it attains ratio-optimality.

In the linear panel literature, researchers often use the first difference to eliminate λ_i. In this case, the natural forecast of Y_{iT+1} in the basic dynamic panel data model (2) would be $\widehat{Y}_{iT+1}^{FD}(\rho) = Y_{iT} + \rho(Y_{iT} - Y_{iT-1})$, which is different from $\widehat{Y}_{iT+1}^{opt}$ in (10). Thus, we can immediately deduce from Theorem 3.1 that $\widehat{Y}_{iT+1}^{FD}(\rho)$ is not an optimal forecast. The quasi-differencing of Y_{it} introduces a predictable moving-average error term that is ignored by the predictor $\widehat{Y}_{iT+1}^{FD}(\rho)$.

4 Implementation of the Optimal Forecast

We will construct a consistent approximation of the posterior mean $\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i,U_{iT+1}}[\lambda_i]$ using a convenient formula which is named after the statistician Maurice Tweedie (though it had been previously derived by the astronomer Arthur Eddington). This formula is presented in Section 4.1. In Section 4.2 we discuss the parametric estimation of the correction term and in Section 4.3 we consider a nonparametric kernel-based estimation. The QMLE and Generalized Method-of-Moments (GMM) estimation of the parameter θ are discussed in Sections 4.4 and 4.5.

4.1 Tweedie's Formula

When the innovations U_{it} are conditionally normally distributed, we can derive a convenient formula for the posterior expectation $\mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ of the individual heterogeneous parameter λ_i.

Assumption 4.1 *The unpredictable shock V_{it} has a standard normal distribution:*

$$V_{it} \mid (Y_i^{1:t-1}, X_i^{0:t-1}, W_{2i}, Z_i, \lambda_i) \sim N(0,1), \quad t = 1, ..., T.$$

The assumption of normally distributed V_{it}'s is not as restrictive as it may seem. Recall that the shocks U_{it} are defined as $V_{it}\sigma_t(X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T}, \gamma_t)$. Thus, due to the potential heteroskedasticity, the distribution of shocks is a mixture of normals. The only restriction is that the random variables characterizing the scale of the mixture component are observed. Moreover, 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. Using Assumption 4.1 we will now further manipulate the density $p(y_i, x_{2,i}, \lambda_i|h_i, \theta)$ in (3).[4] To simplify the notation we will drop the i subscript. Define

$$\tilde{y}_t(\theta) = y_t - \rho'x_{t-1} - \alpha'z_{t-1}, \quad \Sigma(\theta) = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2), \tag{14}$$

and let $\tilde{y}(\theta)$ and w be matrices with rows $\tilde{y}_t(\theta)$ and w'_{t-1}, $t = 1, ..., T$. Because the subsequent calculations condition on θ we will omit the θ-argument from $\tilde{y}$, Σ, and functions thereof.

---

[4]In principle, the normality assumption could be generalized to the assumption that the distribution of V_{it} belongs to the exponential family.

Replacing $\varphi(v)$ in (3) with a Gaussian density function we obtain:

$$
\begin{aligned}
&p(y, x_2, \lambda | h, \theta) \\
&\quad \propto \ \exp\left\{-\frac{1}{2}(\hat{\lambda} - \lambda)'w'\Sigma^{-1}w(\hat{\lambda} - \lambda)\right\} \exp\left\{-\frac{1}{2}(\tilde{y} - w\hat{\lambda})'\Sigma^{-1}(\tilde{y} - w\hat{\lambda})\right\} \pi(\lambda | h).
\end{aligned}
$$

The factorization of $p(y, x_2, \lambda | h, \theta)$ implies that

$$
\hat{\lambda} = (w'\Sigma^{-1}w)^{-1}w'\Sigma^{-1}\tilde{y} \tag{15}
$$

is a sufficient statistic and that we can express the posterior distribution of λ as

$$
p(\lambda | y, x_2, h, \theta) = p(\lambda | \hat{\lambda}, h, \theta) = \frac{p(\hat{\lambda} | \lambda, h, \theta)\pi(\lambda | h)}{p(\hat{\lambda} | h, \theta)},
$$

where

$$
p(\hat{\lambda} | \lambda, h, \theta) = (2\pi)^{-k_w/2}|w'\Sigma^{-1}w|^{1/2}\exp\left\{-\frac{1}{2}(\hat{\lambda} - \lambda)'w'\Sigma^{-1}w(\hat{\lambda} - \lambda)\right\}. \tag{16}
$$

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

$$
\begin{aligned}
0 &= w'\Sigma^{-1}w \int (\lambda - \hat{\lambda})p(\lambda | \hat{\lambda}, h, \theta)d\lambda - \frac{\partial}{\partial\hat{\lambda}}\ln p(\hat{\lambda} | h, \theta) \\
&= w'\Sigma^{-1}w\big(\mathbb{E}_{\theta,\mathcal{Y}}^{\lambda}[\lambda] - \hat{\lambda}\big) - \frac{\partial}{\partial\hat{\lambda}}\ln p(\hat{\lambda} | h, \theta).
\end{aligned}
$$

Solving this equation for the posterior mean yields Tweedie's formula, which is summarized in the following theorem.

Theorem 4.2 *Suppose that Assumptions 2.1 and 4.1 hold. The posterior mean of λ_i has the representation*

$$
\mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] = \hat{\lambda}_i(\theta) + \left(W_i^{0:T-1'}\Sigma^{-1}(\theta)W_i^{0:T-1}\right)^{-1}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\theta) | H_i, \theta). \tag{17}
$$

The optimal forecast is given by

$$
\begin{aligned}
\widehat{Y}_{iT+1}^{opt}(\theta) &= \left(\hat{\lambda}_i(\theta) + \left(W_i^{0:T-1'}\Sigma^{-1}(\theta)W_i^{0:T-1}\right)^{-1}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\theta) | H_i, \theta)\right)' W_{T+1} \\
&\quad + \rho'X_{iT} + \alpha'Z_{iT}. \tag{18}
\end{aligned}
$$

Tweedie's formula was used by Robbins (1951) to estimate a vector of means λ^N for the model $Y_i|\lambda_i \sim N(\lambda_i, 1)$, $\lambda_i \sim \pi(\cdot)$, $i = 1, \ldots, N$. Recently, it was extended by Efron (2011) to the family of exponential distribution, allowing for a unknown finite-dimensional parameter θ. Theorem 4.2 extends Tweedie's formula to the estimation of correlated random effect parameters in a dynamic panel regression setup.

The posterior mean takes the form of the sum of the sufficient statistic $\hat{\lambda}_i(\theta)$ and a correction term that reflects the prior distribution of λ_i. The correction term is expresses as a function of the marginal density of the sufficient statistic $\hat{\lambda}_i(\theta)$ conditional on H_i and θ. Thus, it is not necessary to solve a deconvolution problem that separates the prior density $\pi(\lambda_i|h_i)$ from the distribution of the error terms V_{it}. We expressed Tweedie's formula in (17) in terms of the conditional density $p(\hat{\lambda}_i(\theta)|H_i, \theta)$. However, because the posterior mean is a function of the log density differentiated with respect to $\hat{\lambda}_i(\theta)$, the conditional density can be replaced by a joint density:

$$\frac{\partial}{\partial \hat{\lambda}_i(\theta)} \ln p(\hat{\lambda}_i(\theta)|H_i, \theta) = \frac{\partial}{\partial \hat{\lambda}_i(\theta)} \ln p(\hat{\lambda}_i(\theta), H_i|\theta).$$

The construction of ratio-optimal forecasts relies on replacing the density $p(\hat{\lambda}_i(\theta), H_i|\theta)$ and the common parameter θ by consistent estimates.

4.2 Parametric Estimation of Tweedie Correction

If the random-effects distribution $\pi(\lambda|h_i)$ is Gaussian, then it is possible to derive the marginal density of the sufficient statistic $p(\hat{\lambda}_i(\theta)|h_i, \theta)$ analytically. Let

$$\lambda_i|(H_i, \theta) \sim N\big(\Phi H_i, \underline{\Omega}\big). \tag{19}$$

Moreover, define $\xi = \big(\text{vec}(\Phi), \text{vech}(\underline{\Omega})\big)'$. To highlight the dependence of the correlated random-effects distribution on the hyperparameter ξ we will write $\pi(\lambda_i|h_i, \xi)$. The marginal density (omitting the i subscripts and the θ-argument of $\hat{\lambda}$) is given by

$$\begin{aligned}
p\big(\hat{\lambda}(\theta)\big|h, \theta, \xi\big) &= \int p\big(\hat{\lambda}(\theta)|\lambda, h, \theta\big)\pi(\lambda|h, \xi)d\lambda \\
&= (2\pi)^{-k_w/2}\big|\underline{\Omega}^{-1}\big|^{1/2}\big|w'\Sigma^{-1}w\big|^{1/2}\big|\bar{\Omega}\big|^{1/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\big(\hat{\lambda}'w'\Sigma^{-1}w\hat{\lambda} + h'\Phi'\underline{\Omega}^{-1}\Phi h - \bar{\lambda}'\bar{\Omega}^{-1}\bar{\lambda}\big)\right\}.
\end{aligned} \tag{20}$$

Here, we used the likelihood of $\hat{\lambda}$ in (16), the density associated with the Gaussian prior in (19), and then the properties of a multivariate Gaussian density to integrate out λ. The terms $\bar{\lambda}$ and $\bar{\Omega}$ are the posterior mean and variance of λ, respectively:

$$\bar{\Omega}^{-1} = \underline{\Omega}^{-1} + w'\Sigma^{-1}w, \quad \bar{\lambda} = \bar{\Omega}\big(\underline{\Omega}^{-1}\Phi h + w'\Sigma^{-1}w\hat{\lambda}\big).$$

Conditional on θ the vector of hyperparameters ξ can be estimated by maximizing the marginal likelihood

$$\hat{\xi}(\theta) = \operatorname{argmax}_\xi \prod_{i=1}^N p(\hat{\lambda}_i(\theta)|h_i, \theta, \xi) \tag{21}$$

using the cross-sectional distribution of the sufficient statistic. Tweedie's formula can then be evaluated based on $p\big(\hat{\lambda}_i(\theta)|h_i, \theta, \hat{\xi}(\theta)\big)$. In principle it is possible to replace the Gaussian prior distribution with a more general parametric distribution. However, in general it will not be possible to derive an analytical formula for the marginal likelihood.

4.3   Nonparametric Estimation of Tweedie Correction

A nonparametric implementation of the Tweedie correction can be obtained by replacing $p(\hat{\lambda}_i(\theta), h_i|\theta)$ and its derivative with respect to $\hat{\lambda}_i(\theta)$ with a Kernel density estimate, e.g.,

$$\hat{p}(\hat{\lambda}_i(\theta), h_i|\theta) \tag{22}$$
$$= \frac{1}{N} \sum_{j=1}^N \Bigg[(2\pi)^{-k_w/2}|B_N|^{-k_w}|V_{\hat{\lambda}}|^{-1/2} \exp\left\{ -\frac{1}{2B_N^2}(\hat{\lambda}_i(\theta) - \hat{\lambda}_j(\theta))'V_{\hat{\lambda}}^{-1}(\hat{\lambda}_i(\theta) - \hat{\lambda}_j(\theta)) \right\}$$
$$\times (2\pi)^{-k_h/2}|B_N|^{-k_h}|V_h|^{-1/2} \exp\left\{ -\frac{1}{2B_N^2}(h_i - h_j)'V_h^{-1}(h_i - h_j) \right\} \Bigg],$$

where B_N is the bandwidth and $V_{\hat{\lambda}}$ and V_h are tuning matrices. Note that even if the prior distribution $\pi(\lambda)$ is a pointmass, the sufficient statistic $\hat{\lambda}$ in (15) has a continuous distribution and one can use a kernel density estimator to construct the Tweedie correction.

If the dimension of the conditioning variables H_i is large, the nonparametric estimation suffers from the curse of dimensionality. In this case, one may reduce the dimension of the conditioning set with some smaller dimensional indices, e.g., by assuming that λ_i and H_i dependent only through $\bar{H}_i = \frac{1}{T}\sum_{t=1}^T H_{it}$, that is, $\pi(\lambda|h) = \pi(\lambda|\bar{h})$. In Section 5 we provide a detailed analysis of the Gaussian kernel estimator in the context of the basic dynamic panel data model in (2) with time-homoskedastic innovations.

4.4   QMLE Estimation of θ

Notice that under Assumption 4.1, $\hat{\lambda}_i(\theta)$ in (15) is a sufficient statistic of λ_i conditional on θ, h_i, and $\pi_\lambda(\lambda_i|h_i, \xi)$ is the parametric version of the correlated random effect density. Integrating out λ under a parametric correlated random effect (or prior) distribution $\pi_\lambda(\lambda|x_0, w_2, z, \xi)$, we have (omitting the i subscripts)

$$
\begin{aligned}
p(y, & x_2|h, \theta, \xi) \hspace{6cm} (23)\\
= & \int p(y, x_2|h, \theta, \lambda)\pi_\lambda(\lambda|h, \hat{\xi}(\theta))d\lambda \\
\propto & \ |\Sigma(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}\big(\tilde{y}(\theta) - w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\theta)\big(\tilde{y}(\theta) - w\hat{\lambda}(\theta)\big)\right\} \\
& \times \int \exp\left\{-\frac{1}{2}\big(\hat{\lambda}(\theta) - \lambda\big)'w'\Sigma^{-1}(\theta)w\big(\hat{\lambda}(\theta) - \lambda\big)\right\}\pi_\lambda\big(\lambda(\theta)|h, \hat{\xi}(\theta)\big)d\lambda \\
\propto & \ |\Sigma(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}\big(\tilde{y}(\theta) - w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\theta)\big(\tilde{y}(\theta) - w\hat{\lambda}(\theta)\big)\right\} \\
& \times \big|w'\Sigma^{-1}w\big|^{-1/2}p(\hat{\lambda}(\theta)|h, \theta, \xi).
\end{aligned}
$$

Here, we used the definition of $\tilde{y}(\theta)$ in (14) and the product of Gaussian likelihood and prior in (15). Note that the term $p(\hat{\lambda}(\theta)|h, \theta, \xi)$ in the last line of (23) is identical to the objective function for ξ used in (21). Thus, we can now jointly determine θ and ξ by maximizing the integrated likelihood as a function:

$$
\big(\hat{\theta}_{QMLE}, \hat{\xi}_{QMLE}\big) = \mathrm{argmax}_{\theta, \xi} \prod_{i=1}^{N} p(y_i, x_{2i}|h_i, \theta, \xi). \hspace{2cm} (24)
$$

We refer to this estimator as *quasi* (Q) maximum likelihood estimator (MLE), because the correlated random effects distribution could be misspecified.

4.5   GMM Estimation of θ

Without a convenient assumption about the random effects distribution, one can estimate the parameter θ using a sample analogue of the moment conditions that were used in the identification analysis in Section 2. For $t = 1, \ldots, T - k_w$, define

$$
Y_{it}^* = Y_{it} - \left(\sum_{s=t+1}^{T} Y_{is}W_{is-1}'\right)\left(\sum_{s=t+1}^{T} W_{is-1}W_{is-1}'\right)^{-1}W_{it-1}. \hspace{1.5cm} (25)
$$

Moreover, define X^*_{it-1} and Z^*_{it-1} by replacing $Y_{i.}$ in (25) with $X_{i.}$ and $Z_{i.}$, respectively, and let

$$
g_{it}(\rho, \alpha) = (Y^*_{it} - \rho' X^*_{it-1} - \alpha' Z^*_{it-1}) \begin{bmatrix} X_i^{0:t-1} \\ Z_i^{0:T} \end{bmatrix}, \quad g_i(\rho, \alpha) = \left[g_{i1}(\rho, \alpha)', \dots, g_{iT-k_w}(\rho, \alpha)' \right]'.
$$

The continuous-updating GMM estimator of ρ and α solves

$$
(\hat{\rho}_{GMM}, \hat{\alpha}_{GMM}) = \operatorname*{argmin}_{\rho, \alpha} \left(\sum_{i=1}^N g_i(\rho, \alpha) \right)' \left(\sum_{i=1}^N g_i(\rho, \alpha) g_i(\rho, \alpha)' \right)^{-1} \left(\sum_{i=1}^N g_i(\rho, \alpha) \right). \quad (26)
$$

This estimator was proposed by Arellano and Bover (1995) and we will refer to it as GMM(AB) estimator in the Monte Carlo simulations (Section 6) and the empirical application (Section 7).[5]

To estimate the heteroskedasticity parameter $\gamma = [\gamma_1, ..., \gamma_T]'$ in $\sigma_t^2(H_i, \gamma_t)$, define:

$$
\begin{aligned}
\tilde{Y}_i(\hat{\rho}, \hat{\alpha}) &= Y_i - X_{i,-T}\hat{\rho} - Z_{i,-T}\hat{\alpha}, \quad \Sigma_i^{1/2}(\gamma) = \operatorname{diag}\big(\sigma_1(h_i, \gamma_1), \dots, \sigma_T(h_i, \gamma_T)\big), \\
S_i(\gamma) &= \Sigma_i^{-1/2}(\gamma) W_i, \quad M_i(\gamma) = I - S_i(S_i' S_i)^{-1} S_i',
\end{aligned}
$$

where $\hat{\rho}$ and $\hat{\alpha}$ could be the estimators in (26). We use the sample analogue to a set of moment condition implied by a generalization of (4):

$$
\begin{aligned}
\hat{\gamma}_{GMM} = \quad \operatorname{argmin}_\gamma \frac{1}{N} \sum_{i=1}^N \bigg\| B \operatorname{vec}\bigg(M_i(\gamma) \Sigma_i^{-1/2}(\gamma) \tilde{Y}_i(\hat{\rho}, \hat{\alpha}) \quad &(27) \\
\times \tilde{Y}_i'(\hat{\rho}, \hat{\alpha}) \Sigma_i^{-1/2}(\gamma) M_i(\gamma) - M_i(\gamma) \bigg) \bigg\|^2,
\end{aligned}
$$

where B is a selection matrix that can be used to eliminate off-diagonal elements of the covariance matrix. In population, these off-diagonal elements should be zero, because the U_{it}'s are assumed to be uncorrelated across time.

4.6 Extension to 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.

---

[5]There exists a large literature on the estimation of dynamic panel data models. Alternative estimators include Arellano and Bond (1991) and Blundell and Bond (1998).

We can express

$$Y_{iT+h} = \left(\sum_{s=0}^{h-1} \rho^s\right)\lambda_i + \rho^h Y_{iT} + \sum_{s=0}^{h-1}\rho^2 U_{iT+h-s}.$$

Under the assumption that the oracle knows ρ and $\pi(\lambda_i, Y_{i0})$ we can express the oracle forecast as

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

As in the case of the one-step-ahead forecasts, the posterior mean $\mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ 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.

5   Ratio Optimality in the Basic Dynamic Panel Model

Throughout this section we will consider the basic dynamic panel data model with homoskedastic Gaussian innovations:

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \quad U_{it} \sim iidN(0, \sigma^2), \quad (\lambda_i, Y_{i0}) \sim \pi(\lambda, y_{i0}). \tag{28}$$

We will prove that ratio optimality for a general prior density $\pi(\lambda_i|h_i)$ can be achieved with a Kernel estimator of the joint density of the sufficient statistic and initial condition: $p(\hat{\lambda}_i(\theta), H_i|\theta)$. The proof of the main result is a significant generalization of the proof in Brown and Greenshtein (2009) for a vector of means to the dynamic panel data model with estimated common coefficients.

For the model in (28), the sufficient statistic is given by

$$\hat{\lambda}_i(\rho) = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \rho Y_{it-1}) \tag{29}$$

and the posterior mean of λ_i simplifies to

$$\mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] = \mu\big(\hat{\lambda}_i(\rho), \sigma^2/T, p(\hat{\lambda}_i, Y_{i0})\big) = \hat{\lambda}_i(\rho) + \frac{\sigma^2}{T}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\rho), Y_{i0}). \tag{30}$$

The formula recognizes that the heterogeneous coefficient is a scalar intercept and that the errors are homoskedastic. We simplified the notation by writing $p(\hat{\lambda}_i(\rho), Y_{i0})$ instead of $p(\hat{\lambda}_i(\rho), Y_{i0}|\theta)$. This simplification is justified because we will estimate the density of $(\hat{\lambda}_i(\rho), Y_{i0})$ directly from the data; see (31) below. We will use the notation $\mu(\cdot)$ to refer to the conditional mean as function of the sufficient statistic $\hat{\lambda}$, the scale factor σ^2/T, and the density $p(\hat{\lambda}_i, Y_{i0})$.

To facilitate the theoretical analysis, we make two adjustments to the posterior mean predictor of Y_{iT+1}. First, we replace the kernel density estimator of $(\hat{\lambda}_i(\rho), Y_{i0})$ given in (22) by a leave-one-out estimator of the form:

$$\hat{p}^{(-i)}(\hat{\lambda}_i(\rho), Y_{i0}) = \frac{1}{N-1} \sum_{j \neq i} \frac{1}{B_N} \phi \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \right) \frac{1}{B_N} \phi \left(\frac{Y_{j0} - Y_{i0}}{B_N} \right), \qquad (31)$$

where $\phi(\cdot)$ is the pdf of a $N(0,1)$. 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:

$$
\begin{aligned}
\mathbb{E}_{\theta, \mathcal{Y}_i}^{\mathcal{Y}^{(-i)}} [\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] &= \int \frac{1}{\sqrt{\sigma^2/T + B_N^2}} \phi \left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{\sigma^2/T + B_N^2}} \right) \\
&\quad \times \left[\int \frac{1}{B_N} \phi \left(\frac{y_{i0} - \tilde{y}_{i0}}{B_N} \right) p(\tilde{y}_{i0}|\lambda_i) d\tilde{y}_{i0} \right] p(\lambda_i) d\lambda_i.
\end{aligned}
\qquad (32)
$$

Taking expectations of the kernel estimator leads to a variance adjustment for 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 convolution.

Second, we replace the scale factor $\hat{\sigma}^2/T$ in the posterior mean function $\mu(\cdot)$ by $\hat{\sigma}^2/T + B_N^2$, which is the term that appears in (32). Moreover, we truncate the absolute value of the posterior mean function from above. For $C > 0$ and for any $x \in \mathbb{R}$, define $[x]^C := \text{sgn}(x) \min\{|x|, C\}$. Then

$$\widehat{Y}_{iT+1} = \left[\mu \left(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{-i}(\cdot) \right) \right]^{C_N} + \hat{\rho} Y_{iT}, \qquad (33)$$

where $C_N \longrightarrow \infty$ slowly. Formally, we make the following technical assumptions.

Assumption 5.1 (Marginal distribution of λ_i) *The marginal density of λ_i, $\pi(\lambda)$ has support $\Lambda^\pi \subset [-C_N, C_N]$, where for any $\epsilon > 0$, $C_N = o(N^\epsilon)$.*

Assumption 5.2 (Bandwidth) *Let $C'_N = (1+k)(\sqrt{\ln N} + C_N)$, where k is a constant such that $k > \max\{0, \sqrt{2\sigma^2/T} - 1\}$. The bandwidth for the kernel density estimator, B_N, satisfies the following conditions: (i) for any $\epsilon > 0$, $1/B_N^2 = o(N^\epsilon)$; (ii) $B_N(C'_N + 2C_N) = o(1)$.*

Assumption 5.3 (Conditional distribution of $Y_{i0}|\lambda_i$) *Let $\mathcal{Y}_\lambda^\pi$ be the support of the conditional density $\pi(y_{i0}|\lambda_i)$. The conditional density of Y_{i0} conditioning on $\lambda_i = \lambda$, $\pi(y|\lambda)$, satisfies the following three conditions: (i) $0 < \pi(y|\lambda) < M$ for $y \in \mathcal{Y}_\lambda^\pi$ and $\lambda \in \Lambda^\pi$. (ii) There exists a finite constant $\bar{C}$ such that for any large value $C > \bar{C}$,*

$$\max\left\{\int_C^\infty \pi(y|\lambda)dx, \int_{-\infty}^{-C} \pi(y|\lambda)dy\right\} \leq \exp(-m(C,\lambda)),$$

where the function $m(C,\lambda) > 0$ satisfies the following: $m(C,\lambda)$ is an increasing function of C for each λ and there exists finite constants $K > 0$ and $\epsilon \geq 0$ such that

$$\liminf_{N \longrightarrow \infty} \inf_{|\lambda| \leq C_N} \left(m\left(K(\sqrt{\ln N} + C_N), \lambda\right) - (2+\epsilon)\ln N\right) \geq 0.$$

(iii) The following holds uniformly in $y \in \mathcal{Y}_\lambda^\pi \cap [-C'_N, C_N]$ and $\lambda \in \Lambda^\pi$:

$$\int \frac{1}{B_N}\phi\left(\frac{\tilde{y} - y}{B_N}\right)\pi(\tilde{y}|\lambda)d\tilde{y} = \left(1 + o(1)\right)\pi(y|\lambda).$$

Assumption 5.4 (Estimators of ρ and σ^2) *There exist estimators $\hat{\rho}$ and $\hat{\sigma}^2$ such that for any $\epsilon > 0$, (i) $\mathbb{E}_\theta^{\mathcal{Y}^N}\left[|\sqrt{N}(\hat{\rho} - \rho)|^4\right] \leq o(N^\epsilon)$, (ii) $\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\hat{\sigma}^4\right] \leq o(N^\epsilon)$, and (iii) $\mathbb{E}_\theta^{\mathcal{Y}^N}\left[|\sqrt{N}(\hat{\sigma}^2 - \sigma^2)|^2\right] \leq o(N^\epsilon)$.*

We factorize the correlated random effects distribution as $\pi(\lambda_i, y_{i0}) = \pi(\lambda_i)\pi(y_{i0}|\lambda_i)$ and impose regularity conditions on the marginal distribution of the heterogeneous coefficient and the conditional distribution of the initial condition. In Assumption 5.1 we let the support of $\pi(\lambda_i)$ slowly expand with the sample size by assuming that C_N grows at a subpolynomial rate. Assumption 5.2 provides an upper and a lower bound for the rate at which the bandwidth of the kernel estimator shrinks to zero. Note that for technical reasons the assumed rate is much slower than in typical density estimation problems.[6]

Assumption 5.3 imposes regularity conditions on the conditional density of the initial observation. In (i) we assume that $\pi(y_{i0}|\lambda_i)$ is bounded. In (ii) we control the tails of the distribution. In the first constraint on $m(C,\lambda)$ we essentially assume that the density of y_{i0} has exponential tails. This also guarantees that the fourth moment of Y_{i0} exists. In part

---

[6]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 Definition 3.2 by N^{ϵ_0} if the λ_i coefficients are identical for each cross-sectional unit.

(iii) we assume that $\pi(y|\lambda)$ is sufficiently smooth with respect to y such that the convolution on the left-hand side uniformly converges to $\pi(y|\lambda)$ as the bandwidth B_N tends to zero. We verify in the Appendix that a $\pi(y|\lambda)$ that satisfies Assumption 5.3 is $\pi(y|\lambda) = \phi(y - \lambda)$, where $\phi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$. Finally, Assumption 5.4 postulates the existence of finite sample moments of the estimators of the common parameter. The main result is stated in the following theorem:

Theorem 5.5 *Suppose that Assumptions 2.1, 4.1, and 5.1 to 5.4. Then, for the basic dynamic panel model the predictor $\widehat{Y}_{iT+1}$ defined in (33) satisfies the ratio optimality in Definition 3.2.*

The result in Theorem 5.5 is pointwise with respect to θ. However, the convergence of the predictor $\widehat{Y}_{iT+1}$ to the oracle predictor is uniform with respect to the unobserved heterogeneity and the observed trajectory $\mathcal{Y}_i$ in the sense that the integrated risk (conditional on θ) of the feasible predictor converges to the integrated risk of the oracle predictor. The proof of the theorem is a generalization of the proof in Brown and Greenshtein (2009), allowing for the presence of estimated parameters in the sufficient statistic $\hat{\lambda}(\cdot)$. The remarkable aspect of the results is the acceleration of the convergence (N_0^ϵ instead of N in the denominator of the standardized regret in Definition 3.2) in cases in which the intercepts are identical across units and $\pi(\lambda)$ is a pointmass.

6 Monte Carlo Simulations

We will now conduct several Monte Carlo experiments to illustrate the performance of the empirical Bayes predictor.

6.1 Experiment 1: Gaussian Random Effects Model

The first Monte Carlo experiment is based on the basic dynamic panel data model in (2). The design of the experiment is summarized in Table 1. We assume that the λ_i's are normally distributed and uncorrelated with the initial condition Y_{i0}. The innovations U_{it} and the heterogeneous intercepts λ_i have unit variances. We consider two values for the autocorrelation parameter: $\rho \in \{0.5, 0.95\}$. The panel consists of $N = 1,000$ cross-sectional units and the number of time periods is $T = 3$. Generally, the smaller T relative to number

Table 1: Monte Carlo Design 1

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0, \gamma^2)$. $\rho \in \{0.5, 0.95\}$, $\gamma = 1$
Initial Observations: $Y_{i0} \sim N(0, 1)$
Gaussian Random Effects: $\lambda_i \| Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$, $\phi_0 = 0$, $\phi_1 = 0$, $\underline{\Omega} = 1$
Sample Size: $N = 1,000$, $T = 3$
Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

of right-hand-side variables with heterogeneous coefficients, the larger the gain from using a prior distribution to compute posterior mean estimates of the λ_i's. We will compare the performance of the following predictors:

Oracle Forecast. The oracle knows the parameters $\theta = (\rho, \gamma)$ as well as the random effects distribution $\pi(\lambda_i | Y_{i0}, \xi)$, where $\xi = (\phi_0, \phi_1, \underline{\Omega})$. However, the oracle does not know the specific λ_i values. Its forecast is given by (10).

Posterior Predictive Mean Approximation Based on QMLE. The random effects distribution is correctly modeled as belonging to the family $\lambda_i | (Y_{i0}, \xi) \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$. The estimators $\hat{\theta}_{QMLE}$ and $\hat{\xi}_{QMLE}$ are defined in (24). Tweedie's formula (see (30) for the simplified version) is evaluated based on $p(\hat{\lambda}_i(\hat{\theta}_{QMLE}) | y_{i0}, \hat{\theta}_{QMLE}, \hat{\xi}_{QMLE})$.

Posterior Predictive Mean Approximation Based on GMM Estimator. We use the Arellano-Bover estimator described in Section 4.5. The estimator for ρ is given by (26) and the estimator for γ by (27). The formulas simplify considerably. We have $W_{it} = 1$, $X_{it-1} = Y_{it-1}$, $Z_{it-1} = \emptyset$ and $\alpha = \emptyset$. Moreover, $\Sigma_i^{1/2} = \gamma I$, $M_i(\gamma) = I - \iota\iota'/T$, where ι is a $T \times 1$ vector of ones. Let $\bar{\tilde{Y}}_i(\hat{\rho})$ be the temporal average of $\tilde{Y}_i(\hat{\rho})$. Then

$$\hat{\gamma}_{GMM}^2 = \frac{1}{NT}\frac{T}{T-1}\sum_{i=1}^{N} \text{tr}\left[(\tilde{Y}_i(\hat{\rho}) - \iota\bar{\tilde{Y}}_i(\hat{\rho}))(\tilde{Y}_i(\hat{\rho}) - \iota\bar{\tilde{Y}}_i(\hat{\rho}))'\right].$$

The estimator $\hat{\xi}(\hat{\theta}_{GMM})$ is obtained from (21). Finally, Tweedie's formula is evaluated based on $p(\hat{\lambda}_i(\hat{\theta}_{GMM}) | y_{i0}, \hat{\theta}_{GMM}, \hat{\xi}(\hat{\theta}_{GMM}))$.

GMM Plug-In Predictor. We use the Arellano-Bover estimator to obtain $\hat{\rho}_{GMM}$. Instead of using the posterior mean for λ_i, the plug-in predictor is based on the MLE $\hat{\lambda}_i(\hat{\rho}_{GMM})$. The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_i(\hat{\rho}_{GMM}) + \hat{\rho}_{GMM}Y_{iT}$.

Loss-Function-Based Predictor. We construct an estimator of (ρ, λ^N) based on the

objective function:

$$\hat{\rho}_L = \text{argmin}_\rho \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \rho Y_{it-1} - \hat{\lambda}_i(\rho)\right)^2, \quad \hat{\lambda}_i(\rho) = \frac{1}{T} \sum_{t=1}^{T} Y_{it} - \rho Y_{it-1}. \quad (34)$$

This estimator minimizes the loss function under which the forecasts are evaluated in sample. It is well-known that due to the incidental parameter problem, the estimator $\hat{\rho}_L$ is inconsistent under fixed-N asymptotics. The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_i(\hat{\rho}_L) + \hat{\rho}_L Y_{iT}$.

Pooled-OLS Predictor. Ignoring the heterogeneity in the λ_i's and imposing that $\lambda_i = \lambda$ for all i, we can define

$$(\hat{\rho}_P, \hat{\lambda}_P) = \text{argmin}_{\rho,\lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \rho Y_{it-1} - \lambda\right)^2. \quad (35)$$

The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_P + \hat{\rho}_P Y_{iT}$.

First-Difference Predictor. In the panel data literature it is common to difference-out idiosyncratic intercepts, which suggests to predict ΔY_{iT+1} based on ΔY_{iT}. We evaluate the first-difference predictor at the Arellano-Bover GMM estimator of ρ to obtain $\widehat{Y}_{iT+1}^{FD}(\hat{\rho}_{GMM})$.

In Table 2 we report the regret associated with each predictor relative to the posterior variance of λ_i, averaged over all trajectories $\mathcal{Y}^N$, as specified in Definition 3.2 (setting $N^\epsilon = 1$). For the oracle predictor the regret is by definition zero and we tabulate the risk R_N^{opt} instead (in parentheses). We also report the median forecast error $\hat{e}_{iT+1|T} = Y_{iT+1} - \widehat{Y}_{iT+1}$ to highlight biases in the forecasts.

The columns titled "All Units" correspond to $D_i(\mathcal{Y}^N) = 1$. As expected from the theoretical analysis, the posterior mean predictors have the lowest regret among the feasible predictors. The density of $\hat{\lambda}_i$ is estimated parametrically, using a family of distributions that nests the true random effects distribution. Because it is based on a correctly specified likelihood function, the predictor based on $\hat{\theta}_{QMLE}$ performs slightly better than the predictor based on $\hat{\theta}_{GMM}$. Consider $\rho = 0.5$: for the QMLE-based predictor the regret is 0.5% of the average posterior variance, whereas it is 3% for the GMM-based predictor. The plug-in predictor that replaces the unknown λ_i's by the sufficient statistic $\hat{\lambda}_i$ (which is also the maximum likelihood estimator) instead of the posterior mean is associated with a much larger relative regret, which is about 37%.

The remaining three predictors are also strictly dominated by the posterior mean predictors. Ignoring the serial correlation in ΔY_{it}, the first-difference predictor performs the

Table 2: Monte Carlo Experiment 1: Random Effects, Parametric Tweedie Correction, Selection Bias

Estimator / Predictor	All Units		Bottom Group		Middle Group		Top Group	
	Regret	Median Forec.E.	Regret	Median Forec.E.	Regret	Median Forec.E	Regret	Median Forec.E.
Low Persistence: $\rho = 0.50$								
Oracle Predictor	(1252.7)	0.002	(65.95)	-0.037	(62.48)	0.003	(62.10)	-0.003
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.005	0.005	0.002	-0.030	0.002	0.006	0.018	-0.004
Post. Mean ($\hat{\theta}_{GMM}$, Parametric)	0.030	0.004	0.015	-0.035	0.022	0.008	0.100	0.004
Plug-In Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	0.358	0.005	1.150	0.536	0.045	0.009	1.421	-0.558
Loss-Function-Based Estimator	0.369	0.199	0.275	0.190	0.348	0.197	0.352	0.188
Pooled OLS	0.656	-0.285	1.892	-0.663	0.491	-0.288	0.223	0.044
First-Difference Predictor ($\hat{\theta}_{GMM}$)	2.963	0.001	5.317	0.935	1.936	0.009	5.656	-0.986
High Persistence: $\rho = 0.95$								
Oracle Predictor	(1252.7)	0.002	(67.36)	-0.081	(63.16)	0.007	(61.86)	-0.002
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.009	0.011	0.003	-0.075	0.005	0.016	0.036	0.015
Post. Mean ($\hat{\theta}_{GMM}$, Parametric)	0.046	0.003	0.019	-0.071	0.023	0.010	0.178	-0.005
Plug-In Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	0.380	0.004	1.036	0.498	0.039	0.017	1.546	-0.569
Loss-Function-Based Estimator	0.623	0.357	0.014	0.033	0.522	0.357	1.358	0.597
Pooled OLS	1.015	-0.454	1.066	-0.517	0.967	-0.459	0.872	-0.422
First-Difference Predictor ($\hat{\theta}_{GMM}$)	3.986	0.000	6.582	0.887	2.733	0.013	6.912	-0.939

Notes: The design of the experiment is summarized in Table 1. For the oracle predictor we report the compound risk (in parentheses) instead of the regret. The regret is standardized by the average posterior variance of λ_i, see Definition 3.2.

Figure 1: QMLE Estimation: Distribution of $\widehat{\mathbb{E}}^{\lambda_i}_{\hat\theta, \mathcal{Y}_i}[\lambda_i]$ versus $\hat\lambda_i(\hat\theta)$
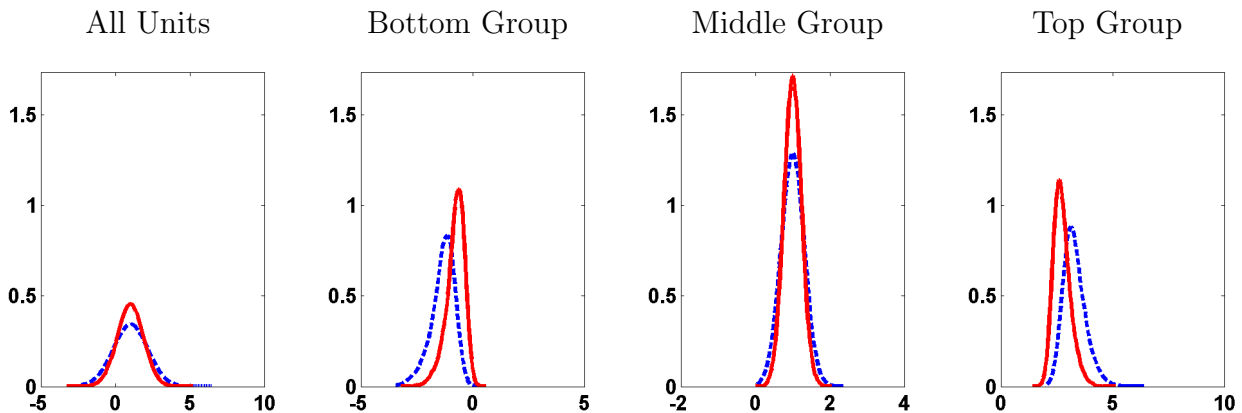
Notes: Solid (red) lines depict cross-sectional densities of posterior mean estimates $\widehat{\mathbb{E}}^{\lambda_i}_{\hat\theta, \mathcal{Y}_i}[\lambda_i]$. Dashed (blue) lines depict cross-sectional densities of sufficient statistic $\hat\lambda_i(\hat\theta)$. The results are based on the QMLE estimator. The Monte Carlo design is described in Table 1.

worst for both choices of ρ. The second-to-worst predictor is the pooled-OLS predictor which ignores the cross-sectional heterogeneity in the λ_i's. A reduction of the variance $\underline{\Omega}$ of the heterogeneous intercepts would improve the relative performance of the pooled-OLS predictor. Finally, the loss-function-based predictor dominates the pooled-OLS and the first difference predictor. As mentioned above, while conceptually appealing, the loss-function-based predictor relies on an inconsistent estimate of ρ, which in comparison to the GMM plug-in predictor is unappealing if the cross-sectional dimension N is very large.

Across all units, the predictions under the loss-function-based estimator and the pooled-OLS estimator appear to be biased. To study this bias further we now consider level-based selection rules $D_i(\mathcal{Y}^i)$. Using the 5%, 47.5%, 52.5%, and 95% quantiles of the population distribution of Y_{iT}, we define cut-offs for a bottom 5% group, a middle 5% group, and a top 5% group. Because the cut-offs are computed from the population distribution of Y_{iT}, for unit i the selection rules only depends on $\mathcal{Y}_{iT}$ and not on Y_{jT} with $j \neq i$.

For the top and bottom groups only the posterior mean predictors lead to unbiased forecast errors. The sufficient statistic $\hat\lambda_i$ tends to overestimate (underestimate) λ_i for the top (bottom) group, because it interprets a sequence of above-average (below-average) U_{iT}'s as evidence for a high (low) λ_i. This is reflected in the bias: the plug-in predictors' forecast errors for the top group are on average positive, whereas the forecast errors for the bottom group tend to be negative. The posterior mean tends to correct these biases because it

Table 3: Monte Carlo Design 2

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0,\gamma^2)$; $\rho = 0.5$, $\gamma = 1$
Initial Observation: $Y_{i0} \sim N\left(\frac{\mu_\lambda}{1-\rho}, V_Y + \frac{V_\lambda}{(1-\rho)^2}\right)$, $V_Y = \gamma^2/(1-\rho^2)$; $\underline{\mu}_\lambda = 1$, $\underline{V}_\lambda = 1$
Non-Gaussian Correlated Random Effects:

$$\lambda_i|Y_{i0} \sim \begin{cases} N\big(\phi_+(Y_{i0}), \underline{\Omega}\big) & \text{with probability } p_\lambda \\ N\big(\phi_-(Y_{i0}), \underline{\Omega}\big) & \text{with probability } 1 - p_\lambda \end{cases},$$

$$\phi_+(Y_{i0}) = \phi_0 + \delta + (\phi_1 + \delta)Y_{i0},$$
$$\phi_-(Y_{i0}) = \phi_0 - \delta + (\phi_1 - \delta)Y_{i0},$$
$$\underline{\Omega} = \left[\frac{1}{(1-\rho)^2}V_Y^{-1} + \underline{V}_\lambda^{-1}\right]^{-1}, \ \phi_0 = \underline{\Omega}\,\underline{V}_\lambda^{-1}\underline{\mu}_\lambda, \ \phi_1 = \frac{1}{1-\rho}\underline{\Omega}V_Y^{-1},$$
$$p_\lambda = 1/2, \ \underline{\Omega} = 1, \ \delta \in \{1/5,\, 1,\, 5\} \ (\delta = 1/\sqrt{\kappa})$$

Sample Size: $N = 1,000$, $T = 3$
Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

shrinks toward the mean of the prior distribution of the λ_i's. This reduces the regrets for the top and bottom groups, and is also reflected in the risk calculated across all units. The bias correction is illustrated in Figure 1, which compares the cross-sectional distribution of the sufficient statistics $\hat{\lambda}_i(\hat{\theta})$ to the distribution of the posterior mean estimates $\widehat{\mathbb{E}}_{\hat{\theta},\mathcal{Y}_i}^{\lambda_i}[\lambda_i]$ obtained with Tweedie's formula. Due to the shrinkage effect of the prior, the distribution of the posterior means, in particular for the top and bottom groups, is more compressed.

6.2   Experiment 2: Non-Gaussian Correlated Random Effects Model

We now change the Monte Carlo design in two dimensions. First, we replace the Gaussian random effects specification with a non-Gaussian specification in which the heterogeneous coefficient λ_i is correlated with the initial condition Y_{i0}. Second, we consider a Tweedie correction based on a kernel density estimate of $p(\hat{\lambda}_i|Y_{i0})$ as discussed in Section 4.3.

The Monte Carlo design is summarized in Table 3. Starting point is a joint normal distribution for (λ_i, Y_{i0}), factorized into a marginal distribution $\pi_*(\lambda_i)$ and a conditional distribution $\pi_*(Y_{i0}|\lambda_i)$. We assumed $\lambda_i \sim N(\underline{\mu}_\lambda, \underline{V}_\lambda)$ and that $Y_{i0}|\lambda_i$ corresponds to the stationary distribution of Y_{it} associated with its autoregressive law of motion. The implied marginal distribution for Y_{i0} is used as $\pi(Y_{i0})$ in the Monte Carlo design. To obtain $\pi(\lambda_i|Y_{i0})$ we took $\pi_*(\lambda_i|Y_{i0})$ from the Gaussian model and replaced it with a mixture of normals described in Table 3. For $\delta = 0$ the mixture reduces to $\pi_*(\lambda_i|Y_{i0})$, whereas for large values of δ it becomes bimodal. This bimodality also translates into the distribution of $\hat{\lambda}|Y_{i0}$, which is depicted in Figure 2 for $\delta = 1/10$ (almost Gaussian) and $\delta = 1$ (bimodal).

Figure 2: QMLE Estimation: Density $p(\hat{\lambda}_i|y_{i0}, \theta)$ for $\delta = 1/10$ versus $\delta = 1$
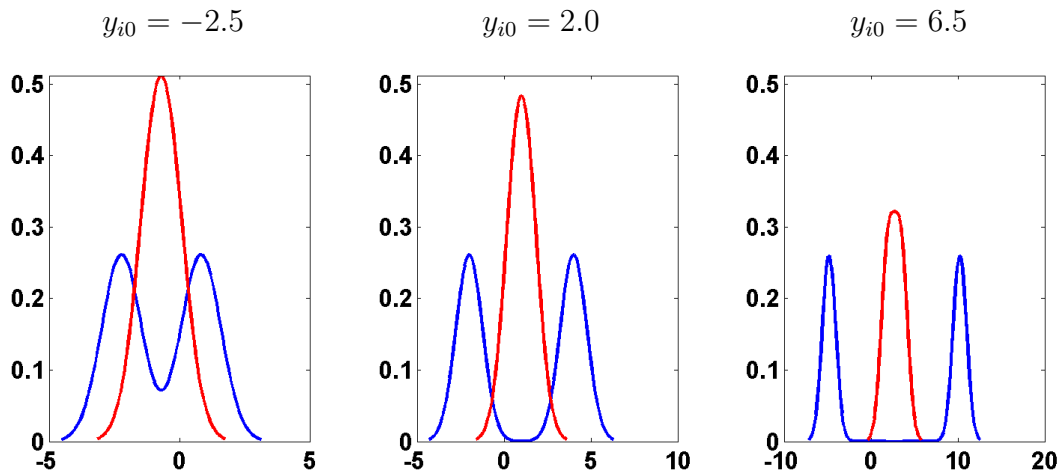
Notes: Solid (blue) line is $\delta = 1$ and solid (red) line is $\delta = 1/10$. The Monte Carlo design is described in Table 3.
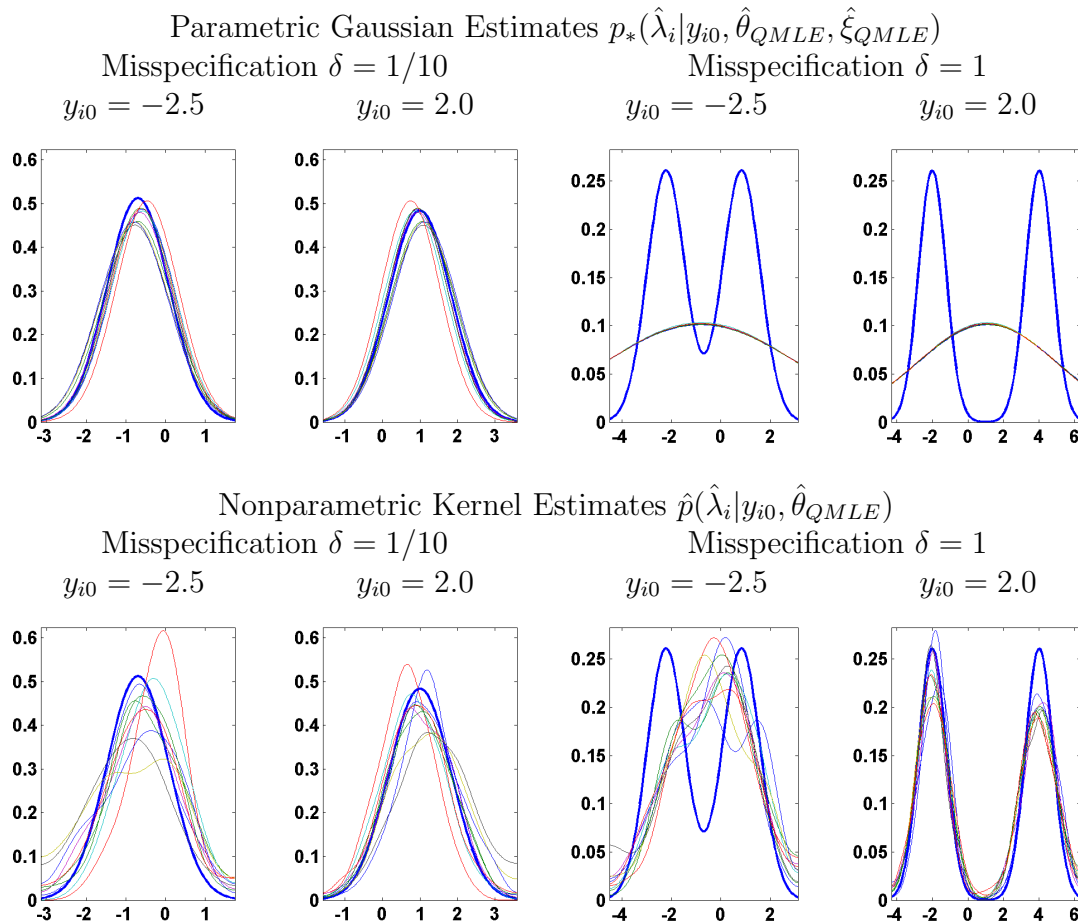
In this experiment we consider a parametric Tweedie correction (same as in Experiment 1, but now misspecified in view of the DGP) and two nonparametric Tweedie corrections. First, we compute the correction based on the simple Gaussian kernel in (22). The bandwidth is chosen in accordance with the theory in Section 5. We set $B_N = c/(\ln N)^{0.55}$, which would be consistent with a truncation of the form $C_N = c\sqrt{\ln N}$, and let $c \in \{1/2, 1, 2\}$.[7] Second, we use the adaptive estimator proposed by Botev, Grotowski, and Kroese (2010), henceforth BGK estimator, which is based on the solution of a diffusion partial differential equation. This estimator is associated with a plug-in bandwidth selection rule that requires no further tuning.[8] Unless otherwise noted, the subsequent results are based on the BGK estimator.

Figure 3 shows the "true" density $p(\hat{\lambda}_i|y_{i0}, \theta)$ as well as Gaussian and nonparametric approximations. Under the Gaussian correlated random effects distribution we can directly calculate the conditional distribution of $\hat{\lambda}_i$ given y_{i0}. The nonparametric approximation is obtained by dividing an estimate of the joint density of $(\hat{\lambda}_i, y_{i0})$ by an estimate of the marginal density of y_{i0} (this normalization is not required for the Tweedie correction). Each hairline in Figure 3 corresponds to a density estimate from a different Monte Carlo run. For $\delta = 1/10$ the Gaussian approximation is accurate and the variability of the estimates is much smaller than that of the kernel estimates. For $\delta = 1$ the Gaussian density is unable

---

[7]The tuning matrices $V_{\hat{\lambda}}$ and V_h are set equal to the sample variances of $\hat{\lambda}_i$ and y_{i0}, respectively.

[8]Our estimates are based on Algorithms 1 and 2 in BGK. We use the authors' MATLAB code to implement the density estimator.

Figure 3: QMLE Estimation: "True" Density $p(\hat{\lambda}_i|y_{i0}, \theta)$ versus Gaussian and Nonparametric Estimates
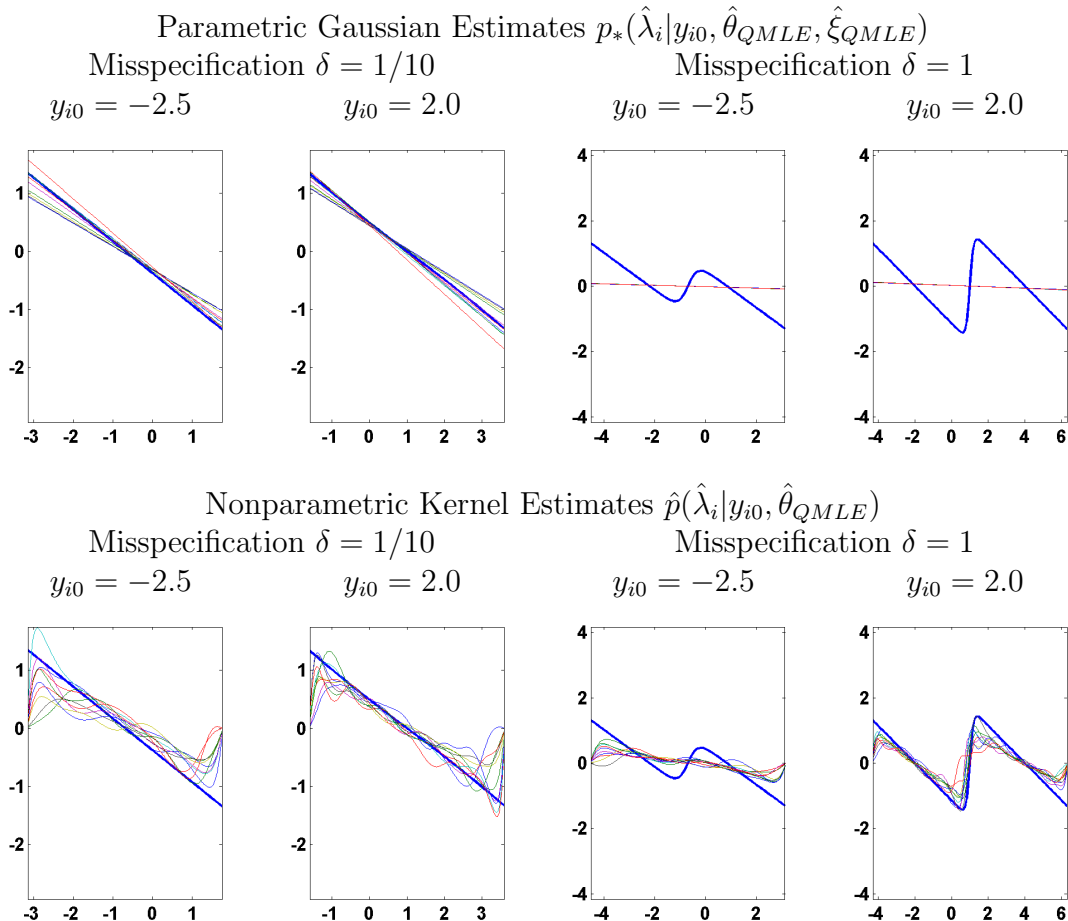
Parametric Gaussian Estimates $p_*(\hat{\lambda}_i|y_{i0}, \hat{\theta}_{QMLE}, \hat{\xi}_{QMLE})$

Misspecification $\delta = 1/10$		Misspecification $\delta = 1$	
$y_{i0} = -2.5$	$y_{i0} = 2.0$	$y_{i0} = -2.5$	$y_{i0} = 2.0$

Nonparametric Kernel Estimates $\hat{p}(\hat{\lambda}_i|y_{i0}, \hat{\theta}_{QMLE})$

Misspecification $\delta = 1/10$		Misspecification $\delta = 1$	
$y_{i0} = -2.5$	$y_{i0} = 2.0$	$y_{i0} = -2.5$	$y_{i0} = 2.0$

Notes: Solid (blue) lines depict "true" $p(\hat{\lambda}_i|y_{i0}, \theta)$. Colored "hairs" depict 10 estimates from the Monte Carlo repetitions. The nonparametric estimates are based on the BGK kernel estimator. The Monte Carlo design is described in Table 3.

to approximate the bimodal $p(\hat{\lambda}_i, y_{i0}|\theta)$, whereas the non-parametric approximation, at least for $y_{i0} = 2.0$ captures the key features of the density of $\hat{\lambda}_i$.

For the prediction, the relevant object is the correction $(\sigma^2/T)\partial \ln p(\hat{\lambda}_i, y_{i0}|\theta)/\partial\hat{\lambda}_i$, which is depicted in Figure 4. Under a Gaussian correlated random effects distribution, the Tweedie correction is linear in $\hat{\lambda}_i$ because the posterior mean is a linear combination of the prior mean and the maximum of the likelihood function. Thus, the corrections based on the Gaussian density estimate are linear regardless of δ. For $\delta = 1/10$ the correction under the "true" random effects distribution is nearly linear, and thus well approximated by the Gaussian correction. The nonparametric correction is fairly accurate for values of $\hat{\lambda}$ in the center of

Figure 4: QMLE Estimation: Gaussian versus Nonparametric Estimates Tweedie Correction

Parametric Gaussian Estimates $p_*(\hat{\lambda}_i|y_{i0},\hat{\theta}_{QMLE},\hat{\xi}_{QMLE})$

Nonparametric Kernel Estimates $\hat{p}(\hat{\lambda}_i|y_{i0},\hat{\theta}_{QMLE})$

Notes: Solid (blue) lines depict Tweedie correction based on $p(\hat{\lambda}_i|y_{i0},\theta)$. Colored "hairs" depict 10 estimates from the Monte Carlo repetitions. The nonparametric estimates are based on the BGK kernel estimator. The Monte Carlo design is described in Table 3.

the conditional distribution $\hat{\lambda}_i|(y_{i0},\theta)$, but it becomes less accurate in the tails. For $\delta = 1$, on the other hand, the kernel-based correction provides a much better approximation of the optimal correction than the Gaussian correction.

Table 4 compares the performance of twelve predictors; half of them based on QMLE and the other half based on GMM. It is well-known that the GMM estimator of θ is consistent under the DGP described in Table 3. We show in the Appendix that the QMLE estimator is also consistent for θ under this DGP, despite the fact that the correlated random effects distribution is misspecified. For each of the two θ estimators we construct posterior mean predictors using four different nonparametric Tweedie corrections as well as the Gaussian Tweedie correction. Moreover, we compute the plug-in predictor based on $\hat{\lambda}_i(\hat{\theta})$.

Table 4: Monte Carlo Experiment 2: Correlated Random Effects, Non-parametric versus Parametric Tweedie Correction

Estimator / Predictor	All Units		Bottom Group		Top Group	
	Regret	Median Forec.E.	Regret	Median Forec.E.	Regret	Median Forec.E
	$\delta = 1/10$					
Oracle Predictor	(1177.6)	0.003	(54.92)	-0.046	(63.97)	-0.010
Post. Mean ($\hat{\theta}_{QMLE}$, BGK Kernel)	0.179	-0.001	0.737	0.159	0.543	-0.119
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 0.5$)	0.635	0.001	1.711	0.438	1.157	-0.360
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 1.0$)	0.454	0.000	1.126	0.345	0.779	-0.279
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 2.0$)	0.416	0.000	0.826	0.267	0.568	-0.183
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.048	0.001	0.053	0.060	0.130	0.127
Plug-in Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.915	0.001	2.323	0.527	1.549	-0.437
Post. Mean ($\hat{\theta}_{GMM}$, BGK Kernel)	0.217	0.002	0.766	0.135	0.566	-0.095
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 0.5$)	0.693	0.002	1.761	0.423	1.182	-0.336
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 1.0$)	0.509	0.001	1.180	0.333	0.813	-0.255
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 2.0$)	0.459	0.002	0.866	0.252	0.601	-0.160
Post. Mean ($\hat{\theta}_{GMM}$, Parametric)	0.091	0.002	0.079	0.043	0.192	0.146
Plug-in Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	0.968	0.003	2.356	0.511	1.558	-0.413
	$\delta = 1$					
Oracle Predictor	(1161.7)	-0.003	(54.43)	-0.056	(65.78)	-0.024
Post. Mean ($\hat{\theta}_{QMLE}$, BGK Kernel)	0.298	0.006	0.756	0.181	0.735	-0.073
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 0.5$)	0.526	0.001	0.857	0.240	0.855	-0.089
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 1.0$)	0.661	0.002	0.894	0.226	0.936	-0.050
Post. Mean ($\hat{\theta}_{QMLE}$, Gaussian Kernel $c = 2.0$)	0.833	0.005	1.080	0.225	1.100	0.000
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	1.025	0.001	1.292	0.233	1.256	-0.012
Plug-in Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	1.068	0.001	1.852	0.388	1.468	-0.158
Post. Mean ($\hat{\theta}_{GMM}$, BGK Kernel)	0.343	0.006	0.906	0.171	0.874	-0.068
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 0.5$)	0.571	0.001	1.015	0.234	0.994	-0.086
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 1.0$)	0.706	0.002	1.050	0.217	1.076	-0.046
Post. Mean ($\hat{\theta}_{GMM}$, Gaussian Kernel $c = 2.0$)	0.930	0.005	1.235	0.218	1.242	0.006
Post. Mean ($\hat{\theta}_{GMM}$, Parametric)	1.071	0.001	1.443	0.228	1.392	-0.005
Plug-in Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	1.115	0.001	2.011	0.383	1.609	-0.154

Notes: The design of the experiment is summarized in Table 3. For the oracle predictor we report the compound risk (in parentheses) instead of the regret. The regret is standardized by the average posterior variance of λ_i, see Definition 3.2. The BGK estimator relies on a adaptive bandwidth choice. For the Gaussian kernel estimator in (22) we set $B_N = c/(\ln N)^{0.49}$.

Among the nonparametric predictors, the one based on the BGK density estimator clearly dominates the ones derived from the simple kernel density estimator. If the random effects distribution is almost normal, i.e., $\delta = 1/10$, setting $c = 2$ is preferable to the other choices of c. For the bimodal random effects distribution, i.e., $\delta = 1$, the best performance of the simple kernel estimator is attained for $c = 1/2$. The predictors that rely on posterior mean approximations generally outperform the naive predictors based on $\hat{\lambda}_i(\hat{\theta})$. The benefits from shrinkage are most pronounced for the bottom and top groups. If the misspecification
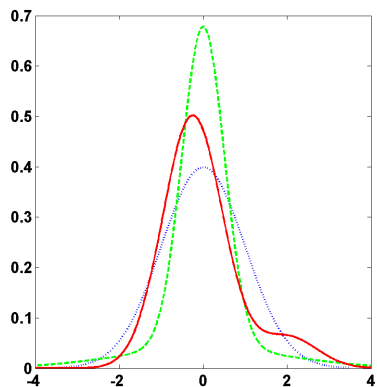
Table 5: Monte Carlo Design 3

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$, $\rho = 0.5$, $\mathbb{E}[U_{it}] = 0$, $\mathbb{V}[U_{it}] = 1$

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

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

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

$\mu_- = 1/4$, $\mu_+ = 2$, $p_u = \mu_u^-/(\mu_u^- + \mu_u^+) = 1/9$,

$\gamma^2 = 1 - p_u(\mu_u^+)^2 - (1 - p_u)(\mu_u^-)^2 = 1/2$

Initial Observations: $Y_{i0} \sim N(0, 1)$

Gaussian Random Effects: $\lambda_i | Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$, $\phi_0 = 0$, $\phi_1 = 0$, $\underline{\Omega} = 1$

Sample Size: $N = 1,000$, $T = 3$

Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

The plot overlays a $N(0, 1)$ density (blue, dotted), the scale mixture (green, dashed), and the location mixture (red, solid).

is small ($\delta = 1/10$), the parametric correction leads to more precise forecasts than the nonparametric correction because it is based on a more efficient density estimator. As the degree of misspecification increases, the nonparametric correction starts to perform better and for $\delta = 1$ it clearly dominates the parametric competitor. This is consistent with the accuracy of the underlying density estimators shown in Figures 3 and 4.

6.3 Experiment 3: Misspecified Likelihood Function

In the third experiment, summarized in Table 5, we consider a misspecification of the Gaussian likelihood function by replacing the Normal distribution in the DGP with two mixtures. 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

Table 6: Monte Carlo Experiment 3: Misspecified Likelihood Function

| Estimator / Predictor | All Units | | Bottom Group | | Top Group | |
	Regret	Median Forec.E.	Regret	Median Forec.E	Regret	Median Forec.E.
Scale Mixture – Excess Kurtosis						
Oracle Predictor	(1153.7)	0.000	(67.98)	0.002	(55.99)	-0.033
Post. Mean ($\hat{\theta}_{QMLE}$, BGK Kernel)	0.977	-0.002	2.031	0.170	2.226	-0.227
Post. Mean ($\hat{\theta}_{GMM}$, BGK Kernel)	1.033	-0.000	2.055	0.162	2.388	-0.211
Plug-In Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	1.605	0.002	3.666	0.555	4.396	-0.642
Loss-Function-Based Estimator	1.615	0.197	1.423	0.206	1.198	0.146
Pooled OLS	2.244	-0.286	4.295	-0.644	2.516	-0.020
Location Mixture – Skewness						
Oracle Predictor	(1200.2)	-0.146	(63.29)	-0.167	(62.31)	-0.162
Post. Mean ($\hat{\theta}_{QMLE}$, BGK Kernel)	0.359	-0.106	0.338	-0.077	0.962	-0.410
Post. Mean ($\hat{\theta}_{GMM}$, BGK Kernel)	0.398	-0.105	0.362	-0.080	1.086	-0.399
Plug-In Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	0.810	-0.091	1.359	0.330	2.784	-0.818
Loss-Function-Based Estimator	0.807	0.099	0.461	0.030	0.497	-0.006
Pooled OLS	1.240	-0.391	3.902	-0.889	0.828	-0.235

Notes: The design of the experiment is summarized in Table 5. For the oracle predictor we report the compound risk (in parentheses) instead of the regret. The regret is standardized by the average posterior variance of λ_i, see Definition 3.2.

$\mathbb{V}[U_{it}] = 1$. For the heterogeneous intercepts λ_i we adopt the Gaussian random effects specification of Experiment 1. In this experiment we compute the relative regret for five predictors:[9] the posterior mean predictor based on the non-parametric Tweedie correction and the plug-in predictor based on $\hat{\theta}_{QMLE}$ and $\hat{\theta}_{MLE}$, respectively. Note that both the QMLE and the GMM estimator of θ remain consistent under the likelihood misspecification. However, the (non-parametric) Tweedie correction no longer delivers a valid approximation of the posterior mean.

The results are summarized in Table 6. The risk of the oracle predictors can be compared to that reported in Table 1. The excess kurtosis of the scale mixture and the skewness of the location mixture slightly reduce the posterior variance of λ compared to the standard normal benchmark in Experiment 1. Due to the misspecification of the likelihood function, the relative regret of the various predictors increases considerably, but the relative ranking is essentially unchanged. The posterior mean predictors based on the nonparametric Tweedie correction dominate all the other predictor, attaining a relative regrets of about 1 and 0.4, respectively. Compared to the plug-in and loss-function based predictors, the

---

[9]The computation of the oracle predictor and the normalization of the regret by the posterior variance of λ require a Gibbs sampler which is described in the Appendix.

Tweedie correction still reduces the regret 40% to 50%. The predictor based on the pooled OLS estimation performs the worst among the five predictors in this experiment.

7 Empirical Application

We will now use the previously-developed predictors to forecast pre-provision net revenues (PPNR) of bank holding companies (BHC). The stress tests that have become mandatory under the 2010 Dodd-Frank Act require banks to establish how PPNR varies in stressed macroeconomic and financial scenarios. A first step toward building and estimating models that provide trustworthy projections of PPNR 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 seems particularly attractive for stress-test applications.

We generate a collection of panel data sets in which pre-provision net revenue as a fraction of consolidated assets (the ratio is scaled by 400 to obtain annualized percentages) is the key dependent variable. The data sets are based on the FR Y-9C consolidated financial statements for bank holding companies for the years 2002 to 2014, which are available through the website of the Federal Reserve Bank of Chicago. Because the balance sheet data exhibit strong seasonal features, we time-aggregate the quarterly observations into annual observations and take the time period t to be one year.

We construct rolling samples that consist of $T + 2$ observations, where T is the size of the estimation sample and varies between $T = 3$ and $T = 11$ years. 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. For instance, with data from 2002 to 2014 we can construct $M = 9$ samples of size $T = 3$ with forecast origins running from $\tau = 2005$ to $\tau = 2013$. Each rolling sample is indexed by the pair (τ, T). The cross-sectional dimension N varies from sample to sample and ranges from approximately $= 460$ to 725. Further details about the data as well as a description of our procedure to create balanced panels and eliminate outliers are provided in the Appendix.

In Section 7.1 we use the basic dynamic panel data model to generate PPNR forecasts. In Section 7.2 we extend the model to include covariates and compare forecasts under the

Table 7: MSE for Basic Dynamic Panel Model

	Rolling Samples				
	$T = 3$	$T = 5$	$T = 7$	$T = 9$	$T = 11$
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.74	0.69	0.58	0.48	0.45
Post. Mean ($\hat{\theta}_{QMLE}$, BGK Kernel)	0.84	0.74	0.59	0.50	0.46
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.90	0.79	0.60	0.51	0.48
Post. Mean ($\hat{\theta}_{GMM}$, Parametric)	1.08	0.83	0.60	0.49	0.43
Post. Mean ($\hat{\theta}_{GMM}$, BGK Kernel)	1.16	0.93	0.61	0.50	0.44
Plug-In Predictor ($\hat{\theta}_{GMM}$, $\hat{\lambda}_i(\hat{\theta}_{GMM})$)	1.17	0.89	0.61	0.51	0.46
Loss-Function-Based Estimator	0.91	0.84	0.63	0.53	0.42
Pooled OLS	0.71	0.68	0.57	0.48	0.45

Notes: The MSEs are computed across the different forecast origins τ associated with each sample size T.

actual realization of the covariates and stressed scenarios in which we set the covariantes to counterfactual levels.

7.1 Results from the Basic Dynamic Panel Model

We begin by evaluating forecasts from the basic dynamic panel model in (28). The parametric Tweedie correction is based on $\lambda_i|(H_i, \theta) \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\omega}^2)$. The forecast evaluation criterion is the mean-squared error (MSE) computed across institutions and across time:

$$MSE = \frac{1}{M} \sum_{\tau=\tau_1}^{\tau_1+M-1} \left(\frac{\frac{1}{N_\tau} \sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau})(Y_{i\tau+1} - \widehat{Y}_{i\tau+1})^2}{\frac{1}{N_\tau} \sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau})} \right), \tag{36}$$

where M is the number of rolling samples. Table 7 summarizes the MSEs for different estimators and different sizes T of the estimation samples. Recall that the unit of $\widehat{Y}_{i\tau}$ is annual revenue as fraction of total assets converted into annualized percentages.

For the short samples, i.e., $T = 3$ and $T = 5$, the QMLE-based predictors are more accurate than the GMM-based predictors. This discrepancy vanishes as the sample size is increased to $T = 11$. The posterior mean predictors computed with the Tweedie correction are more accurate than the plug-in predictors. As expected, the MSE differential is largest in the small T samples, because the unit-specific likelihood function contains fairly little information and the prior strongly influences the posterior. The parametric Tweedie correction delivers more accurate predictions than the non-parametric Tweedie correction,

Figure 5: Tweedie Corrections for $T = 5$ and $\tau = 2012$
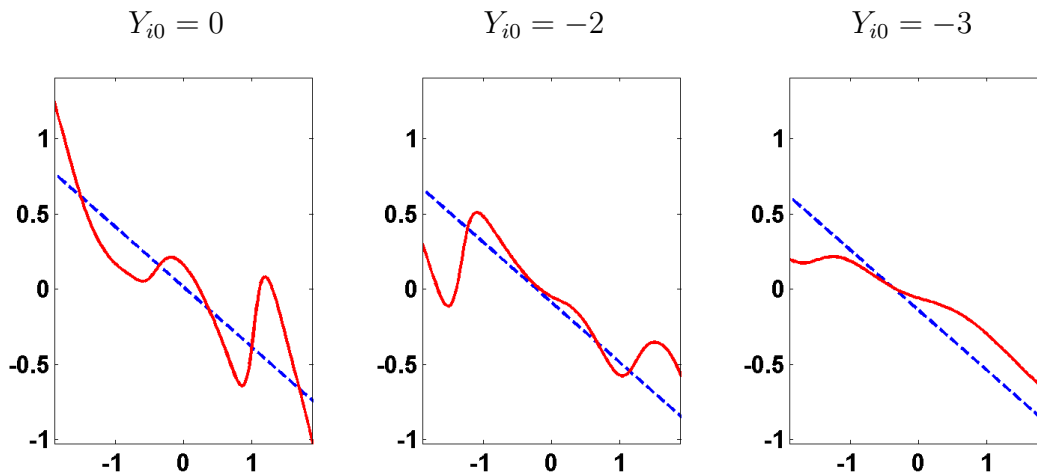
Notes: Each panel shows the parametric (dashed blue) and the non-parametric (solid red) Tweedie correction for $\hat{\theta}_{QMLE}$.

in particular for small T. In Figure 5 we compare the Tweedie corrections for $T = 5$ and $\tau = 2012$. While the corrections are quite similar for values of the sufficient statistic $\hat{\lambda}_i(\rho) = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \rho Y_{it-1})$ between -1% and 1%, the non-parametric correction behaves somewhat erratic outside of this interval which hurts the predictive performance.

Returning to the MSE results in Table 7, the posterior mean predictor yields roughly the same MSE as pooled OLS. This suggests that *a posteriori* the data sets contain only weak evidence for heterogeneous intercepts. In this regard, the parametric specification is more efficient in shrinking the intercept estimates toward a common value. Finally, for all sample sizes except $T = 11$, the posterior-mean predictor based on $\hat{\theta}_{QMLE}$ and the parametric Tweedie correction is more accurate than the loss-function-based predictor.

In Table 8 we focus on the sample size $T = 5$. In addition to averaging forecast errors across all $T = 5$ samples, we also report results for specific forecast origins, namely choices of τ that correspond to the years 2007, the onset of the Great Recession, and 2012, which is during the recovery period. Moreover, we compute MSEs based on cross-sectional selection rules that depend on the level of PPNR at the forecast origin τ. We focus on institutions with PPNR less than 0%, -1%, -2%, and -3%, respectively. Because the QMLE predictors dominate the GMM predictors and the parametric Tweedie correction was preferable to the nonparametric correction, we now restrict our attention to the posterior-mean predictor based on $\hat{\theta}_{QMLE}$ and the parametric Tweedie correction, the $\hat{\theta}_{QMLE}$ plug-in predictor, and

Table 8: MSE for Basic Dynamic Panel Model for $T = 5$

| | Selection $D_i(\mathcal{Y}_{i\tau})$ | | | | |
	All	$y_{i\tau} \leq 0$	$y_{i\tau} \leq -1$	$y_{i\tau} \leq -2$	$y_{i\tau} \leq -3$
Rolling Sample $\tau = 2007$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.90	0.90	1.04	1.29	1.72
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	1.26	1.21	1.39	1.65	2.08
Loss-Function-Based Estimator	1.17	1.17	1.54	2.31	1.99
Pooled OLS	0.91	0.91	1.04	1.28	1.71
Rolling Sample $\tau = 2012$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.51	0.56	0.83	0.91	1.01
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.55	0.51	0.75	0.85	1.05
Loss-Function-Based Estimator	0.63	0.69	0.98	1.02	1.00
Pooled OLS	0.48	0.57	0.85	0.97	1.12
All Rolling Samples $\tau = 2007, \ldots, 2013$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.69	0.88	1.12	1.43	1.69
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.79	1.00	1.32	1.72	2.16
Loss-Function-Based Estimator	0.84	1.00	1.24	1.54	1.63
Pooled OLS	0.71	0.90	1.16	1.50	1.80

Notes: For the last panel (all rolling samples) the MSEs are computed across the different forecast origins τ.

predictors constructed from loss-function-based estimates and pooled OLS, respectively.

For the 2007 sample, the plug-in and the loss-function-based predictor are dominated by the other two predictors. The performance of the posterior-mean and the pooled-OLS predictor are essentially identical. For the 2012 sample, the posterior-mean predictor performs better than the plug-in predictor if we average across all institutions or if we condition on BCHs with PPNR of less than -3%. In the other cases the ranking is reversed. Across all rolling samples, the posterior mean predictor dominates. Across all institutions its performance is only slightly better than pooled OLS, but if we condition on BCHs with PPNR of less than -1%, -2%, or -3% then the accuracy relative to pooled OLS is more pronounced.

Table A-3 in the Appendix provides point estimates of the parameters of the basic dynamic panel model and the parametric correlated random effects distribution for $T = 5$ and $\tau = 2007, \ldots, 2013$. Until 2010 the estimated variance of the correlated random effects distribution is essentially zero, which implies that $\lambda_i \approx \phi_0 + \phi_1 Y_{i0}$. Because of a non-zero $\hat{\phi}_1$ the resulting predictor is not exactly pooled OLS but it is very similar as we have seen from the results in Table 8. Starting in 2011, we obtain non-trivial estimates of $\hat{\underline{\omega}}^2$ which imply non-trival *a priori* dispersion of the intercepts (that is not due to the dispersion in initial conditions). Overall, the estimates $\hat{\underline{\omega}}^2$ imply a large degree of shrinkage. The positive

Table 9: Parameter Estimates for $T = 5$: $\hat{\theta}_{QMLE}$, Parametric Tweedie Correction

τ	$\hat{\rho}$	$\hat{\sigma}^2$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\underline{\omega}}^2$	N
2007	0.90	0.61	0.03	0.01	6E-8	537
2008	0.83	0.55	0.11	0.05	2E-8	598
2009	0.76	0.76	0.01	0.10	4E-8	613
2010	0.80	0.67	-0.05	0.09	2E-7	606
2011	0.79	0.58	-0.02	0.07	0.07	582
2012	0.71	0.53	0.04	0.13	0.16	587
2013	0.79	0.58	-0.05	0.12	0.09	608

Notes: Point estimates for the model $Y_{it+1} = \lambda_i + \rho Y_{it} + U_{it+1}$, $U_{it+1} \sim N(0, \sigma^2)$, $\lambda_i | Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\omega}^2)$.

estimate $\hat{\phi}_1$ generates positive correlation between λ_i and Y_{i0}. The intercept of the corre-lated random effects distribution drops during the Great Recession[10], which is consistent with the fact that bank revenues eroded during the financial crisis. The estimated common autoregressive coefficients range from 0.7 to 0.9.

7.2 Results from Models with Covariates

To analyze the performance of the banking sector under stress scenarios it is necessary to add predictors to the dynamic panel data model that reflect macroeconomic and financial conditions. We consider three aggregate variables: the unemployment rate, the federal funds rate, and the spread between the federal funds rate and the 10-year treasury bill. Because these predictors are not bank-specific, the effect of the predictors on PPNR has to be identified from time-series variation, which is challenging given the short time-dimension of our panels. We consider two specifications: the first model only includes the unemployment rate as additional predictor and we focus on the $T = 5$ data sets. The second model includes all three aggregate predictors and we estimated it based on the $T = 11$ sample.

We generate forecasts using the actual values of the aggregate predictors (which we can evaluate based on the actual PPNR realizations for the forecast perior) and compare these forecasts to predictions under a stressed scenario, in which we use hypothetical values for the predictors. 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}. We are implicitly assuming that there is no feedback from

---

[10]Recall that the $\tau = 2010$ estimation sample comprises the observations for 2006-2010.

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.

Results From a Model with Unemployment. We use the unemployment rate (UN-RATE) from the FRED database maintained by the Federal Reserve Bank of St. Louis and convert it to annual frequency by temporal averaging. We begin by computing MSEs, which are reported in Table 10. This table has the same format as Table 8: we consider MSEs for 2007, 2012, and averaged across all rolling samples. Moreover, we compute MSEs conditional on the level of PPNR at the forecast origin. A few observations stand out. First, the MSE for the posterior mean predictor is slightly reduced by including unemployment for the 2007 and 2012 samples, but across all of the rolling samples it slightly increases. Second, the gain of using the Tweedie correction, that is, the MSE differential between the plug-in predictor and the posterior mean predictor, becomes larger as we include unemployment. This is very intuitive: the more coefficients need to be estimated based on a given time-series dimension, the more important the shrinkage induced from the prior distribution. Third, the performance of the posterior-mean predictor and the pooled-OLS predictors remain very similar, meaning that the Tweedie correction shrinks toward pooled OLS.[11]

We now impose stress by increasing the unemployment rate by 5%. This corresponds to the unemployment movement in the *severely adverse* macroeconomic scenario in the Federal Reserve's CCAR 2016. In Figure 6 we are comparing one-year-ahead predictions for forecast origins $\tau = 2007$ and $\tau = 2012$ under the actual period $\tau + 1$ unemployment rate and the stressed unemployment rate. Each circle in the graphs corresponds to a particular BHC. We indicate institutions with assets greater than 50 billion dollars[12] by red circles, while the other BHCs appear as blue circles. The large institutions have in general smaller revenues than the smaller BHCs. According to the plug-in predictor (the two right panels), the response to the unemployment shock is very heterogeneous. For about half of the intitutions a rise in unemployment leads to a drop in revenues, whereas for the other half higher unemployment is associated with larger revenues. However, we know from Table 8 that forecasts from the plug-in predictor are fairly inaccurate. The stress-test implications of the posterior mean predictor are markedly different. Due to the strong shrinkage the effect is more homogeneous across institutions and appears to be slightly positive.

---

[11] This is supported by the estimates of $\hat{\omega}_1^2$ and $\hat{\omega}_2^2$ reported in the Online Appendix.

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

Table 10: MSE for Model with Unemployment for $T = 5$

	All	$y_{i\tau} \leq 0$	$y_{i\tau} \leq -1$	$y_{i\tau} \leq -2$	$y_{i\tau} \leq -3$
			Selection $D_i(\mathcal{Y}_{i\tau})$		
Rolling Sample $\tau = 2007$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.88	0.95	1.11	1.40	1.72
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	1.38	1.62	2.23	2.61	3.29
Loss-Function-Based Estimator	1.44	1.23	1.55	2.14	1.92
Pooled OLS	0.88	0.93	1.06	1.31	1.70
Rolling Sample $\tau = 2012$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.49	0.55	0.80	0.92	1.09
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.64	0.67	0.98	1.27	1.73
Loss-Function-Based Estimator	0.84	1.12	1.56	1.66	1.60
Pooled OLS	0.49	0.58	0.85	0.97	1.12
All Rolling Samples $\tau = 2007, \ldots, 2013$					
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.72	0.92	1.16	1.45	1.70
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	2.52	3.90	4.39	6.07	5.88
Loss-Function-Based Estimator	2.14	3.22	3.71	4.91	4.56
Pooled OLS	0.72	0.96	1.23	1.56	1.86

Notes: For the last panel (all rolling samples) the MSEs are computed across the different forecast origins τ.

A Model with Unemployment, Federal Funds Rate, and Spread. We now expand the list of covariates and in addition to the unemployment rate include the federal funds rate and the spread between the federal funds rate and the 10-year treasury bill. Both series are obtained from the FRED database (FEDFUNDS and DGS10). We convert the series into annual frequency by temporal averaging. Because we now have three regressors that do not vary across units (meaning all BHCs are operating within the same macroeconomic conditions, but may have hetereogeneous responses to these conditions), we focus on the data set with the largest time series dimension, namely $T = 11$. MSEs are presented in Table 11. The forecast origin is $\tau = 2013$. As before, the posterior mean predictor with the Tweedie correction strongly dominates the plug-in predictor. Moreover, the posterior mean predictor is also slightly more accurate than the predictor based on pooled OLS.[13] Unlike in the previous cases, the predictor constructed from the loss-function-based estimate of the model coefficients now performs slightly better than the posterior mean predictor.

Figure 7 compares PPNR predictions under the actual macroeconomic conditions and a stressed macroeconomic scenario. The stressed scenario comprises an increase in the unem-

---

[13]While the estimates of the conditional variances of the λ_{ij} coefficients are close to zero, the estimated conditional means of λ_{ij} vary with Y_{i0}. This explains the difference between the posterior mean and the pooled-OLS predictor.

Figure 6: Predictions under Actual and Stressed Scenario for $T = 5$

Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)   Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)
Rolling Sample $\tau = 2007$
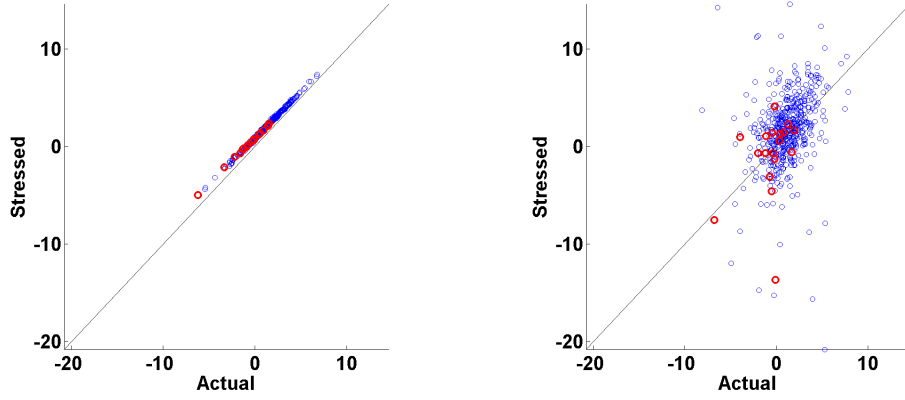
Rolling Sample $\tau = 2012$
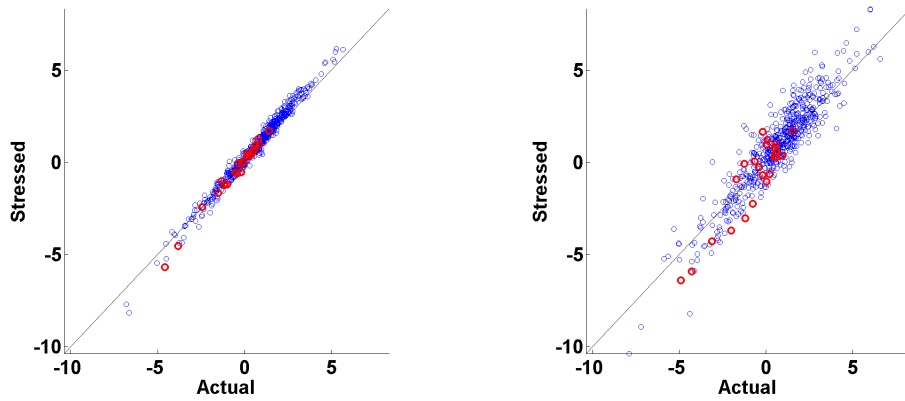
Notes: Each dot corresponds to a BHC in our dataset. We plot point predictions of PPNR under the actual macroeconomic conditions (the unemployment rate is at its observed level in period $\tau + 1$) and a stressed scenario (unemployment rate is 5% higher than its actual level).

Table 11: MSE for Model with Unemployment, Fed Funds Rate, and Spread for $T = 11$

		Selection $D_i(\mathcal{Y}_{i\tau})$			
	All	$y_{i\tau} \leq 0$	$y_{i\tau} \leq -1$	$y_{i\tau} \leq -2$	$y_{i\tau} \leq -3$
Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)	0.49	0.64	0.94	1.00	1.08
Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)	0.78	1.35	2.14	2.04	1.61
Loss-Function-Based Estimator	0.47	0.61	0.88	0.88	0.78
Pooled OLS	0.50	0.68	1.00	1.04	1.10

Notes: The MSEs are computed for the forecast origin $\tau = 2013$.

Figure 7: Predictions under Actual and Stressed Scenario for $T = 11$ and $\tau = 2013$

Post. Mean ($\hat{\theta}_{QMLE}$, Parametric)   Plug-In Predictor ($\hat{\theta}_{QMLE}$, $\hat{\lambda}_i(\hat{\theta}_{QMLE})$)
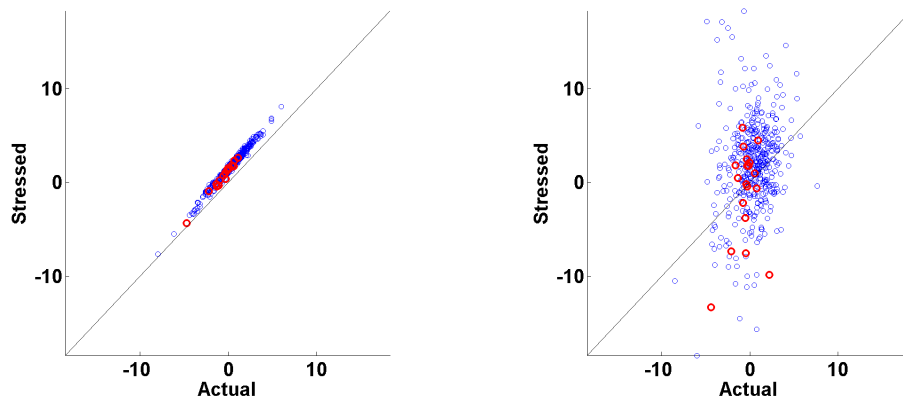
Notes: Each dot corresponds to a BHC in our dataset. We plot point predictions of PPNR under the actual macroeconomic conditions (the unemployment rate, federal funds rate, and spread are at their observed 2014 levels) and a stressed scenario (the unemployment rate, federal funds rate, and spread are 5% higher than their actual level in 2014).

ployment 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. The plug-in predictor generates very heterogeneous responses to the macroeconomic stress scenario. Some banks benefit from the monetary tightening and others experience a substantial fall in revenues. The posterior mean predictor implies a much more homogeneous response of the banking sector under which there is a very small (relative to the cross-sectional dispersion) increase in predicted revenues.

Discussion. 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 model complexity 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 decided to time-aggregate the revenue data to smooth out irregular and non-Gaussian features of the accounting data at the quarterly frequency. This limits the ability to precisely measure the potentially heterogeneous effects of macroeconomic conditions on bank performance. Prior information is used to discipline the inference. In our empirical Bayes procedure, this prior information is essentially extracted from the cross-sectional variation in the data set.

While we *a priori* allowed for heterogeneous responses, it turned out *a posteriori*, trading-off model complexity and fit, that the estimated coefficients exhibited very little heterogeneity. Third, our empirical results indicate that relative to the cross-sectional dispersion of PPNR, 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 and consider a broader set of covariates. Finally, it would be desirable to allow for a feedback from the performance of the banking sector into the aggregate conditions.

8 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, 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 consider both parametric and nonparametric techniques to estimate this distribution. We provide a theorem that establishes a ratio-optimality property for the nonparametric estimator of the Tweedie correction. The nonparametric estimation works well in environments in which the cross-sectional distribution of heterogeneous coefficients is irregular. If it is well approximated by a Gaussian distribution, then a parametric implementation of the Tweedie correction is preferable. We illustrate in an application that our forecasting techniques may be useful to execute bank stress tests. Our paper focuses on one-step-ahead point forecasts. We leave extensions to multi-step forecasting and density forecasting for future work.

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. (2003): *Panel Data Econometrics*. Oxford University Press.

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 S. BONHOMME (2012): "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 79(3), 987–1020.

ARELLANO, M., AND O. BOVER (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of econometrics*, 68(1), 29–51.

ARELLANO, M., AND B. HONORÉ (2001): "Panel data models: some recent developments," *Handbook of econometrics*, 5, 3229–3296.

BALTAGI, B. (1995): *Econometric Analysis of Panel Data*. John Wiley & Sons, New York.

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.

BOTEV, Z. I., J. F. GROTOWSKI, AND D. P. KROESE (2010): "Kernel Density Estimation via Diffusion," *Annals of Statistics*, 38(5), 2916–2957.

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 (2016a): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Economics (Forthcoming)*.

———— (2016b): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Business & Economic Statistics (Forthcoming).*

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

HSIAO, C. (2014): *Analysis of panel data*, no. 54. Cambridge university press.

JIANG, W., C.-H. ZHANG, ET AL. (2009): "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 37(4), 1647–1684.

LANCASTER, T. (2002): "Orthogonal parameters and panel data," *The Review of Economic Studies*, 69(3), 647–666.

LIU, L. (2016): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *Manuscript, University of Pennsylvania.*

NEYMAN, J., AND E. L. SCOTT (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

NICKELL, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49(6), 1417–1426.

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.

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 2

Lemma A.1 *Suppose that* $T \geq k_w + 1 \geq 2$. *Suppose that* W *is a* $T \times k_w$ *matrix with* $\text{rank}(W) = k_w$. *Let* Σ *be a* $T \times T$ *matrix of rank* T. *Let* $S = \Sigma W$. *Then,* $\text{rank}(M_{S \otimes S} B) = T$, *where* $M_{S \otimes S}$ *and* B *are defined in the proof of Theorem 2.3.*

Proof of Lemma A.1. Notice that the matrix B is a $T^2 \times T$ selection matrix that has one at positions $(1, 1), (T + 2, 2), (2T + 3, 3), ..., (T^2, T)$ and zeros at the other positions. Notice that since Σ is full rank, $\text{rank}(S) = \text{rank}(\Sigma W) = \text{rank}(W) = k_w$. If $\text{rank}(S) = k_w$, then $\text{rank}(S \otimes S) = k_w^2$. Since the rank of the projection matrix is the same as its trace, we have $\text{rank}(M_{S \otimes S}) = tr(M_{S \otimes S}) = T^2 - k_w^2$.

By the spectral decomposition, we can decompose $M_{S \otimes S} = F \Lambda F'$, where F is a $T^2 \times T^2$ orthogonal matrix and Λ is a $T^2 \times T^2$ diagonal matrix whose first $T^2 - k_w^2$ elements are one and the rest are zero. Since F is full rank, $\text{rank}(M_{S \otimes S} B) = \text{rank}(F \Lambda F' B) = \text{rank}(\Lambda F' B)$. Notice that $F' B$ is a $T^2 \times T$ matrix that collects the columns of F' in the positions of $1, T + 2, 2T + 3, ..., T^2$. Since the columns of F' are linearly independent, $\text{rank}(F' B) = T$. Notice that $\Lambda F' B$ is a submatrix of $F' B$ that selects the first $T^2 - k_w^2$ rows. Since $T - 1 \geq k_w$ and $T \geq 2$ implies that $T^2 - k_w^2 \geq 2T - 1 > T$, the $(T^2 - k_w^2) \times T$ submatrix of $F' B$, $\Lambda F' B$, has rank T. $\square$

The matrix $\mathbb{E}\big[(W_{it}', X_{it}', Z_{it}')'(W_{it}', X_{it}', Z_{it}')\big]$ has full rank for $t = 1, \ldots, T$. The matrices $\sum_{s=t+1}^{T} W_{is-1} W_{is-1}'$ are invertible with probability one for all $t = 1, \ldots, T - k_w$ and $i = 1, \ldots, N$.

Proof of Theorem 2.3. (i) The parameters α and ρ are identifiable by Assumption 2.2.

(ii) Let Y_i, W_i, X_i, Z_i and U_i denote the matrices vectors that stack Y_{it}, W'_{it-1}, X'_{it-1}, Z'_{it-1}, and U_{it}, respectively, for $t = 1, \ldots, T$. Define

$$
\begin{aligned}
\Sigma_i^{1/2}(\gamma) &= \operatorname{diag}\big(\sigma_1(h_i, \gamma_1), \ldots, \sigma_T(h_i, \gamma_T)\big), \\
S_i(\gamma) &= \Sigma_i^{-1/2}(\gamma)W_i, \quad M_i(\gamma) = I - S_i(S'_i S_i)^{-1} S'_i.
\end{aligned}
$$

Using the same manipulation as in the main text, we obtain the condition

$$
M_i(\tilde{\gamma})\big(\Sigma_i^{-1/2}(\tilde{\gamma})\Sigma_i(\gamma)\Sigma_i^{-1/2}(\tilde{\gamma}) - I\big)M'_i(\tilde{\gamma}) = 0. \tag{A.1}
$$

for each h_i. Taking expectations with respect to H_i and using Assumption 2.2(ii), we deduce that

$$
\mathbb{E}\big[M_i(\tilde{\gamma})\big(\Sigma_i^{-1/2}(\tilde{\gamma})\Sigma_i(\gamma)\Sigma_i^{-1/2}(\tilde{\gamma}) - I\big)M'_i(\tilde{\gamma})\big] = 0. \tag{A.2}
$$

if and only if $\tilde{\gamma} = \gamma$.

(iii) The subsequent argument is similar to the proof of Theorem 2 in Arellano and Bonhomme (2012). Conditional on ρ, α, and γ we can remove the effect of X_i and Z_i from Y_i and define

$$
\tilde{Y}_i = \Sigma_i^{-1/2}(\gamma)(Y_i - X_i\rho - Z_i\alpha) = S_i(\gamma)\lambda_i + V_i. \tag{A.3}
$$

To simplify the notation, we will omit the i subscripts and the γ argument in the remainder of the proof.

Because $S(\gamma)$, λ and V are independent conditional on H (and γ), we have

$$
\ln \Psi_{\tilde{Y}}(\tau|h) = \ln \Psi_\lambda(S'\tau|h) + \ln \Psi_V(\tau) \tag{A.4}
$$

Taking the second derivative with respect to τ leads to

$$
\begin{aligned}
\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_{\tilde{Y}}(\tau|h) &= \frac{\partial^2}{\partial\tau\partial\tau'}\big(\ln\Psi_\lambda(S'\tau|h)\big) + \frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau) \\
&= S\left(\frac{\partial^2}{\partial\xi\partial\xi'}\ln\Psi_\lambda(S'\tau|h)\right)S' + \frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau).
\end{aligned} \tag{A.5}
$$

Using the assumption that the V_ts are independent over t, we can write

$$
\ln\Psi_V(\tau) = \sum_{t=1}^T \ln\Psi_{V_t}(\tau_t),
$$

where Ψ_{V_t} is the characteristic function of V_t. Then,

$$\text{vec}\left(\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau)\right) = \text{vec}\left(\text{diag}\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)\right) \quad (A.6)$$

$$= B\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)'$$

for a suitably chosen matrix B. Let

$$M_{S\otimes S} = I - S(S'S)^{-1}S' \otimes S(S'S)^{-1}S'.$$

Then,

$$M_{S\otimes S}\text{vec}(\ln\Psi_{\tilde{Y}}(\tau|h)) = M_{S\otimes S}B\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)'. \quad (A.7)$$

Because $\Sigma(\gamma)$ is of full rank T (Assumption 2.2(iii)) and W is of full rank of k_w (Assumption 2.2(iv)), $S(\gamma)$ has full rank k_w. Notice that $T \geq k_w + 1$. Then, according to Lemma A.1, $M_{S\otimes S}B$ is also full rank. In turn, from (A.7), we can identify $\ln\Psi_{V_t}(\tau_t)$ uniquely for $t = 1, ..., T$. Also using the restrictions that $\frac{\partial}{\partial\tau_t}\ln\Psi_{V_t}(0) = 0$ ($\mathbb{E}(V_{it}) = 0$) and $\ln\Psi_{V_t}(0) = 0$, we can deduce that the characteristic function of V_t is uniquely identified.

Next, we show how to identify $\ln\Psi_\lambda(\tau|h)$. Because $\ln\Psi_{\tilde{Y}}(\tau|h)$ and $\ln\Psi_V(\tau)$ are identified, from (A.4) we obtain

$$\ln\Psi_{\tilde{Y}}(\tau|h) - \ln\Psi_V(\tau) = \ln\Psi_\lambda(S'\tau|h). \quad (A.8)$$

Taking second derivatives, we obtain

$$\frac{\partial^2}{\partial\tau\partial\tau'}\left(\ln\Psi_{\tilde{Y}}(\tau|h) - \sum_{t=1}^{T}\ln\Psi_V(\tau_t)\right) = S\left(\frac{\partial^2}{\partial\xi\partial\xi'}\ln\Psi_\lambda(S'\tau|h)\right)S'. \quad (A.9)$$

Because S is of full rank, we can identify

$$\frac{\partial^2}{\partial\xi\partial\xi'}\ln\Psi_\lambda(S'\tau|h) = (S'S)^{-1}S'\left[\frac{\partial^2}{\partial\tau\partial\tau'}\left(\ln\Psi_{\tilde{Y}}(\tau|h) - \sum_{t=1}^{T}\ln\Psi_V(\tau_t)\right)\right]S(S'S)^{-1}. \quad (A.10)$$

The mean $\mathbb{E}(\lambda|h)$ can be identified as follows. Note that

$$\hat{\lambda} = (S'S)^{-1}S'\tilde{Y} = \lambda + (S'S)^{-1}S'V. \quad (A.11)$$

Taking expectations yields

$$\mathbb{E}(\lambda|h) = \mathbb{E}[\hat{\lambda}|h], \tag{A.12}$$

because $\mathbb{E}[(S'S)^{-1}S'V|h] = (S'S)^{-1}S'\mathbb{E}[V|h] = 0$. Once the mean has been determined, we can identify $\ln \Psi_\lambda(\xi|h)$ using $\frac{\partial}{\partial \xi} \ln \Psi_\lambda(0|h) = \mathbb{E}(\lambda|h)$ and $\ln \Psi_\lambda(0|h) = 0$. ∎

Discussion of Assumption 2.2(i). We discuss an example of how to identify α and ρ based on moment conditions in the general model (1). Under the model (1) we can remove the effect of λ_i with the following within projections:

$$Y_{it}^* = Y_{it} - \left(\sum_{s=t+1}^{T} Y_{is} W_{is-1}'\right) \left(\sum_{s=t+1}^{T} W_{is-1} W_{is-1}'\right)^{-1} W_{it-1}$$

$$X_{it-1}^* = X_{it-1} - \left(\sum_{s=t+1}^{T} X_{is-1} W_{is-1}'\right) \left(\sum_{s=t+1}^{T} W_{is-1} W_{is-1}'\right)^{-1} W_{it-1}$$

$$Z_{it-1}^* = Z_{it-1} - \left(\sum_{s=t+1}^{T} Z_{is-1} W_{is-1}'\right) \left(\sum_{s=t+1}^{T} W_{is-1} W_{is-1}'\right)^{-1} W_{it-1}$$

for $t = 1, \ldots, T - k_w$. Because $\mathbb{E}[U_{it}|Y_i^{1:t-1}, H_i, \lambda_i] = 0$, we obtain the moment condition

$$\mathbb{E}\left[\left(Y_{it}^* - \begin{bmatrix} \tilde{\rho}' & \tilde{\alpha}' \end{bmatrix} \begin{bmatrix} X_{it-1}^* \\ Z_{it-1}^* \end{bmatrix}\right) \begin{bmatrix} X_{it-s-1}' & Z_{it-s-1}' \end{bmatrix}\right] = 0 \tag{A.13}$$

for $s \geq 0$. To simplify the exposition, suppose that we choose $[X_{it-1}, Z_{it-1}]$ as instrumental variables. In this case, for the moment conditions to be only satisfied only at $\tilde{\rho} = \rho$ and $\tilde{\alpha} = \alpha$ it is necessary that the matrix

$$\mathbb{E} \begin{bmatrix} X_{it-1}^* X_{it-1}' & X_{it-1}^* Z_{it-1}' \\ Z_{it-1}^* X_{it-1}' & Z_{it-1}^* Z_{it-1}' \end{bmatrix} \tag{A.14}$$

is full rank. Consider, for instance, the upper-left element. We can write

$$\mathbb{E}[X_{it-1}^* X_{it-1}']$$

$$= \mathbb{E}\left[\left(X_{it-1} - \left(\sum_{s=t+1}^{T} X_{is-1}W_{is-1}'\right)\left(\sum_{s=t+1}^{T} W_{is-1}W_{is-1}'\right)^{-1} W_{it-1}\right) X_{it-1}'\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(X_{it-1} - \left(\sum_{s=t+1}^{T} X_{is-1}W_{is-1}'\right)\left(\sum_{s=t+1}^{T} W_{is-1}W_{is-1}'\right)^{-1} W_{it-1}\right) X_{it-1}' \,\Big|\, W_i^{t:T-1}\right]\right]$$

$$= \mathbb{E}[X_{it-1}X_{it-1}'] - \frac{1}{T-h}\left(\sum_{s=t+1}^{T} \mathbb{E}\left[\mathbb{E}[X_{is-1}X_{it-1}'|W_i^{t:T-1}]\right.\right.$$

$$\times W_{is-1}'\left(\frac{1}{T-h}\sum_{s=t+1}^{T} W_{is-1}W_{is-1}'\right)^{-1} W_{it-1}\right]\bigg)$$

$$= \mathbb{E}[X_{it-1}X_{it-1}'] - \frac{1}{T-h}\sum_{s=t+1}^{T} \kappa_s \mathbb{E}[X_{is-1}X_{it-1}'] = I + II, \text{ say.}$$

The fourth equality is based on the assumption that the W_{it}'s are strictly exogenous. The completion of the identification argument requires a moment bound for

$$\kappa_s = \mathbb{E}\left[W_{is-1}'\left(\frac{1}{T-h}\sum_{s=t+1}^{T} W_{is-1}W_{is-1}'\right)^{-1} W_{it-1}\right],$$

a full rank condition on $\mathbb{E}[X_{it-1}X_{it-1}']$, and a condition that ensures that term II does not induce a rank deficiency in term I. Similar conditions need to be imposed on the terms that appear in the other submatrices of (A.14).

A.2 Proofs for Section 5

A.2.1 Sufficient Conditions for Assumption 5.3(iii)

The high-level condition in Assumption 5.3(iii) is satisfied if the following two conditions hold:

(a) There exists a sequence $D_N \to \infty$ such that $B_N D_N = o(1)$ and

$$\exp\left(-\frac{D_N^2}{2}\right) = o(1)\left(\inf_{y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N], \lambda \in \Lambda^\pi} \pi(y|\lambda)\right).$$

(b) There exists a shrinking neighborhood of y and a function $\delta(y, \lambda)$ such that for any $|a| \leq \kappa_N \to 0$,

$$|\pi(y|\lambda) - \pi(y + a|\lambda)| \leq \delta(y, \lambda)|a|,$$

where

$$\sup_{y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N], \lambda \in \Lambda^\pi} \left| B_N \frac{\delta(y, \lambda)}{\pi(y|\lambda)} \right| = o(1).$$

The claim can be verified as follows. For $|y| \leq \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N]$ and $\lambda \in \Lambda^\pi$, by the change-of-variable with $y^* = \frac{\tilde{y}-y}{B_N}$, we have

$$\int \frac{1}{B_N} \phi\left(\frac{\tilde{y} - y}{B_N}\right)\left(\frac{\pi(\tilde{y}|\lambda)}{\pi(y|\lambda)} - 1\right) d\tilde{y} = \int \phi(y^*)\left(\frac{\pi(y + B_N y^*|\lambda) - \pi(y|\lambda)}{\pi(y|\lambda)}\right) dy^*.$$

Split the integration into two, one over $|y^*| \leq D_N$ and other one over $|y^*| > D_N$. By Assumption 5.3(i) and (iii)-(a), uniformly in $|y^*| \leq D_N$ and other one over $|y^*| > D_N$,

$$\left|\int_{|y^*|>D_N} \phi(y^*)\left(\frac{\pi(y + B_N y^*|\lambda) - \pi(y|\lambda)}{\pi(y|\lambda)}\right) dy^*\right| \leq \frac{M \int_{|y^*|>D_N} \phi(y^*) dy^*}{\inf_{y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N], \lambda \in \Lambda^\pi} \pi(y|\lambda)}$$

$$\leq \frac{M \exp\left(-\frac{D_N^2}{2}\right)}{\inf_{y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N], \lambda \in \Lambda^\pi} \pi(y|\lambda)}$$

$$= o(1)$$

Also, notice that since $|y^*| \leq D_N$, $|B_N y^*| \leq B_N D_N = o(1)$. Then, by Assumption (iii)-(b),

$$\left|\int_{|y^*|\leq D_N} \phi(y^*)\left(\frac{\pi(y + B_N y^*|\lambda) - \pi(y|\lambda)}{\pi(y|\lambda)}\right) dy^*\right| \leq \int \phi(y^*) y^* dy^* \left|\frac{\delta(y, \lambda)}{\pi(y|\lambda)} B_N\right|$$

$$= Mo(1) = o(1)$$

uniformly in $y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N]$ and $\lambda \in \Lambda^\pi$.

A.2.2 An Example of a $\pi(y|\lambda)$ That Satisfies Assumption 5.3

Consider $\pi(y|\lambda) = \phi(y - \lambda)$, where $\phi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$. First, since $0 < \phi(x) < 1$, Assumption 5.3(i) is satisfied. To verify Assumption 5.3(ii), notice that because $Y_{i0}|\lambda_i \sim N(\lambda_i, 1)$, we have for $C \geq 0$,

$$\mathbb{P}\{Y_{i0} \geq C|\lambda_i = \lambda\} \leq \exp\left(-\frac{(C - \lambda)^2}{2}\right).$$

In this case, $m(C, \lambda) = (C - \lambda)^2/2$. Choose $K \geq \max\{1, \sqrt{2(2 + \epsilon)}\}$ with any $\epsilon \geq 0$. Then,

$$\liminf_{N \longrightarrow \infty} \inf_{|\lambda| \leq C_N} \left(m(K(\sqrt{\ln N} + C_N), \lambda) - (2 + \epsilon) \ln N \right) \geq 0,$$

as required for Assumption 5.3(ii), regardless of the specific rate of C_N. To verify Assumption 5.3(iii) we can use the closed-form expression for the convolution:

$$\int \frac{1}{B_N} \phi\left(\frac{\tilde{y} - y}{B_N}\right) \pi(\tilde{y}|\lambda) d\tilde{y} = \frac{1}{\sqrt{1 + B_N^2}} \phi\left(\frac{y - \lambda}{\sqrt{1 + B_N^2}}\right).$$

Note that we can write

$$\phi\left(\frac{y - \lambda}{\sqrt{1 + B_N^2}}\right) = \phi(y - \lambda) \exp\left(\frac{(B_N(y - \lambda))^2}{2(1 + B_N^2)}\right).$$

Thus,

$$\sup_{y \in \mathcal{Y}_\lambda^\pi \cap [-C_N', C_N], \lambda \in \Lambda^\pi} \exp\left(\frac{(B_N(y - \lambda))^2}{2(1 + B_N^2)}\right) - 1 \leq \exp\left((B_N(C_N' + C_N))^2\right) - 1 = o(1),$$

according to Assumption 5.2.

A.2.3 Main Theorem

Proof of Theorem 5.5. The goal is to prove that for a given $\epsilon_0 > 0$

$$\limsup_{N \to \infty} \frac{R_N(\widehat{Y}_{T+1}^N) - R_N^{\text{opt}}}{N \mathbb{E}_\theta^{\mathcal{Y}^i, \lambda_i} \left[(\lambda_i - \mathbb{E}_{\theta, \mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 0, \tag{A.15}$$

where

$$R_N(\widehat{Y}_{T+1}^N) = N \mathbb{E}_\theta^{\mathcal{Y}^N, \lambda_i} \left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right] + N\sigma^2$$

$$R_N^{\text{opt}} = N \mathbb{E}_\theta^{\mathcal{Y}_i, \lambda_i} \left[\left(\lambda_i - \mathbb{E}_{\theta, \mathcal{Y}^i}^{\lambda_i}[\lambda_i]\right)^2\right] + N\sigma^2.$$

Here we used the fact that there is cross-sectional independence and symmetry in terms of i. The statement is equivalent to

$$\limsup_{N \to \infty} \frac{N \mathbb{E}_\theta^{\mathcal{Y}^N, \lambda_i}\left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right]}{N \mathbb{E}_\theta^{\mathcal{Y}^i, \lambda_i}\left[(\lambda_i - \mathbb{E}_{\theta, \mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right] + N^{\epsilon_0}} \le 1. \tag{A.16}$$

Forecast Error Decomposition. We decompose the forecast error as follows: Using the previously developed notation, we expand the prediction error due to parameter estimation as follows:

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

We define the density $p_*(\hat{\lambda}_i(\rho), Y_{i0})$ as the expected value of the kernel density estimator:

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

It can be calculated as follows. Taking expectations with respect to $(\hat{\lambda}_j, y_{j,0})$ for $j \ne i$ yields

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

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

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

and change the order of integration. This leads to:

$$\mathbb{E}_{\theta,\mathcal{Y}_i}^{\mathcal{Y}^{(-i)}}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})]$$

$$= \int\int\left[\int \frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_i - \hat{\lambda}_j}{B_N}\right)p(\hat{\lambda}_j|\lambda_j)d\hat{\lambda}_j\right]\frac{1}{B_N}\phi\left(\frac{y_{i0} - y_{j0}}{B_N}\right)\pi(\lambda_j, y_{j0})d\lambda_j dy_{j0}$$

$$= \int\int \frac{1}{\sqrt{\sigma^2/T + B_N^2}}\phi\left(\frac{\hat{\lambda}_i - \lambda_j}{\sqrt{\sigma^2/T + B_N^2}}\right)\frac{1}{B_N}\phi\left(\frac{y_{i0} - y_{j0}}{B_N}\right)\pi(\lambda_j, y_{j0})d\lambda_j dy_{j0}$$

$$= \int \frac{1}{\sqrt{\sigma^2/T + B_N^2}}\phi\left(\frac{\hat{\lambda}_i - \lambda_j}{\sqrt{\sigma^2/T + B_N^2}}\right)\left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0} - y_{j0}}{B_N}\right)\pi(y_{j0}|\lambda_j)dy_{j0}\right]\pi(\lambda_j)d\lambda_j.$$

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

$$p_*(\hat{\lambda}_i, y_{i0})$$

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

Risk Decomposition. Write

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

We deduce from the C_r inequality that the statement of the theorem follows if we can show that for the $\epsilon_0 > 0$ given in Definition 3.2:

$$\begin{array}{ll}
\text{(i)} & N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{1i}^2\right] = o(N^{\epsilon_0}) \\[2mm]
\text{(ii)} & \limsup\limits_{N\to\infty} \dfrac{N\mathbb{E}_\theta^{\mathcal{Y}^N,\lambda_i}\left[A_{2i}^2\right]}{N\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i - \mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 1 \\[2mm]
\text{(iii)} & N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{3i}^2\right] = o(N^{\epsilon_0}).
\end{array}$$

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

A.2.4 Three Important Lemmas

Truncations. The remainder of the proof involves a number of truncations that we will apply when analyzing the risk terms. For now, $L_N = o(N^\epsilon)$ will be a sequence such that

$L_N \longrightarrow \infty$ as $N \longrightarrow \infty$. We will specify the rate at which L_N diverges below.

1. Define the truncated region $\mathcal{T}_1 = \{|\hat{\sigma}^2 - \sigma^2| \leq 1/L_N\}$. By Chebyshev's inequality and Assumption 5.4, we can bound

$$N\mathbb{P}(\mathcal{T}_1^c) = N\mathbb{P}\{|\hat{\sigma}^2 - \sigma^2| > 1/L_N\} \leq L_N^2 \mathbb{E}[N(\hat{\sigma}^2 - \sigma^2)^2] = o(N^\epsilon),$$

provided that $L_N^2 = o(N^\epsilon)$ for any ϵ.

2. Define the truncated region $\mathcal{T}_2 = \{|\hat{\rho} - \rho| \leq 1/L_N^2\}$. By Chebyshev's inequality and Assumption 5.4, we can bound

$$N\mathbb{P}(\mathcal{T}_2^c) = N\mathbb{P}\{|\hat{\rho} - \rho| > 1/L_N^2\} \leq L_N^4 \mathbb{E}\left[N(\hat{\rho} - \rho)^2\right] = o(N^\epsilon),$$

provided that $L_N^4 = o(N^\epsilon)$ for any ϵ.

3. 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 $\mathcal{T}_3 = \left\{\max_{1 \leq i \leq N} |\bar{U}_{i,-1}(\rho)| \leq M_3 L_N\right\}$ for some constant M_3. Notice that $\bar{U}_{i,-1}(\rho) \sim iidN(0, \sigma_{\bar{U}}^2)$ with $0 < \sigma_{\bar{U}}^2 < \infty$. Thus, we have

$$
\begin{aligned}
N\mathbb{P}(\mathcal{T}_3^c) &= N\mathbb{P}\{\max_{1 \leq i \leq N} |\bar{U}_{i,-1}(\rho)| \geq L_N\} \\
&\leq N \sum_{i=1}^{N} \mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} \\
&= N^2 \mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} \\
&\leq 2\exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N\right). \quad\quad\quad\text{(A.18)}
\end{aligned}
$$

4. Define the truncated region $\mathcal{T}_4 = \{\max_{1 \leq i \leq N} |Y_{i0}| \leq L_N\}$. Then,

$$
\begin{aligned}
N\mathbb{P}\mathcal{T}_4^c &= N\mathbb{P}\{\max_{1 \leq i \leq N} |Y_{i0}| \geq L_N\} \\
&\leq N \sum_{i=1}^N \mathbb{P}\{|Y_{i0}| \geq L_N\} \\
&= N^2 \int \left[\int_{L_N}^\infty \pi(y_0|\lambda) dy_0 + \int_{-\infty}^{-L_N} \pi(y_0|\lambda) dy_0 \right] \pi_\lambda(\lambda) d\lambda \\
&\leq 2N^2 \int \exp\left[-m\left(L_N, \lambda\right)\right] \pi(\lambda) d\lambda \\
&\leq 2C_N \left(\sup_{|\lambda| \leq C_N} \exp\left[-m\left(L_N, \lambda\right) + 2\ln N\right] \right),
\end{aligned}
\tag{A.19}
$$

where the last three lines hold by Assumptions 5.1 and 5.3.

5. Let $\bar{Y}_{i,-1} = C_1(\rho) Y_{i0} + C_2(\rho)\lambda_i + \bar{U}_{i,-1}(\rho)$, where $C_1(\rho) = \frac{1}{T}\sum_{t=1}^T \rho^{t-1}$, $C_2(\rho) = \frac{1}{T}\sum_{t=2}^T (1 + \cdots + \rho^{t-2})$. According to Assumption 5.1 the support of λ_i is contained in $[-C_N, C_N]$. Moreover, because T is finite, $|C_1(\rho)| \leq 1$ and $|C_2(\rho)| < T$. Then, in the region $\mathcal{T}_3 \cap \mathcal{T}_4$:

$$
\begin{aligned}
\max_{1 \leq i \leq N} |\bar{Y}_{i,-1}| &\leq |C_1(\rho)| \max_{1 \leq i \leq N} |\lambda_i| + |C_2(\rho)| \max_{1 \leq i \leq N} |Y_{i0}| + \max_{1 \leq i \leq N} |\bar{U}_{i,-1}(\rho)| \\
&\leq C_N + T L_N + \exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N \right)
\end{aligned}
$$

which leads to

$$
\max_{1 \leq i,j \leq N} |\bar{Y}_{j,-1} - \bar{Y}_{i,-1}| \leq 2 \max_{1 \leq i \leq N} |\bar{Y}_{i,-1}| \leq 2\left(C_N + T L_N + \exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N \right) \right).
\tag{A.20}
$$

6. For the region $\mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$ we obtain the bound

$$
\max_{1 \leq i,j \leq N} |(\hat{\rho} - \rho)(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})| \leq \frac{2\left(C_N + T L_N + \exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N \right) \right)}{L_N^2}.
\tag{A.21}
$$

Recall that $C_N = o(N^\epsilon)$ is the truncation for the support of the prior of λ (Assumption 5.1). We will choose

$$
L_N = o(N^\epsilon) \text{ such that } L_N = \max\left\{ \sigma_{\bar{U}} \sqrt{2(2+\epsilon)\ln N}, K(\sqrt{\ln N} + C_N), \frac{1}{B_N}, C_N \right\},
\tag{A.22}
$$

so that we can deduce

$$N\mathbb{P}\mathcal{T}_1^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_2^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_3^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_4^c = o(N^\epsilon)$$

$$(A.20) = o(N^\epsilon), \quad (A.21) = o(N^\epsilon). \tag{A.23}$$

for any ϵ.

A.2.4.1 Term A_{1i}

Lemma A.2 *Suppose the assumptions in Theorem 5.5 hold. Then,*

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

Proof of Lemma A.2. We begin with the following bound:

$$
\begin{aligned}
|A_{1i}| &= \left|\left[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\right]^{C_N} - \mu\big(\hat{\lambda}_i(\rho), \sigma^2/T + B_N^2, p_*(\hat{\lambda}_i(\rho), Y_{i0})\big)\right| \\
&\leq \left|\left[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0})\big)\right]^{C_N}\right| + \left|\mu\big(\hat{\lambda}_i(\rho), \sigma^2/T + B_N^2, p_*(\hat{\lambda}_i(\rho), Y_{i0})\big)\right| \\
&\leq 2C_N. \tag{A.24}
\end{aligned}
$$

The last equality follows from the fact that the second term can be interpreted as a posterior mean under the likelihood function

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

and the prior distribution $\pi(\lambda)$. Because, according to Assumption 5.1, the prior has support on the interval $[-C_N, C_N]$, we can deduce that the posterior mean has to be bounded by C_N as well. Then,

$$
\begin{aligned}
N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2] &\leq N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2\mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)] + C_N^2 N\left(\mathbb{P}\mathcal{T}_1^c + \mathbb{P}\mathcal{T}_2^c + \mathbb{P}\mathcal{T}_3^c + \mathbb{P}\mathcal{T}_4^c\right) \\
&\leq N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2\mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)] + o(N^{\epsilon_0}). \tag{A.25}
\end{aligned}
$$

The bound for the second term follows from the fact that (A.23) and (A.24) hold for any $\epsilon > 0$, including ϵ_0. In the remainder of the proof we will construct a bound for the first term on the right-hand side of (A.25). We proceed in two steps.

Step 1. We introduce two additional trunctation regions, $\mathcal{T}_{5i}$ and $\mathcal{T}_{6i}$, which are defined as follows:

$$
\begin{aligned}
\mathcal{T}_{5i} &= \left\{ (\hat{\lambda}_i, Y_{i0}) \,\middle|\, -C_N' \leq \hat{\lambda}_i \leq C_N',\ -C_N' \leq Y_{i0} \leq C_N' \right\} \\
\mathcal{T}_{6i} &= \left\{ (\hat{\lambda}_i, Y_{i0}) \,\middle|\, p(\hat{\lambda}_i, Y_{i0}) \geq \frac{N^{\epsilon'}}{N} \right\},
\end{aligned}
$$

where $C_N' > C_N$ will be defined in (A.28) below and it is assumed that $0 < \epsilon' < \epsilon_0$. 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 "high." We will show that

$$
N \mathbb{E}_\theta^{\mathcal{Y}^N} [A_{1i}^2 \mathbb{I}(\mathcal{T}_{5i}) \mathbb{I}(\mathcal{T}_{6i}^c)] \leq o(N^{\epsilon_0}) \tag{A.26}
$$

$$
N \mathbb{E}_\theta^{\mathcal{Y}^N} [A_{1i}^2 \mathbb{I}(\mathcal{T}_{5i}^c)] \leq o(N^{\epsilon_0}). \tag{A.27}
$$

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.26). Using the bound for $|A_{1i}|$ in (A.24) we obtain:

$$
\begin{aligned}
N \mathbb{E}_\theta^{\mathcal{Y}^N} \left[A_{1i}^2 \mathbb{I}(\mathcal{T}_{5i}) \mathbb{I}(\mathcal{T}_{6i}^c) \right] &\leq 4 N C_N^2 \mathbb{P}(\mathcal{T}_{5i} \cap \mathcal{T}_{6i}^c) \\
&= 4 N C_N^2 \int_{\hat{\lambda}_i = -C_N'}^{C_N'} \int_{y_{i0} = -C_N'}^{C_N'} \mathbb{I}\left\{ p(\hat{\lambda}_i, y_{i0}) < \frac{N^{\epsilon'}}{N} \right\} p(\hat{\lambda}_i, y_{i0}) d(\hat{\lambda}_i, y_{i0}) \\
&\leq 4 N C_N^2 \int_{\hat{\lambda}_i = -C_N'}^{C_N'} \int_{y_{i0} = -C_N'}^{C_N'} \left(\frac{N^{\epsilon'}}{N} \right) dy_{i0} d\hat{\lambda}_i \\
&\leq 4 C_N^2 (C_N')^2 N^{\epsilon'} \\
&= o(N^{\epsilon_0}).
\end{aligned}
$$

The last equality holds by the definition of C_N' found in (A.28) below. This establishes (A.26).

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:"

$$
\begin{aligned}
&N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{1i}^2\mathbb{I}(\mathcal{T}_{5i}^c)\right]\\
&\leq\; 4NC_N^2\int_{\mathcal{T}_5^c}p(\hat{\lambda}_i, y_{i0})d(\hat{\lambda}_i, y_{i0})\\
&=\; 4NC_N^2\int_{\mathcal{T}_5^c}\left[\int_{\lambda_i}\frac{1}{\sigma/\sqrt{T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)\pi(y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i\right]d(\hat{\lambda}_i, y_{i0})\\
&\leq\; 4NC_N^2\int_{\lambda_i}\left[\int_{|\hat{\lambda}_i|>C'_N}\frac{1}{\sigma/\sqrt{T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)\pi(y_{i0}|\lambda_i)d(\hat{\lambda}_i, y_{i0})\right.\\
&\qquad\left.+\int_{|y_{i0}|>C'_N}\frac{1}{\sigma/\sqrt{T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)\pi(y_{i0}|\lambda_i)d(\hat{\lambda}_i, y_{i0})\right]\pi(\lambda_i)d\lambda_i\\
&=\; 4NC_N^2\int_{|\lambda_i|<C_N}\left[\int_{|\hat{\lambda}_i|>C'_N}\frac{1}{\sigma/\sqrt{T}}\phi\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)d\hat{\lambda}_i\right]\pi(\lambda_i)d\lambda_i\\
&\quad+4NC_N^2\int_{|\lambda_i|<C_N}\left[\int_{|y_{i0}|>C'_N}\pi(y_{i0}|\lambda_i)dy_{i0}\right]\pi(\lambda_i)d\lambda_i\\
&=\; B_1+B_2,\quad\text{say.}
\end{aligned}
$$

The second equality is obtained by integrating out y_{i0} and $\hat{\lambda}_i$, recognizing that the integrant is a properly scaled probability density function that integrates to one. We are able to restrict the range of integration for λ_i to the set $|\lambda_i| < C_N$ because, by assumption, that is the support of the prior density $\pi(\lambda)$

We will first analyze term B_1. Note 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 $\int_x^\infty \phi(\lambda)d\lambda \leq \phi(x)/x$. Assuming that N is sufficiently large such that

$$\sqrt{T}(C_N' - |\lambda_i|)/\sigma > 1$$

for $|\lambda_i| < C_N$, we obtain

$$B_1 \leq 8NC_N^2 \int_{|\lambda_i|<C_N} \exp\left(-\frac{T}{2\sigma^2}(C_N' - |\lambda_i|)^2\right) \pi(\lambda_i)d\lambda_i.$$

We can deduce that $B_1 = o(N^\epsilon)$ for any $\epsilon > 0$ (including ϵ_0) if

$$\inf_{|\lambda_i|<C_N} \frac{T}{2\sigma^2}(C_N' - |\lambda_i|)^2 > \ln N,$$

which follows if we choose

$$C_N' = (1+k)\left(\sqrt{\ln N} + C_N\right), \quad k > \max\{0, \sqrt{2\sigma^2/T} - 1\}. \tag{A.28}$$

This is the rate that appears in Assumption 5.2.

For B_2, notice that under Assumption 5.3(ii) we obtain

$$
\begin{aligned}
B_2 &= 4NC_N^2 \int_{|\lambda_i|<C_N} \left[\int_{|y_{i0}|>C_N'} \pi(y_{i0}|\lambda_i)dy_{i0}\right] \pi(\lambda_i)d\lambda_i \\
&\leq 4NC_N^2 \int_{|\lambda_i|<C_N} 2\exp\left(-m(C_N', \lambda_i)\right)\pi(\lambda_i)d\lambda_i \\
&\leq 8C_N^2 \left[\sup_{|\lambda_i|\leq C_N} \exp\left(-m(C_N', \lambda_i) + \ln N\right)\right] \int_{|\lambda_i|<C_N} \pi(\lambda_i)d\lambda_i \\
&\leq o(N^\epsilon)
\end{aligned}
$$

for any ϵ. This leads to the desired bound in (A.27).

Step 2. It remains to be shown that

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{1i}^2 \mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)\mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i})\right] \leq o(N^{\epsilon_0}). \tag{A.29}$$

We introduce the following notation:

$$
\begin{aligned}
\widetilde{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) & \text{(A.30)} \\
d\widetilde{p}_i^{(-i)} &= \frac{1}{\partial\hat{\lambda}_i(\hat{\rho})}\partial\hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \\
\hat{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\rho), Y_{i0}) \\
d\hat{p}_i^{(-i)} &= \frac{1}{\partial\hat{\lambda}_i(\rho)}\partial\hat{p}^{-i}(\hat{\lambda}_i(\rho), Y_{i0}) \\
p_i &= p(\hat{\lambda}_i(\rho), Y_{i0}) \\
p_{*i} &= p_*(\hat{\lambda}_i(\rho), Y_{i0}) \\
dp_{*i} &= \frac{1}{\partial\hat{\lambda}_i(\rho)}\partial p_*(\hat{\lambda}_i(\rho), Y_{i0}).
\end{aligned}
$$

Using the fact that $|\mu\big(\hat{\lambda}_i(\rho), Y_{i0}, \sigma^2/T + B_N^2, p_*(\hat{\lambda}_i(\rho), Y_{i0})\big)| \leq C_N$ and the triangle inequality, we obtain

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

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

$$
N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{1ji}^2 \mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)\mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i})\right] \leq o(N^{\epsilon_0}), \quad j = 1, 2, 3.
$$

First, using a slightly more general argument than the one used in the proof of Lemma A.4, we can show that

$$
N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{11i}^2\right] = \mathbb{E}_\theta^{\mathcal{Y}^N}\left[N(\hat{\rho} - \rho)^2\bar{Y}_{i,-1}\right] = o(N^{\epsilon_0}).
$$

Second, in the region $\mathcal{T}_{5i}$ we can bound

$$
\left(\frac{\sigma^2}{T} + B_N^2\right)\left|\frac{dp_{*i}}{p_{*i}}\right| = \left|\hat{\lambda}_i(\rho) - \mathbb{E}_\theta\left[\lambda_i\big|\hat{\lambda}_i(\rho), Y_{i0}; p_*(\hat{\lambda}_i(\rho), Y_{i0})\right]\right| \leq C_N' + C_N, \qquad \text{(A.31)}
$$

where $\mathbb{E}_\theta[\lambda_i|\cdot]$ is the posterior expectation of λ_i conditional on $(\hat{\lambda}_i(\rho), Y_{i0})$ under the prior distribution $p_*(\hat{\lambda}_i(\rho), Y_{i0})$. Using Assumption 5.4 we obtain the bound

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[A_{12i}^2 \mathbb{I}(\mathcal{T}_{5i})\right] \le \frac{1}{(\sigma^2/T + B_N^2)^2}\mathbb{E}_\theta^{\mathcal{Y}^N}\left[N(\hat{\sigma}^2 - \sigma^2)^2\right](C_N' + C_N)^2 = o(N^{\epsilon_0}).$$

Finally, note that

$$A_{13i}^2\mathbb{I}(\mathcal{T}_1) \le \left(\frac{\sigma^2}{T} + B_N^2 + \frac{1}{L_N}\right)^2\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2.$$

Thus, the desired result follows if we show

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)\mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i})\right] = o(N^{\epsilon_0}) \qquad \text{(A.32)}$$

To show (A.32), we have to control the denominator and consider the following truncation region:

$$\mathcal{T}_{7i} = \left\{(\hat{\lambda}_i, Y_{i0})\,\bigg|\,\widetilde{p}_i^{(-i)} > \frac{p_{*i}}{2}\right\}. \qquad \text{(A.33)}$$

We first analyze (A.32) on $\mathcal{T}_{7i}$ (Step 2.1) and then on $\mathcal{T}_{7i}^c$ (Step 2.2). We will use the following decomposition:

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

We also will abbreviate $\mathbb{I}(\mathcal{T}_l)\mathbb{I}(\mathcal{T}_k) = \mathbb{I}(\mathcal{T}_l\mathcal{T}_k)$.

Step 2.1. For the region $\mathcal{T}_{7i}$ we have

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]$$

$$\le 2N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{d\widetilde{p}_i^{(-i)} - dp_{*i}}{\widetilde{p}_i^{(-i)} - p_{*i} + p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]$$

$$+2o(N^{\epsilon_0})N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{\widetilde{p}_i^{(-i)} - p_{*i}}{\widetilde{p}_i^{(-i)} - p_{*i} + p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]$$

$$= 2B_{1i} + 2o(N^{\epsilon_0})B_{2i},$$

say. The $o(N^{\epsilon_0})$ bound follows from (A.31). 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

$$
\begin{aligned}
R_{1i}(\rho) &= -\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N^2}\phi\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)^2(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right) \\
&\quad +\frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N^3}\phi\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right), \\
R_{2i}(\rho) &= \frac{1}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right),
\end{aligned}
$$

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

We proceed with the analysis of B_2. Using the lower bound for $\widetilde{p}_i^{(-i)}$ over the region $\mathcal{T}_{7i}$, the C_r inequality, and the law of iterated expectations, we obtain

$$
\begin{aligned}
B_{2i} &\leq 8\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^2}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{p}_i^{(-i)} - p_{*i})^2\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\right] \\
&\quad +8\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^2}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{\rho} - \rho)^2 R_{2i}^2(\tilde{\rho})\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\right] \\
&= 8\mathbb{E}_\theta^{\mathcal{Y}^i}[B_{21i} + B_{22i}],
\end{aligned}
$$

say.

According to Lemma A.7(c) (see Section A.2.5)

$$
\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N(\hat{p}_i^{(-i)} - p_{*i})^2\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right] \leq \frac{M}{B_N^2}p_i\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}).
$$

This leads to

$$
\mathbb{E}_\theta^{\mathcal{Y}^i}[B_{21i}] \leq \frac{M}{B_N^2}\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{p_i}{p_{*i}^2}\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\right] = \frac{M}{B_N^2}\int_{\mathcal{T}_{5i}\cap\mathcal{T}_{6i}}\frac{p_i^2}{p_{*i}^2}d\hat{\lambda}_i dy_{i0}.
$$

According to Lemma A.7(e) (see Section A.2.5)

$$\int_{\mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{p_i^2}{p_{*i}^2} d\hat{\lambda}_i dy_{i0} = o(N^\epsilon).$$

Because $1/B_N^2 = o(N^\epsilon)$ according to Assumption 5.2, we can deduce that

$$\mathbb{E}_\theta^{\mathcal{Y}^i}[B_{21i}] \leq o(N^{\epsilon_0}).$$

Using the Cauchy-Schwarz Inequality, we obtain

$$B_{22i} \leq \frac{1}{p_{*i}^2} \sqrt{\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[N^2(\hat{\rho}-\rho)^4\right]} \sqrt{\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]}.$$

Using the inequality once more leads to

$$
\begin{aligned}
\mathbb{E}_\theta^{\mathcal{Y}^i}[B_{22i}] &\leq \sqrt{\mathbb{E}_\theta^{\mathcal{Y}^N}\left[N^2(\hat{\rho}-\rho)^4\right]} \sqrt{\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^4}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\right]} \\
&\leq M\sqrt{\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{1}{p_{*i}^4}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\right]}.
\end{aligned}
$$

The second inequality follows from Assumption 5.4. According to Lemma A.7(a) (see Section A.2.5)

$$\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}^4(\tilde{\rho})\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right] \leq ML_N^4 p_i^4\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}),$$

where $L_N = o(N^{\epsilon_0})$ was defined in (A.22). This leads to the bound

$$
\begin{aligned}
\mathbb{E}_\theta^{\mathcal{Y}^i}[B_{22i}] &\leq ML_N^2\sqrt{\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\left(\frac{p_i}{p_{*i}}\right)^4\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\right]} \\
&= ML_N^2\sqrt{\int_{\mathcal{T}_{5i}\cap\mathcal{T}_{6i}}\left(\frac{p_i}{p_{*i}}\right)^4 p_i d\hat{\lambda}_i dy_{i0}} \\
&\leq M_*L_N^2\sqrt{\int_{\mathcal{T}_{5i}\cap\mathcal{T}_{6i}}\left(\frac{p_i}{p_{*i}}\right)^4 d\hat{\lambda}_i dy_{i0}} \\
&\leq o(N^{\epsilon_0}).
\end{aligned}
$$

The second inequality holds because the density p_i is bounded from above. The last inequality is proved in Lemma A.7(e) (see Section A.2.5).

We deduce that $B_{2i} = o(N^{\epsilon_0})$. A similar argument can be used to establish that $B_{1i} = o(N^{\epsilon_0})$.

Step 2.2. Over the set $\mathcal{T}_{7i}^c$, since $|A_{1i}| \leq o(N^{\epsilon_0})$, we have

$$N\mathbb{E}_{\theta}^{\mathcal{Y}^N}\left[\left(\frac{d\widehat{p}_i^{(-i)}}{\widehat{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2 \mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i}^c)\right] \leq o(N^{\epsilon_0})N\mathbb{P}_{\theta}^{\mathcal{Y}^N}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i}^c).$$

Notice that

$$
\begin{aligned}
\mathcal{T}_{7i}^c &= \left\{\hat{p}_i^{(-i)} - p_{*i} + (\hat{\rho} - \rho)R_{1i}(\widetilde{\rho}) < -\frac{p_{*i}}{2}\right\} \\
&\subset \left\{\hat{p}_i^{(-i)} - p_{*i} - |\hat{\rho} - \rho||R_{1i}(\widetilde{\rho})| < -\frac{p_{*i}}{2}\right\} \\
&\subset \left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} \cup \left\{|\hat{\rho} - \rho||R_{1i}(\widetilde{\rho})| > \frac{p_{*i}}{4}\right\}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&N\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i}^c) \\
&\leq N\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} + N\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[\left\{|\hat{\rho} - \rho||R_{2i}(\widetilde{\rho})| > \frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i})\right] \\
&\leq N\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} + \frac{16L_N^4}{p_{*i}^2}\mathbb{E}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left[R_{2i}(\widetilde{\rho})^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right] \\
&\leq N\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\} + \frac{ML_N^4}{p_{*i}^2}p_i\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}).
\end{aligned}
$$

The first inequality is based on the superset of $\mathcal{T}_{7i}^c$ from above. The second inequality is based on Chebychev's inequality and trucation $\mathcal{T}_2$. The third inequality uses a version of the result in Lemma A.7(a) in which the remainder is raised to the power of two instead of to the power of four. Moreover, we use the fact that p_i is bounded from above to absorb one of the p_i terms in the constant M.

In Lemma A.7(f) (see Section A.2.5) we apply Bernstein's inequality to bound the probability $\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}$ uniformly over $(\hat{\lambda}_i, Y_{i0})$ in the region $\mathcal{T}_{5i}$, showing that

$$N\mathbb{E}_{\theta}^{\mathcal{Y}^i}\left[\mathbb{P}_{\theta,\mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}\left\{\hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4}\right\}\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\right] = o(N^{\epsilon_0}),$$

as desired. Moreover, according to Lemma A.7(f) (see Section A.2.5)

$$\mathbb{E}_\theta^{\mathcal{Y}^i}\left[\frac{p_i}{p_{*i}^2}\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\right] = \int_{\mathcal{T}_{5i}\cap\mathcal{T}_{6i}}\left(\frac{p_i}{p_{*i}}\right)^2 d\hat{\lambda}_i dy_{i0} = o(N^{\epsilon_0}),$$

which gives us the required result for Step 2.2. Combining the results from Steps 2.1 and 2.2 yields (A.29).

The bound in (A.25) now follows from (A.26), (A.27), and (A.29), which completes the proof of the lemma. ∎

A.2.4.2   Term A_{2i}

Lemma A.3 *Suppose the assumptions in Theorem 5.5 hold. Then,*

$$\limsup_{N\to\infty}\frac{N\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]}{N\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i-\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right]+N^{\epsilon_0}}\le 1$$

Proof of Lemma A.3. Notice that $\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)$ can be interpreted $\mu(\cdot)$ as the posterior mean of λ_i under the $p_*(\cdot)$ measure. We use $\mathbb{E}_{*,\theta}^{\mathcal{Y}^i,\lambda_i}[\cdot]$ to denote the joint distribution of $\mathcal{Y}^i$ and λ_i under the $p_*(\cdot)$ measure. Let $\{\tau_N\}$ be a non-negative sequence such that $\tau_N = o(N^{\epsilon_0})$. The desired result follows if we can show that

$$(i)\qquad\limsup_{N\to\infty}\frac{N\mathbb{E}_{*,\theta}^{\mathcal{Y}^i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]+\tau_N}{N\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i-\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right]+N^{\epsilon_0}}\le 1$$

$$(ii)\qquad\limsup_{N\to\infty}\frac{N\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]}{N\mathbb{E}_{*,\theta}^{\mathcal{Y}^i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]+\tau_N}\le 1,$$

where

$$\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i-\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right]=\mathbb{E}_\theta^{\mathcal{Y}^i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right].$$

Part (i): We will construct an upper bound for the numerator. Using the fact that the posterior mean minimizes the integrated risk, we obtain

$$
N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i} \left[\left(\mu\big(\hat{\lambda}_i(\rho), Y_{i0}, \sigma^2/T + B_N^2, p_*(\hat{\lambda}_i(\rho), Y_{i0})\big) - \lambda_i \right)^2 \right]
$$

$$
\leq \quad N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i} \left[\left(\mu\big(\hat{\lambda}_i(\rho), Y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), Y_{i0})\big) - \lambda_i \right)^2 \right]
$$

$$
= \quad N \int\int p_*(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 d\hat{\lambda}_i dy_{i0}
$$

$$
\leq \quad N \int\int p_*(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) d\hat{\lambda}_i dy_{i0}
$$

$$
+ N4C_N^2 \mathbb{P}(\mathcal{T}_{5i}^c \cup \mathcal{T}_{6i}^c)
$$

$$
= \quad N \int\int p_*(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) d\hat{\lambda}_i dy_{i0} + o(N^{\epsilon_0}).
$$

The second inequality uses the fact that $|\lambda_i| \leq C_N$ and therefore the posterior mean has to be bounded in absolute value by C_N as well. The last line follows from an argument similar to that used in Step 1 of the proof of Lemma A.2.

According to Lemma A.6, we obtain the following uniform bound over the region $\mathcal{T}_{5i} \cap \mathcal{T}_{6i}$:

$$
p_*(\hat{\lambda}_i, y_{i0}) \leq (1 + o(1))p(\hat{\lambda}_i, y_{i0}).
$$

Therefore,

$$
\int\int p_*(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) d\hat{\lambda}_i dy_{i0}
$$

$$
= \quad (1 + o(1)) \int\int p(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) d\hat{\lambda}_i dy_{i0}.
$$

In turn, we obtain the following bound:

$$
N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i} \left[\left(\mu\big(\hat{\lambda}_i(\rho), Y_{i0}, \sigma^2/T + B_N^2, p_*(\hat{\lambda}_i(\rho), Y_{i0})\big) - \lambda_i \right)^2 \right] + \tau_N
$$

$$
\leq \quad (1 + o(1))N \int\int p(\hat{\lambda}_i, y_{i0}) \left(\mu\big(\hat{\lambda}_i(\rho), y_{i0}, \sigma^2/T, p(\hat{\lambda}_i(\rho), y_{i0})\big) - \lambda_i \right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) d\hat{\lambda}_i dy_{i0} + o(N^{\epsilon_0})
$$

$$
\leq \quad (1 + o(1))N\mathbb{E}_{\theta}^{\mathcal{Y}^i,\lambda_i} \left[(\lambda_i - \mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2 \right] + o(N^{\epsilon_0})
$$

$$
\leq \quad (1 + o(1))N\mathbb{E}_{\theta}^{\mathcal{Y}^i,\lambda_i} \left[(\lambda_i - \mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2 \right] + N^{\epsilon_0},
$$

which yields the required result for Part (i).

Part (ii): Similar to the proof of Part (i), we construct an upper bound for the numerator as follows

$$
N\mathbb{E}_\theta^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat\lambda_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat\lambda_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]
$$

$$
= N\int\int p(\hat\lambda_i,y_{i0})\left(\mu\big(\hat\lambda_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat\lambda_i(\rho),y_{i0})\big)-\lambda_i\right)^2 d\hat\lambda_i dy_{i0}
$$

$$
\leq \int\int p_*(\hat\lambda_i,y_{i0})\frac{p(\hat\lambda_i,y_{i0})}{p_*(\hat\lambda_i,y_{i0})}\left(\mu\big(\hat\lambda_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat\lambda_i(\rho),y_{i0})\big)-\lambda_i\right)^2\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat\lambda_i dy_{i0}
$$

$$
\quad +N4C_N^2\mathbb{P}(\mathcal{T}_{5i}^c\cup\mathcal{T}_{6i}^c)
$$

$$
= (1+o(1))N\int\int p_*(\hat\lambda_i,y_{i0})\left(\mu\big(\hat\lambda_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat\lambda_i(\rho),y_{i0})\big)-\lambda_i\right)^2
$$

$$
\quad \times\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat\lambda_i dy_{i0}+o(N^\epsilon),\quad \text{any }\epsilon>0
$$

$$
\leq (1+o(1))N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat\lambda_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat\lambda_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]+\tau_N.
$$

For the last line we used the fact that $\tau_N=o(N^{\epsilon_0})$. We now have the required result for Part (ii).

A.2.4.3 Term A_{3i}

Lemma A.4 *Suppose the assumptions in Theorem 5.5 hold. Then, for any $\epsilon>0$:*

$$
N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[(\hat\rho-\rho)^2 Y_{iT}^2\right]=o(N^\epsilon).
$$

Proof of Lemma A.4. Using the Cauchy-Schwarz inequality, we can bound

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

By Assumption 5.4, we have

$$
\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\big(\sqrt{N}(\hat\rho-\rho)\big)^4\right]\leq o(N^\epsilon)
$$

for any $\epsilon>0$.

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 the assumptions that $|\rho| < 1$ and $U_{it} \sim iidN(0, \sigma^2)$, we deduce that there are finite constants M_1, M_2, M_3 such that

$$
\begin{aligned}
\mathbb{E}_\theta^{\mathcal{Y}^N}\left[Y_{iT}^4\right] &\leq M_1 \mathbb{E}_\theta^{\mathcal{Y}^N}\left[Y_{i0}^4\right] + M_2 \mathbb{E}_\theta^{\mathcal{Y}^N}\left[\lambda_i^4\right] + M_3 \mathbb{E}_\theta^{\mathcal{Y}^N}\left[U_{i1}^4\right] \\
&= M_1 \mathbb{E}_\theta^{\mathcal{Y}^N}\left[Y_{i0}^4\right] + o(N^{\epsilon_0}) + o(N^\epsilon)
\end{aligned}
$$

for any ϵ, where the last line holds because $|\lambda_i| \leq C_N$ according to Assumption 5.1 and U_{i1} is normally distributed and therefore all its moments are finite.

The desired $o(N^\epsilon)$ bound for the fourth moment of Y_{i0} can be obtained as follows (we are dropping subscripts and superscripts from expectation and probability operators):

$$
\begin{aligned}
\mathbb{E}\left[|Y_{i0}|^4\right] &= 4\mathbb{E}\left[\int_0^\infty \mathbb{I}\{|Y_{i0}| \geq \tau\}\tau^3 d\tau\right] \\
&= 4\mathbb{E}\left[\int_0^\infty \mathbb{P}\{|Y_{i0}| \geq \tau|\lambda_i\}\tau^3 d\tau\right] \\
&= 4\mathbb{E}\left[\int_0^{\bar{C}} \mathbb{P}\{|Y_{i0}| \geq \tau|\lambda_i\}\tau^3 d\tau\right] + \mathbb{E}\left[\int_{\bar{C}}^\infty \mathbb{P}\{|Y_{i0}| \geq \tau|\lambda_i\}\tau^3 d\tau\right] \\
&\leq M + \int\left[\int_{\bar{C}}^\infty \exp\left(-m(\tau, \lambda)\right)\tau^3 d\tau\right]\pi_\lambda(\lambda)d\lambda
\end{aligned}
$$

for some finite constant M, where $\bar{C}$ is the constant in Assumption 5.3(ii).

Notice that on the domain $[\bar{C}, \infty)$, the function $\exp\left(-m(\tau, \lambda)\right)$ in decreasing in τ, while the function τ^3 is increasing in τ. W.l.o.g, suppose that $\bar{C} = (1 + k)(\sqrt{\ln N^*} + C_{N^*})$ and $(1 + k)(\sqrt{\ln N} + C_N) > 2\ln N$ for all $N \geq N^*$. Now, let $\tau_N = (1 + k)(\sqrt{\ln N} + C_N)$ and bound the integral with a Riemann sum:

$$
\begin{aligned}
\int_{\bar{C}}^\infty \exp\left(-m(\tau, \lambda)\right)\tau^3 d\tau &\leq \sum_{N=N^*}^\infty \exp\left(-m(\tau_N, \lambda)\right)\tau_{N+1}^3(\tau_{N+1} - \tau_N) \\
&\leq \sum_{N=N^*}^\infty \exp\left(-m(\tau_N, \lambda)\right)\tau_{N+1}^4 \\
&= \sum_{N=N^*}^\infty \exp\left(-m(\tau_N, \lambda) + 4\ln\tau_{N+1}\right) \\
&\leq \sum_{N=N^*}^\infty \exp\left(-(2 + \epsilon)\ln N + 4\ln\tau_{N+1}\right) \\
&= \sum_{N=N^*}^\infty \frac{\tau_{N+1}^4}{N^{2+\epsilon}},
\end{aligned}
$$

for some constant $\epsilon \geq 0$. The last inequality holds by Assumption 5.3(ii). Because $\tau_N^4 = o(N^\epsilon)$, there exists a finite constant M such that

$$\sum_{N=N^*}^{\infty} \frac{\tau_{N+1}^4}{N^{2+\epsilon}} \leq M \sum_{N=N^*}^{\infty} \frac{1}{N^2} < \infty.$$

This leads to the desired result

$$\mathbb{E}\left[|Y_{i0}|^4\right] < \infty. \quad \blacksquare$$

A.2.5 Further Details

We now provide more detailed derivations for some of the bounds used in Section A.2.4. Recall that

$$
\begin{aligned}
R_{1i}(\rho) &= -\frac{1}{N-1} \sum_{j \neq i}^{N} \frac{1}{B_N^2} \phi\left(\frac{\hat\lambda_j(\rho) - \hat\lambda_i(\rho)}{B_N}\right) \left(\frac{\hat\lambda_j(\rho) - \hat\lambda_i(\rho)}{B_N}\right)^2 (\bar{Y}_{j,-1} - \bar{Y}_{i,-1}) \frac{1}{B_N} \phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right) \\
&\quad + \frac{1}{N-1} \sum_{j \neq i}^{N} \frac{1}{B_N^3} \phi\left(\frac{\hat\lambda_j(\rho) - \hat\lambda_i(\rho)}{B_N}\right) (\bar{Y}_{j,-1} - \bar{Y}_{i,-1}) \frac{1}{B_N} \phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right) \\
R_{2i}(\rho) &= \frac{1}{N-1} \sum_{j \neq i}^{N} \frac{1}{B_N} \phi\left(\frac{\hat\lambda_j(\rho) - \hat\lambda_i(\rho)}{B_N}\right) \left(\frac{\hat\lambda_j(\rho) - \hat\lambda_i(\rho)}{B_N}\right) (\bar{Y}_{j,-1} - \bar{Y}_{i,-1}) \frac{1}{B_N} \phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right)
\end{aligned}
$$

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 $\mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$. First, using (A.21) we can bound

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

Thus,

$$\phi \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} - (\tilde{\rho} - \rho) \left(\frac{\bar{Y}_{j,-1} - \bar{Y}_{i,-1}}{B_N} \right) \right) \mathbb{I}(\mathcal{T}_2 \mathcal{T}_3 \mathcal{T}_4)$$

$$\leq \ \phi \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} + \left(\frac{M}{L_N B_N} \right) \right) \mathbb{I} \left\{ \frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \leq -\frac{M}{L_N B_N} \right\}$$

$$+ \phi(0) \mathbb{I} \left\{ \left| \frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \right| \leq \frac{M}{L_N B_N} \right\}$$

$$+ \phi \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} - \left(\frac{M}{L_N B_N} \right) \right) \mathbb{I} \left\{ \frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \geq \frac{M}{L_N B_N} \right\}$$

$$= \ \bar{\phi} \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \right),$$

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

Second, we can use the bound

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

Third, for the region $\mathcal{T}_3 \cap \mathcal{T}_4$ we can deduce from (A.20) that

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

Therefore,

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

Now, define the function

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

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.22), 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 $R_{2i}(\widetilde{\rho})$:

$$\left| R_{2i}(\widetilde{\rho})\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4) \right| \leq \frac{ML_N}{N-1} \sum_{\substack{j \neq i}}^{N} \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \right) \frac{1}{B_N} \phi \left(\frac{Y_{j0} - Y_{i0}}{B_N} \right). \tag{A.34}$$

For the subsequent analysis it is convenient define the function

$$f(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) = \frac{1}{B_N^2} \bar{\phi}_* \left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N} \right) \phi \left(\frac{Y_{j0} - Y_{i0}}{B_N} \right). \tag{A.35}$$

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.2.4. We will abbreviate $\mathbb{E}_{\theta, \mathcal{Y}^i}^{\mathcal{Y}^{(-i)}}[\cdot] = \mathbb{E}_i[\cdot]$ and simply use $\mathbb{E}[\cdot]$ to denote $\mathbb{E}_\theta^{\mathcal{Y}^N}[\cdot]$.

Lemma A.5 *Suppose the assumptions required for Theorem 5.5 are satisfied. Then, for a finite positive integer m, over the region $\mathcal{T}_{5i}$, we have*

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

Proof of Lemma A.5. We have

$$\mathbb{E}_i \left[f^m(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0}) \right]$$
$$= \int \left(\frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right) \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right) \right)^m p(\hat{\lambda}, y_0) d(\hat{\lambda}, y_0)$$
$$= \frac{1}{B_N^{2(m-1)}} \int \left\{ \int \frac{1}{B_N} \bar{\phi}_* \left(\frac{\hat{\lambda} - \hat{\lambda}_i}{B_N} \right)^m \frac{1}{B_N} \phi \left(\frac{y_0 - Y_{i0}}{B_N} \right)^m p(\hat{\lambda}, y_0|\lambda) d(\hat{\lambda}, y_0) \right\} \pi(\lambda) d\lambda.$$

The inner integral is

$$\int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m p(\hat{\lambda},y_0|\lambda)d(\hat{\lambda},y_0)$$

$$= \int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)d\hat{\lambda}$$

$$\times \int \frac{1}{B_N}\phi\left(\frac{y_0-Y_{i0}}{B_N}\right)^m \pi(y_0|\lambda)dy_0$$

$$= I_1 \times I_2,$$

say.

Notice that

$$I_1 = \int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)d\hat{\lambda}$$

$$= \int \bar{\phi}_*(\lambda^*)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i+B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^*$$

$$= \int \bar{\phi}_*(\lambda^*)^m \exp\left(-\left((\hat{\lambda}_i-\lambda_i)B_N\lambda^*\right)\frac{1}{\sigma^2/T}\right)\exp\left(-\frac{1}{2}\left(\frac{B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^*$$

$$\times \left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right]$$

$$\leq M\left(\int \bar{\phi}_*(\lambda^*)^m \exp\left(v_N\lambda^*\right)d\lambda^*\right)\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right]$$

$$\leq M\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] = Mp(\hat{\lambda}_i|\lambda_i,Y_{i0}).$$

We used the change-of-variable $\lambda_* = (\hat{\lambda}-\hat{\lambda}_i)/B_N$ to replace $\hat{\lambda}$. Here the second 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 a constant. Moreover, under truncation $\mathcal{T}_{5i}$, $|\hat{\lambda}_i| \leq C_N'$ and the support of λ_i is bounded by $[-C_N, C_N]$ (under Assumption 5.1). Thus, $v_N = B_N(C_N' + 2C_N)$. According to Assumption 5.2 $v_N = B_N(C_N' + 2C_N) = o(1)$. Thus, the last inequality holds because $\int \bar{\phi}_*(\lambda^*)^m \exp\left(v_N\lambda^*\right)d\lambda^*$ is finite. Finally, note that $p(\hat{\lambda}_i|\lambda_i, Y_{i0}) = p(\hat{\lambda}_i|\lambda_i)$.

We now proceed with a bound for the second integral, I_2. Using the fact that the Gaussian pdf $\phi(x)$ is bounded, 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 $|\lambda| \leq C_N$. Here the last equality follows from Assumption 5.3(iii). Combining the bounds for I_1 and I_2 and integrating over λ, we obtain

$$
\begin{aligned}
\mathbb{E}_i\big[f^m(\hat{\lambda}_j - \hat{\lambda}_i, Y_{j0} - Y_{i0})\big] &= \frac{1}{B_N^{2(m-1)}} \int I_1 \times I_2 \pi(\lambda_i)d\lambda_i \\
&\leq \frac{1}{B_N^{2(m-1)}} M\big(1 + o(1)\big) \int p(\hat{\lambda}_i|\lambda_i, Y_{i0})p(Y_{i0}|\lambda_i)\pi(\lambda_i)d\lambda_i \\
&= \frac{1}{B_N^{2(m-1)}} M\big(1 + o(1)\big)p_i,
\end{aligned}
$$

as required.

Lemma A.6 *Suppose the assumptions required for Theorem 5.5 are satisfied. Then,*

$$
\sup_{(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{p_i}{p_{*i}} = 1 + o(1) \tag{A.36}
$$

$$
\sup_{(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{p_{*i}}{p_i} = 1 + o(1). \tag{A.37}
$$

Proof of Lemma A.6. We begin by verifying (A.36). Let

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

such that

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

Because $|\lambda_i| \leq C_N$ by Assumption 5.1 and $|\hat{\lambda}_i| \leq C_N'$ in the region $\mathcal{T}_{5i}$, for some finite

constant M we have

$$
\begin{aligned}
\frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{\sigma^2/T}}\right) &= \frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right) \\
&\quad \times \frac{\sqrt{B_N^2 + \sigma^2/T}}{\sqrt{\sigma^2/T}}\exp\left\{-\frac{1}{2}\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right)^2\frac{B_N^2}{\sigma^2/T}\right\} \\
&\leq \frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right) \\
&\quad \times \sqrt{1 + MB_N^2}\exp(-M(C_N' + C_N)^2 B_N^2) \\
&= (1 + o(1))\frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right), \qquad \text{(A.38)}
\end{aligned}
$$

where $o(1)$ is uniform in $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}$. Here we used Assumption 5.2 which implies that $v_N = (C_N' + C_N)B_N = o(1)$.

According to Assumption 5.3(iii),

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

uniformly in $|y_{i0}| \leq C_N'$ and $|\lambda_i| \leq C_N$. This implies that

$$
\pi(y_{i0}|\lambda_i) \leq (1 + o(1))\int \frac{1}{B_N}\phi\left(\frac{y_{i0} - \tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}. \qquad \text{(A.39)}
$$

uniformly in $|y_{i0}| \leq C_N'$ and $|\lambda_i| \leq C_N$.

Then, by combining the bounds in (A.38) and (A.39) we deduce

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

Note that the $o(1)$ term does not depend on $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}$.

We deduce that

$$
\begin{aligned}
\sup_{(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{p_i}{p_{*i}} &= 1 + \sup_{(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{p_i - p_{*i}}{p_{*i}} \\
&= 1 + \sup_{(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \frac{\int \left[p(\hat{\lambda}_i, y_{i0} | \lambda_i) - p_*(\hat{\lambda}_i, y_{i0} | \lambda_i) \right] \pi(\lambda_i) d\lambda_i}{p_{*i}} \\
&= 1 + o(1).
\end{aligned}
$$

This proves (A.36). A similar argument can be used to establish (A.37). ■

Lemma A.7 *Under the assumptions required for Theorem 5.5, we obtain the following bounds:*

(a) $\mathbb{E}_i \left[R_{2i}^4(\tilde{\rho}) \mathbb{I}(\mathcal{T}_2 \mathcal{T}_3 \mathcal{T}_4 \mathcal{T}_{5i} \mathcal{T}_{6i} \mathcal{T}_{7i}) \right] \leq M L_N^4 p_i^4 \mathbb{I}(\mathcal{T}_{5i} \mathcal{T}_{6i})$

(b) $\mathbb{E}_i \left[R_{1i}^4 \mathbb{I}(\mathcal{T}_2 \mathcal{T}_3 \mathcal{T}_4 \mathcal{T}_{5i} \mathcal{T}_{6i} \mathcal{T}_{7i}) \right] \leq M \frac{L_N^4}{B_N^4} p_i^4 \mathbb{I}(\mathcal{T}_{5i} \mathcal{T}_{6i})$

(c) $\mathbb{E}_i \left[N(\hat{p}_i^{(-i)} - p_{*i})^2 \mathbb{I}(\mathcal{T}_2 \mathcal{T}_3 \mathcal{T}_4 \mathcal{T}_{5i} \mathcal{T}_{6i} \mathcal{T}_{7i}) \right] \leq \frac{M}{B_N^2} p_i \mathbb{I}(\mathcal{T}_{5i} \mathcal{T}_{6i})$

(d) $\mathbb{E}_i \left[N(d\hat{p}_i^{(-i)} - dp_{*i})^2 \mathbb{I}(\mathcal{T}_2 \mathcal{T}_3 \mathcal{T}_4 \mathcal{T}_{5i} \mathcal{T}_{6i} \mathcal{T}_{7i}) \right] \leq \frac{M}{B_N^2} p_i \mathbb{I}(\mathcal{T}_{5i} \mathcal{T}_{6i})$

(e) $\int_{\mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \left(\frac{p_i}{p_{*i}} \right)^m d\hat{\lambda}_i dy_{i0} = o(N^\epsilon)$, $m > 1$.

(f) $N \mathbb{E} \left[\mathbb{P}_i \{ \hat{p}_i^{(-i)} - p_{*i} < -p_{*i}/4 \} \mathbb{I}(\mathcal{T}_{5i} \mathcal{T}_{6i}) \right] = o(N^\epsilon)$

Proof of Lemma A.7. **Part (a).** Recall the following definitions

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

First, recall that according to (A.34), in the region $\mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$

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

Then,

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

say. The second inequality holds because $|x+y|^4 \le 8(|x|^4 + |y|^4)$.

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

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

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

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

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.5, 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 $\mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4 \cap \mathcal{T}_{5i} \cap \mathcal{T}_{6i}$

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

The second inequality holds because over $\mathcal{T}_{6i}$, $p_i \geq \frac{N^{\epsilon'}}{N} \geq \frac{M}{N B_N^2}$. 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.6 we can choose a constant c such that

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

over truncations $\mathcal{T}_{5i}$ and $\mathcal{T}_{6i}$. 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

$$\int_{\mathcal{T}_{5i} \cap \mathcal{T}_{6i}} \left(\frac{p_i}{p_{*i}} \right)^m d\hat{\lambda}_i dy_{i0} \leq (1 + c)^m \int_{\mathcal{T}_{5i} \cap \mathcal{T}_{6i}} d\hat{\lambda}_i dy_{i0} = \left(2 C_N' \right)^2 = o(N^\epsilon),$$

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

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

Notice that for $\psi_i(\lambda_j, Y_{j0}) \sim iid$ across $j \neq i$ with $|\psi_i(\hat{\lambda}_j, Y_{j0})| \leq M$ for some finite constant M. Then, by Bernstein's inequality [14] (e.g., Lemma 19.32 in van der Vaart (1998)),

$$
\begin{aligned}
&N\mathbb{P}_i\left\{ \hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4} \right\} \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) \\
&= N\mathbb{P}_i\left\{ \frac{1}{B_N^2(N-1)} \sum_{j\neq i}^{N} \left(\psi_i(\hat{\lambda}_j, Y_{j0}) - \mathbb{E}_i[\psi_i(\hat{\lambda}_j, Y_{j0})] \right) < -\frac{p_{*i}}{4} \right\} \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) \\
&\leq 2N\exp\left(-\frac{1}{4} \frac{B_N^4(N-1)p_{*i}^2/16}{\mathbb{E}_i[\psi_i(\hat{\lambda}_j, Y_{j0})^2] + MB_N^2 p_{i*}/4} \right) \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}).
\end{aligned}
$$

Using an argument similar to the proof of Lemma A.5 one can show that

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

In turn

$$
N\mathbb{P}_i\left\{ \hat{p}_i^{(-i)} - p_{*i} < -\frac{p_{*i}}{4} \right\} \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}) \leq 2\exp\left(-MNB_N^2 \frac{p_{*i}^2}{p_i + p_{*i}} + \ln N \right) \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i}).
$$

From Lemma A.6 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

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

---

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

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

Then, on the region $\mathcal{T}_{6i}$

$$
\begin{aligned}
N\mathbb{E}\Big[\mathbb{P}_i\Big\{\hat{p}_i^{(-i)} &- p_{*i} < -\frac{p_{*i}}{4}\Big\}\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\Big] \\
&\leq\ 2\mathbb{E}\left[\exp\left(-MNB_N^2\frac{p_{*i}^2}{p_i+p_{*i}}+\ln N\right)\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\right] \\
&\leq\ 2\mathbb{E}\big[\exp\left(-MNB_N^2 p_i+\ln N\right)\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})\big] \\
&\leq\ 2\exp\left(-MB_N^2 N^{\epsilon'}+\ln N\right) \\
&=\ o(N^\epsilon),
\end{aligned}
$$

as desired. ∎

A.3   Derivations for Section 6

A.3.1   Consistency of QMLE in Experiments 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

$$(\theta_*, \xi_*) = \mathrm{argmax}_{\theta, \xi}\, \mathbb{E}^{\mathcal{Y}}_{\theta_0} \left[\ln p(Y, X_2 | H, \theta, \xi) \right], \tag{A.40}$$

and verifying that $\theta_* = \theta_0$. Here, $p(y, x_2 | h, \theta, \xi)$ is given in (23). 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 $\phi = \mathrm{vec}(\Phi')$ and $\tilde{h}' = I \otimes h'$, such that we can write $\Phi h = \tilde{h}'\phi$. Using this notation, we can write

$$
\begin{aligned}
& \ln p(y, x_2 | h, \theta_1, \gamma, \phi, \omega) \tag{A.41}\\
&= \; C - \frac{1}{2}\ln |\Sigma(\gamma)| - \frac{1}{2}\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\gamma)\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big) \\
&\quad -\frac{1}{2}\ln |\underline{\Omega}| + \frac{1}{2}\ln |\bar{\Omega}(\gamma, \omega)| \\
&\quad -\frac{1}{2}\Big(\hat{\lambda}(\theta)'w'\Sigma^{-1}(\gamma)w\hat{\lambda}(\theta) + \phi'\tilde{h}\underline{\Omega}^{-1}\tilde{h}'\phi - \bar{\lambda}'(\theta,\xi)\bar{\Omega}^{-1}(\gamma,\omega)\bar{\lambda}(\theta,\xi)\Big),
\end{aligned}
$$

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, $w = \iota$, $\Sigma(\gamma) = \gamma I$, $x_2 = \emptyset$, $z = \emptyset$,

$h = [1, y_0]'$, $\underline{\Omega} = \omega^2$. Thus, splitting the $(T-1)(\ln \gamma^2)/2$, we can write

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

where

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

Note that

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

In turn, we can write

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

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

$$
\mathbb{E} \left[\ln p(Y|H, \rho, \gamma, \phi, \omega) \right] \tag{A.42}
$$
$$
= C - \frac{T-1}{2} \ln |\gamma^2| - \frac{1}{2\gamma^2} \mathbb{E} \left[\tilde{Y}(\rho)'(I - \iota\iota'/T)\tilde{Y}(\rho) \right] - \frac{1}{2} \ln |\omega^2 + \gamma^2/T|
$$
$$
- \frac{1}{2(\omega^2 + \gamma^2/T)} \Bigg(\left(\phi - \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\hat{\lambda}(\rho)] \right)' \mathbb{E}[\tilde{H}\tilde{H}'] \left(\phi - \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\hat{\lambda}(\rho)] \right)
$$
$$
- \mathbb{E}[\hat{\lambda}(\rho)\tilde{H}'] \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\hat{\lambda}(\rho)] + \mathbb{E}[\hat{\lambda}^2(\rho)] \Bigg).
$$

We deduce that

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

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,

$$\phi_*(\rho_0) = \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.44}$$

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

Plugging the pseudo-true value for ϕ into (A.42) yields the concentrated objective function

$$\mathbb{E}\big[\ln p(Y|H, \rho, \gamma, \phi_*(\rho), \omega)\big] \tag{A.45}$$
$$= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\mathbb{E}\big[\tilde{Y}(\rho)'(I - \iota\iota'/T)\tilde{Y}(\rho)\big]$$
$$- \frac{1}{2}\ln|\omega^2 + \gamma^2/T| - \frac{1}{2(\omega^2 + \gamma^2/T)}\big(\mathbb{E}[\hat{\lambda}^2(\rho)] - \mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big).$$

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

$$\gamma_*^2(\rho) = \frac{1}{T-1}\mathbb{E}\big[\tilde{Y}(\rho)'(I - \iota\iota'/T)\tilde{Y}(\rho)\big] \tag{A.46}$$
$$\omega_*^2(\rho) + \gamma_*^2(\rho)/T = \big(\mathbb{E}[\hat{\lambda}^2(\rho)] - \mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big).$$

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,

$$\gamma_*^2(\rho_0) = \gamma_0^2, \quad \omega_*^2(\rho_0) = \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.47}$$

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

$$\mathbb{E}\big[\ln p(Y|H, \rho, \gamma_*(\rho), \phi_*(\rho), \omega_*(\rho))\big] \tag{A.48}$$
$$= C - \frac{T-1}{2}\ln\big|\mathbb{E}\big[\tilde{Y}(\rho)'(I - \iota\iota'/T)\tilde{Y}(\rho)\big]\big|$$
$$- \frac{1}{2}\ln\big|\mathbb{E}[\tilde{Y}'(\rho)\iota\iota'\tilde{Y}(\rho)] - \mathbb{E}[\tilde{Y}'(\rho)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho)]\big|.$$

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.48) with respect to ρ yields

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

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

$$
\begin{aligned}
\text{F.O.C.}(\rho_0) \;=\; & (T-1)\mathbb{E}\big[X'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \hspace{3cm} \text{(A.49)}\\
& \times\left(\mathbb{E}[\tilde{Y}'(\rho_0)\iota\iota'\tilde{Y}(\rho_0)]-\mathbb{E}[\tilde{Y}'(\rho_0)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\right) \\
& +\mathbb{E}\big[\tilde{Y}(\rho_0)'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \\
& \times\left(\mathbb{E}[X'\iota\iota'\tilde{Y}(\rho_0)]-\mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\right).
\end{aligned}
$$

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.49) as follows:

$$
\begin{aligned}
\text{F.O.C.}(\rho_0) \quad &\hspace{6cm} (\text{A.50}) \\
&= -(T-1)\big(\mathbb{E}[X'\iota\iota'u]\big)\Big(\gamma_0^2 + T\big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big) \\
&\quad + \Big(\mathbb{E}[X'\iota\iota'u] + T\big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big)(T-1)\gamma_0^2 \\
&= T(T-1)\Big[\gamma_0^2\Big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big) \\
&\quad -\mathbb{E}[X'\iota\iota'u]\Big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big)\Big].
\end{aligned}
$$

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,

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

Thus, we obtain

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

Using these expressions, most terms that appear in (A.50) cancel out and the condition

simplifies to

$$\text{F.O.C.}(\rho_0) = T(T-1)\gamma_0 a_T \left(\mathbb{E}[Y_0 \lambda] - \mathbb{E}[Y_0 \tilde{H}'] \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\lambda] \right). \tag{A.51}$$

Now consider

$$\mathbb{E}[Y_0 \tilde{H}'] \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\lambda]$$

$$= \frac{1}{\mathbb{E}[Y_0^2] - (\mathbb{E}[Y_0])} \left[\begin{array}{cc} \mathbb{E}[Y_0] & \mathbb{E}[Y_0^2] \end{array} \right] \left[\begin{array}{cc} \mathbb{E}[Y_0^2] & -\mathbb{E}[Y_0] \\ -\mathbb{E}[Y_0] & 1 \end{array} \right] \left[\begin{array}{c} \mathbb{E}[Y_0] \\ \mathbb{E}[Y_0^2] \end{array} \right]$$

$$= \mathbb{E}[Y_0 \lambda].$$

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

$$\rho_* = \rho_0, \quad \gamma_*^2 = \gamma_0, \quad \phi_* = \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\lambda], \tag{A.52}$$

$$\omega_*^2 = \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda \tilde{H}'] \left(\mathbb{E}[\tilde{H}\tilde{H}'] \right)^{-1} \mathbb{E}[\tilde{H}\lambda]. \quad \blacksquare$$

A.3.2 Computation of the Oracle Predictor in Experiment 3

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

Scale Mixture. Let $a_{it} = 1$ if U_{it} is generated from the mixture component with variance γ_+^2 and $a_{it} = 0$ if U_{it} is generated from the mixture component with variance γ_-^2. Omitting i subscripts from now on, define

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

such that

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

Now let

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \tilde{Y}_t \sim N\left(\lambda, \bar{\gamma}^2(a_{1:T})/T \right),$$

where

$$\bar{\gamma}^2(a_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \gamma^2(a_t).$$

Under the prior distribution

$$\lambda|Y_0 \sim N(\phi_0 + \phi_1 Y_0, \underline{\Omega}),$$

we obtain a posterior distribution of the form

$$\lambda|(a_{1:T}, Y_{0:T}) \sim N\big(\bar{\lambda}(a_{1:T}), \bar{\Omega}(a_{1:T})\big), \tag{A.53}$$

where

$$
\begin{aligned}
\bar{\Omega}(a_{1:T}) &= \big(\underline{\Omega}^{-1} + T/\bar{\gamma}^2(a_{1:T})\big)^{-1} \\
\bar{\lambda}(a_{1:T}) &= \bar{\Omega}(a_{1:T})\big((\phi_0 + \phi_1 Y_0) + (T/\bar{\gamma}^2(a_{1:T}))\hat{\lambda}\big).
\end{aligned}
$$

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

$$\mathbb{P}\big(a_t = 1|\lambda, Y_{0:T}\big) \tag{A.54}$$

$$= \frac{p_u(\gamma_+)^{-1} \exp\left\{-\frac{1}{2\gamma_+^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\}}{p_u(\gamma_+)^{-1} \exp\left\{-\frac{1}{2\gamma_+^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\} + (1 - p_u)(\gamma_-)^{-1} \exp\left\{-\frac{1}{2\gamma_-^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\}}.$$

The posterior mean $\mathbb{E}[\lambda|\mathcal{Y}_i]$ can be approximated with the following Gibbs sampler. Generate a sequence of draws $\{\lambda^s, a_{1:T}^s\}_{s=1}^{N_{sim}}$ by iterating over the conditional distributions given in (A.53) and (A.54). Then,

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

Location Mixture. Let $a_{it} = 1$ if U_{it} is generated from the mixture component with mean μ_+ and $a_{it} = 0$ if U_{it} is generated from the mixture component with mean $-\mu_-$. Omitting i subscripts from now on, define

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

such that

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

Now let

$$\hat{\lambda}(a_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \tilde{Y}_t(a_t) \sim N\big(\lambda, \gamma^2/T\big).$$

Under the prior distribution

$$\lambda|Y_0 \sim N(\phi_0 + \phi_1 Y_0, \underline{\Omega}),$$

we obtain a posterior distribution of the form

$$\lambda|(a_{1:T}, Y_{0:T}) \sim N\big(\bar{\lambda}(a_{1:T}), \bar{\Omega}\big), \tag{A.56}$$

where

$$\bar{\Omega} = \big(\underline{\Omega}^{-1} + T/\gamma^2\big)^{-1}$$

$$\bar{\lambda}(a_{1:T}) = \bar{\Omega}\big((\phi_0 + \phi_1 Y_0) + (T/\gamma^2)\hat{\lambda}(a_{1:T})\big).$$

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

$$\mathbb{P}\big(a_t = 1|\lambda, Y_{0:T}\big) \tag{A.57}$$

$$= \frac{p_u \exp\left\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda - \mu_+)^2\right\}}{p_u \exp\left\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda - \mu_+)\right\} + (1 - p_u) \exp\left\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda + \mu_-)^2\right\}}.$$

The posterior mean $\mathbb{E}[\lambda|Y_{0:T}]$ can be approximated with the following Gibbs sampler. Generate a sequence of draws $\{\lambda^s, a_{1:T}^s\}_{s=1}^{N_{sim}}$ by iterating over the conditional distributions given in (A.56) and (A.57). Then,

$$\widehat{\mathbb{E}}[\lambda|Y_{0:T}] = \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \bar{\lambda}(a_{1:T}^s), \tag{A.58}$$

$$\widehat{\mathbb{V}}[\lambda|Y_{0:T}] = \left(\bar{\Omega} + \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \bar{\lambda}^2(a_{1:T}^s)\right) - \left(\frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \bar{\lambda}(a_{1:T}^s)\right)^2.$$

B   Data Set

The construction of our data is based on Covas, Rump, and Zakrajsek (2014). We down-loaded FR Y-9C BHC finanical statements for the years 2002 to 2014 using the web portal of the Federal Reserve Bank of Chicago. The financial statements are available at quarterly frequency. We define PPNR (relative to assets) as follows

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

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. Because some of the quarter-over-quarter changes in the income and expense flows are a reflection of accounting practices rather than economic conditions of the institutions, we aggregate the quarterly data to annual data. However, prior to the temporal aggregation we eliminate certain types of outliers. Before describing our outlier removal procedure, we briefly discuss the structure of the rolling samples used for the forecast evaluation.

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 year, with data from 2002 to 2014 we can construct $M = 9$ samples of size $T = 3$ with forecast origins running from $\tau = 2005$ to $\tau = 2013$. 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 adjustment procedure is applied to quarterly observations. After the sample adjustments we aggregate from quarterly to annual frequency by averaging the PPNR ratios over the four quarters of the calendar year. 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 the extreme cases $T = 3$ (short sample) and $T = 11$ (long sample). The column labeled N_0 provides the number of raw data for each sample. In columns N_j, $j = 1, \ldots, 4$, 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 900 to 1,400 institutions. Roughly 35% to 65% 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.

Table A-1: Size of Adjusted Rolling Samples

Sample							
T	τ	N_0	N_1	N_2	N_3	N_4	N
3	2005	6,731	2,629	882	580	580	551
3	2006	6,673	2,591	959	650	650	615
3	2007	6,619	2,537	1,024	693	693	655
3	2008	6,519	2,456	1,074	716	716	670
3	2009	6,399	1,281	1,139	693	693	653
3	2010	6,223	1,287	1,157	683	683	639
3	2011	6,518	1,396	1,273	704	704	656
3	2012	6,343	1,413	1,301	755	755	710
3	2013	6,154	1,407	1,291	772	771	725
11	2013	8,011	2,957	1,431	497	496	461

The "Adjustment Step" header spans columns N_0 through N.

Table A-2: Descriptive Statistics for Rolling Samples

Sample		Statistics						
T	τ	Min	Mean	Median	Max	StdD	Skew	Kurt
3	2005	-8.81	1.48	1.65	8.46	2.07	-0.80	5.36
3	2006	-7.61	1.50	1.54	8.46	1.95	-0.43	4.90
3	2007	-9.55	1.36	1.42	7.75	1.94	-0.61	5.51
3	2008	-9.55	1.12	1.22	7.75	1.93	-0.72	5.62
3	2009	-10.44	0.98	1.08	7.00	1.84	-0.82	6.01
3	2010	-7.46	0.87	0.96	6.60	1.74	-0.63	4.76
3	2011	-8.87	0.84	0.96	7.17	1.77	-0.70	5.04
3	2012	-7.65	0.79	0.90	7.81	1.86	-0.46	4.41
3	2013	-8.11	0.82	0.95	7.73	1.87	-0.53	4.62
11	2013	-8.89	1.15	1.23	7.00	1.82	-0.65	5.02

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.

Descriptive statistics for the $T = 3$ and $T = 11$ rolling samples are reported in Table A-1. For each rolling sample we pool observations across institutions and time periods. We do not weight the observations by the size of the institution. Focusing on the $T = 3$ samples, notice that the mean PPNR falls from about 1.5% for the 2005 and 2006 samples to 0.80% for the 2012 sample, which includes observations starting in 2009. In the 2013 sample the mean increased again to 1.15%. The means are generally smaller than the medians, suggesting that the samples are left-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 4.4 to 6.0.

C Additional Empirical Results

Table A-3: Parameter Estimates: $\hat{\theta}_{QMLE}$, Parametric Tweedie Correction

τ	$\hat{\rho}$	$\hat{\sigma}^2$	Intercept $\hat{\phi}_{10}$	$\hat{\phi}_{11}$	$\underline{\hat{\omega}}_1^2$	Unemployment $\hat{\phi}_{20}$	$\hat{\phi}_{21}$	$\underline{\hat{\omega}}_2^2$	N
2007	0.91	1.10	-0.99	0.08	4E-7	0.18	-0.01	9E-9	537
2008	0.86	1.09	-1.25	-0.05	3E-6	0.28	0.02	1E-7	598
2009	0.86	1.14	-0.27	-0.06	1E-7	0.05	0.02	5E-9	613
2010	0.86	1.14	-0.38	-0.03	2E-8	0.07	0.01	1E-9	606
2011	0.94	1.12	-0.22	-0.17	2E-7	0.03	0.02	3E-9	582
2012	0.94	1.12	0.01	-0.30	2E-8	0.00	0.03	1E-9	587
2013	0.93	1.12	-0.47	-0.30	3E-7	0.05	0.04	2E-9	608

Notes: Point estimates for the model $Y_{it+1} = \lambda_{1i} + \lambda_{2i}UR_t + \rho Y_{it} + U_{it+1}$, $U_{it+1} \sim N(0, \sigma^2)$, $\lambda_{ji}|Y_{i0} \sim N(\phi_{j0} + \phi_{j1}Y_{i0}, \underline{\omega}_j^2)$ for $j = 1, 2$. The time-series dimension of the estimation sample is $T = 5$.