2017

Point And Density Forecasts In Panel Data Models

Laura Liu
University of Pennsylvania, yuliu4@sas.upenn.edu

Point And Density Forecasts In Panel Data Models

Abstract
This dissertation develops econometric methods that facilitate estimation and improve forecasting performance in panel data models. The panel considered in this paper features large cross-sectional dimension (N) but short time series (T). It is modeled by a dynamic linear model with common and heterogeneous coefficients and cross-sectional heteroskedasticity. Due to short T, traditional methods have difficulty in disentangling the heterogeneous parameters from the shocks, which contaminates the estimates of the heterogeneous parameters. To tackle this problem, the methods developed in this dissertation assume that there is an underlying distribution of the heterogeneous parameters and pool the information from the whole cross-section together via this distribution. Chapter 2, coauthored with Hyungsik Roger Moon and Frank Schorfheide, constructs point forecasts using an empirical Bayes method that builds on Tweedie's formula to obtain the posterior mean of the heterogeneous coefficients under a correlated random effects distribution. 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. In Chapter 3, I focus on density forecasts and use a full Bayes approach, where the distribution of the heterogeneous coefficients is modeled nonparametrically allowing for correlation between heterogeneous parameters and initial conditions as well as individual-specific regressors. I develop a simulation-based posterior sampling algorithm specifically addressing the nonparametric density estimation of unobserved heterogeneous parameters. I prove that both the estimated common parameters and the estimated distribution of the heterogeneous parameters achieve posterior consistency, and that the density forecasts asymptotically converge to the oracle forecast. Monte Carlo simulations and an application to young firm dynamics demonstrate improvements in density forecasts relative to alternative approaches.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Economics

First Advisor
Francis X. Diebold

Second Advisor
Frank Schorfheide

Keywords
Bank Stress Tests, Bayesian, Density Forecasts, Panel Data, Point Forecasts, Young Firms Dynamics

Subject Categories
Economics

POINT AND DENSITY FORECASTS IN PANEL DATA MODELS

Laura Liu

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Francis X. Diebold
Professor of Economics

Co-Supervisor of Dissertation

Frank Schorfheide
Professor of Economics

Graduate Group Chairperson

Jesús Fernández-Villaverde
Professor of Economics

Dissertation Committee

Francis X. Diebold, Professor of Economics
Frank Schorfheide, Professor of Economics
Xu Cheng, Associate Professor of Economics
Francis J. DiTraglia, Assistant Professor of Economics

POINT AND DENSITY FORECASTS IN PANEL DATA MODELS

Yu Liu

ACKNOWLEDGEMENT

I am immensely indebted to my advisors, Francis X. Diebold and Frank Schorfheide, and other members of my committee, Xu Cheng and Francis J. DiTraglia. Their interesting lectures in the first year triggered my compassion in econometrics. As I went from the courses to the research, their invaluable advice and support guided me through step by step. Looking back, I cannot imagine myself making such progress and completing this dissertation without them. It is a blessing to have professors as insightful and caring as them, who are also role models for my future endeavor as an economist.

I am grateful to many other professors for their great lectures that broaden my view of economics and their patience with my every naive question. In particular, I greatly benefited from helpful discussions with Timothy Christensen, Benjamin Connault, and Jeremy Greenwood.

I would like to express my appreciation to my coauthors, Mert Demirer, Maria Grith, Christian Matthes, Hyungsik Roger Moon, Katerina Petrova, Kamil Yilmaz, and Molin Zhong. I have learned a lot from them beyond the scope of the projects, and hard work has become enjoyable experiences thanks to them.

I would like to thank all my friends, classmates, and everyone in the econometrics group. Thanks to Yunan Li for being my best friend since the first year of graduate school. Thanks to Nicolas Janetos, Ami Ko, and Jan Tilly for the good time and space we shared together and the support both in research and in life. Thanks to Ross Askanazi, Lorenzo Braccini, Minsu Chang, Pengfei Han, Yang Liu, Paul Sangrey, Minchul Shin, and Jacob Warren for all the thought-provoking conversations on econometrics and economics in general.

I am also very thankful to all the administrative staff in the department, especially Ms. Kelly Quinn. The life of a graduate student is not very easy, but their kindness, considerateness, and helpfulness have made it much easier.

Last but not least, I would like to extend my special thanks to my parents, Luming Liu and Guina Yu, for their unconditional love and continuous support, as well as my beloved husband, Evan Chan, for being there with me, inspiring me, and encouraging me all the time.

ABSTRACT

POINT AND DENSITY FORECASTS IN PANEL DATA MODELS

Laura Liu

Francis X. Diebold

Frank Schorfheide

This dissertation develops econometric methods that facilitate estimation and improve forecasting performance in panel data models. The panel considered in this paper features large cross-sectional dimension (N) but short time series (T). It is modeled by a dynamic linear model with common and heterogeneous coefficients and cross-sectional heteroskedasticity. Due to short T, traditional methods have difficulty in disentangling the heterogeneous parameters from the shocks, which contaminates the estimates of the heterogeneous parameters. To tackle this problem, the methods developed in this dissertation assume that there is an underlying distribution of the heterogeneous parameters and pool the information from the whole cross-section together via this distribution. Chapter 2, coauthored with Hyungsik Roger Moon and Frank Schorfheide, constructs point forecasts using an empirical Bayes method that builds on Tweedie's formula to obtain the posterior mean of the heterogeneous coefficients under a correlated random effects distribution. 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. In Chapter 3, I focus on density forecasts and use a full Bayes approach, where the distribution of the heterogeneous coefficients is modeled nonparametrically allowing for correlation between heterogeneous parameters and initial conditions as well as individual-specific regressors. I develop

a simulation-based posterior sampling algorithm specifically addressing the nonparametric density estimation of unobserved heterogeneous parameters. I prove that both the estimated common parameters and the estimated distribution of the heterogeneous parameters achieve posterior consistency, and that the density forecasts asymptotically converge to the oracle forecast. Monte Carlo simulations and an application to young firm dynamics demonstrate improvements in density forecasts relative to alternative approaches.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

This dissertation develops econometric methods that facilitate estimation and improve forecasting performance in panel data models. Panel data, such as a collection of firms or households observed repeatedly for a number of periods, are widely used in empirical studies and can be useful for forecasting individuals' future outcomes, which is interesting and important in many cases. For example, in the context of banks, stress tests involve forecasting pre-provision net revenues (PPNR) and other balance sheet variables under counterfactual stressed macroeconomic and financial scenarios; in the context of young firms, accurate forecasts can help investors select promising startups and assist policymakers in regulating entrepreneur funding.

For illustrative purposes, let us consider a simple dynamic panel data model:

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. The y_{it}'s are observed individual outcomes, β and σ^2 are common parameters, and λ_i's are unobserved individual effects. The general model studied in this dissertation extends this baseline setup to account for many important features of real-world empirical studies, including regressors with common effects, correlated random coefficients, and cross-sectional heteroskedasticity. Based on the observed panel up to time T, I am interested in providing point and density forecasts of $y_{i,T+1}$.

The panel considered in this paper features large cross-sectional dimension (N) but short time series (T). This framework is appealing to the bank stress tests example due to changes in the regulatory environment in the aftermath of the recent financial crisis as well as frequent mergers in the banking industry. It also fits the young firm dynamics example well as the

number of observations for each young firm is restricted by its age.

Due to short T, traditional methods have difficulty in disentangling the unobserved individual effects from the shocks, which contaminates the estimates of the individual effects. The naive estimators that only utilize the individual-specific observations are inconsistent even if N goes to infinity. To tackle this problem, the methods developed in this dissertation assume that there is an underlying distribution of the individual effects. Moreover, the individual effects are allowed to be correlated with the initial condition y_{i0}, i.e. correlated random effects model. Then, we can pool the information from the whole cross-section together via this distribution in an efficient and flexible way, and provide better estimates of the individual effects and more accurate forecasts of the individual-specific future outcomes.

The methods proposed in this dissertation are general to many other problems beyond forecasting. Here estimating heterogeneous parameters is important because we want to generate good forecasts, but in other cases, the heterogeneous parameters themselves can possibly be the objects of interest. For example, people may be interested in individual-specific treatment effects, and the technique developed here can be applied to those questions.

Chapter 2, coauthored with Hyungsik Roger Moon and Frank Schorfheide, constructs point forecasts using an empirical Bayes method that builds on Tweedie's formula to obtain the posterior mean of the 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). In other words, the regret of forecasts is negligible comparing to the part of the optimal risk that is due to uncertainty about the heterogeneous coefficients.

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. Results show that the impact of stressed macroeconomic conditions (characterized by unemployment, federal funds rate, and spread) on bank revenues is relatively small with respect to the cross-sectional dispersion of revenues.

In Chapter 3, I tackle a different problem in a similar panel data setup as described in Chapter 2. Instead of providing point forecasts via an empirical Bayes method, here I focus on density forecasts and use a full Bayes approach, where the distribution of the heterogeneous coefficients is modeled nonparametrically by a mixture model allowing for correlation between heterogeneous parameters and initial conditions as well as individual-specific regressors. Once this distribution is estimated by exploring the information from the whole cross-section, I can, intuitively speaking, use it as a prior distribution and combine it with individual-specific data and obtain the individual-specific posterior. This individual-specific posterior helps provide better inference about the heterogeneous parameters of each individual.

In this framework, it is natural to construct density forecasts. Basically, it is a predictive distribution of future performance of a specific firm, which summarizes all sources of future uncertainties. Especially, in this setup of dynamic panel data model, the density forecasts reflect uncertainties due to future shocks, individual heterogeneity, and estimation uncertainty, where the part of uncertainties due to individual heterogeneity arises from the lack of time-series information available to infer the heterogeneous parameters of each individual. Moreover, based on density forecasts, it is straightforward to derive point forecasts and interval forecasts.

I develop a simulation-based posterior sampling algorithm specifically addressing the nonparametric density estimation of unobserved heterogeneous parameters. I prove that both the estimated common parameters and the estimated distribution of the heterogeneous pa-

rameters achieve posterior consistency, and that the density forecasts asymptotically converge to the oracle forecast, an (infeasible) benchmark that is defined as the individual-specific posterior predictive distribution under the assumption that the common parameters and the distribution of the heterogeneous parameters are known.

Monte Carlo simulations demonstrate improvements in density forecasts relative to alternative approaches. There are three key factors for better density forecasts: in order of importance, nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients. An application to young firm dynamics also shows that the proposed predictor provides more accurate density predictions, and the estimated model helps shed light on the latent heterogeneity structure.

CHAPTER 2

Point Forecasts and Bank Stress Tests[1]

2.1 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:

$$Y_{it} = \lambda_i' W_{it-1} + \rho' X_{it-1} + \alpha' Z_{it-1} + U_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{2.1.1}$$

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: $X_{it-1} = Y_{it-1}$.

[1] This chapter builds on Liu *et al.* (2016), coauthored with Hyungsik Roger Moon and Frank Schorfheide.

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 gener-

alized 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 (2.1.1). 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 and Zhang (2009) to state our optimality result. Because our setup nests the linear dynamic panel data model, we utilize results on the consistent estimation of ρ in dynamic panel data models with fixed effects when T is small, e.g., Anderson and Hsiao (1981), Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998), Alvarez and Arellano (2003). 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 *et al.* (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 2.2 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 2.3. Section 2.4 discusses feasible implementation strategies for the predictor and we show in Section 2.5 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 2.6. The empirical application is presented in Section 2.7 and Section 2.8 concludes. Technical derivations, proofs, the description of the data set used in the empirical analysis, and further empirical results are relegated to the Appendix.

2.2 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 (2.1.1). 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 $W_{it} = (w_{1,t+1}, W_{2,it})$.[2] 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.

Our main goal is to construct optimal forecasts of $(Y_{1T+1}, ..., Y_{NT+1})$ conditional on the entire panel observations $\{(Y_{it}, W_{it-1}, X_{it-1}, Z_{it-1}),\ i = 1, \ldots, N$ and $t = 1, ..., T$ using the forecasting model (2.1.1). An important special case of model (2.1.1) is the basic dynamic panel data model

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \tag{2.2.1}$$

which is obtained by setting $W_{it} = 1$, $X_{it} = Y_{it}$ and $\alpha = 0$. The restricted model (2.2.1) 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 2.2.1 we specify the likelihood function for model (2.1.1) and in Section 2.2.2 we establish the identifiability of the model parameters, including the distribution of the heterogeneous coefficients λ_i.

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

2.2.1 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 2.2.1.

(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 $t = 1, \ldots, T$, *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* $t = 1, \ldots, T+1$ *and* $(V_{i1}, \ldots, V_{iT})$ *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 2.2.1(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 2.2.1(ii) formalizes the correlated random effects assumption. The subsequent Assumptions 2.2.1(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 (2.2.1) – and that the strictly exogenous predictors do not depend on

λ_i. In Assumption 2.2.1(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 2.2.1(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

$$p(y_i, x_{2,i}, \lambda_i|h_i, \theta) \propto \left(\prod_{t=1}^{T} \frac{1}{\sigma_t(h_i, \gamma_t)} \varphi\left(\frac{y_{it} - \lambda_i' w_{it-1} - \rho' x_{it-1} - \alpha' z_{it-1}}{\sigma_t(h_i, \gamma_t)}\right)\right) \pi(\lambda_i|h_i). \tag{2.2.2}$$

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 (2.2.2).

2.2.2 Identification

We now provide conditions under which the forecasting model (2.1.1) 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 (2.2.1) 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 $\mathbb{E}\big[(Y_{it}^* - \rho X_{it-1}^*)Y_{it-1}\big] = 0$ 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 $\Sigma_i^{1/2}(\tilde{\gamma}) = \operatorname{diag}\big(\sigma_1(h_i, \tilde{\gamma}_1), \ldots, \sigma_T(h_i, \tilde{\gamma}_T)\big)$, $S_i(\tilde{\gamma}) = \Sigma_i^{-1/2}(\tilde{\gamma})\iota$, and $M_i(\tilde{\gamma}) = I - S_i(S_i'S_i)^{-1}S_i'$. Using this notation, we obtain

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

This leads to the conditional moment condition

$$\mathbb{E}\big[M_i(\tilde{\gamma})\Sigma_i^{-1/2}(\tilde{\gamma})\big(Y_i - X_i\rho\big)\big(Y_i - X_i\rho\big)'\Sigma_i^{-1/2}(\tilde{\gamma})M_i'(\tilde{\gamma}) - M_i(\tilde{\gamma})\big|H_i\big] = 0 \tag{2.2.3}$$

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

$$\tilde{Y}_i = \Sigma_i^{-1/2}(\gamma)\big(Y_i - X_i\rho\big) = S_i(\gamma)\lambda_i + V_i. \tag{2.2.4}$$

The identification of $\pi(\lambda_i|h_i)$ can be established using a characteristic function argument similar to that in Arellano and Bonhomme (2012a). For the general model (2.1.1) we make the following assumptions:

Assumption 2.2.2.

(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 2.2.2(i) that the homogeneous parameters are identifiable.[3] 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 (2.1.1). Assumption 2.2.2(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 2.2.3. *Suppose that Assumptions 2.2.1 and 2.2.2 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 (2.1.1) are identified.*

2.3 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 2.3.1 provides a formal definition of the compound risk (expected loss). In Section 2.3.2 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 2.3.3 we introduce the concept of ratio optimality,

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

which describes forecasts that asymptotically (as $N \longrightarrow \infty$) attain the same risk as the oracle forecast.

2.3.1 Compound Risk

Let $L(\widehat{Y}_{iT+1}, Y_{iT+1})$ denote the loss associated with forecast $\hat{Y}_{i,T+1}$ 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

$$D_i = D_i(\mathcal{Y}^N) \in \{0, 1\}, \quad i = 1, \ldots, N, \tag{2.3.1}$$

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 $(Y_{1T}, ..., Y_{NT})$ 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

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_\theta^{\mathcal{Y}^N, \lambda^N, U_{T+1}^N} \left[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N) \right]. \tag{2.3.2}$$

We use the θ subscript for the expectation operator to indicate that the expectation is condi-

tional on θ.[4]. 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 $U_{T+1}^N = (U_{1T+1}, \ldots, U_{NT+1})$.

2.3.2 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

$$R_N(\widehat{Y}_{T+1}^N) = \mathbb{E}_{\theta}^{\mathcal{Y}^N}\left[\mathbb{E}_{\theta,\mathcal{Y}^N}^{\lambda^N, U_{T+1}^N}[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)]\right]. \tag{2.3.3}$$

The inner expectation can be interpreted as posterior risk, which is obtained by conditioning on the observations $\mathcal{Y}^N$ and integrating over the heterogeneous parameter λ^N and the shocks U_{T+1}^N. The outer expectation averages over the possible trajectories $\mathcal{Y}^N$.

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

$$\begin{aligned} &\mathbb{E}_{\theta,\mathcal{Y}^N}^{\lambda^N, U_{T+1}^N}[L_N(\widehat{Y}_{T+1}^N, Y_{T+1}^N)] \\ &= \sum_{i=1}^{N} D_i(\mathcal{Y}^N)\left\{\left(\widehat{Y}_{iT+1} - \mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i, U_{iT+1}}[Y_{iT+1}]\right)^2 + \mathbb{V}_{\theta,\mathcal{Y}_i}^{\lambda_i, U_{iT+1}}[Y_{iT+1}]\right\} \end{aligned} \tag{2.3.4}$$

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

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

predictor. Because U_{iT+1} is mean-independent of λ_i and $\mathcal{Y}_i$, we obtain

$$\widehat{Y}_{iT+1}^{opt} = \mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i, U_{iT+1}}[Y_{iT+1}] = \mathbb{E}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i]' W_{iT} + \rho' X_{iT} + \alpha' Z_{iT}. \tag{2.3.5}$$

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 2.3.1 (Optimal Forecast)**.** *Suppose Assumptions 2.2.1 are satisfied. The optimal forecast that minimizes the composite risk in (2.3.2) is given by $\widehat{Y}_{iT+1}^{opt}$ in (2.3.5). The compound risk of the optimal forecast is*

$$R_N^{opt} = \mathbb{E}_\theta^{\mathcal{Y}^N}\left[\sum_{i=1}^N D_i(\mathcal{Y}^N)\left(W_{iT}' \mathbb{V}_{\theta,\mathcal{Y}_i}^{\lambda_i}[\lambda_i] W_{iT} + \sigma_{T+1}^2(H_i, \gamma_{T+1})\right)\right]. \tag{2.3.6}$$

According to (2.3.6), 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.

2.3.3 Ratio Optimality

The identification result presented in Section 2.2.2 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 2.3.2. For a given $\epsilon_0 > 0$, we say that forecast $\widehat{Y}_{T+1}^N$ achieves ϵ_0-ratio optimality,

if

$$\limsup_{N\to\infty} \frac{R_N(\widehat{Y}^N_{T+1}) - R^{\text{opt}}_N}{\mathbb{E}^{\mathcal{Y}^N}_\theta \left[\sum_{i=1}^N D_i(\mathcal{Y}^N) W'_{iT} \mathbb{V}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i] W_{iT}\right] + N^{\epsilon_0}} \leq 0. \tag{2.3.7}$$

Using (2.3.5), the risk differential in the numerator (called regret) can be written as

$$R_N(\widehat{Y}^N_{T+1}) - R^{\text{opt}}_N = \mathbb{E}^{\mathcal{Y}^N}_\theta \left[\sum_{i=1}^N D_i(\mathcal{Y}^N) \left(\widehat{Y}_{iT+1} - \mathbb{E}^{\lambda_i, U_{iT+1}}_{\theta,\mathcal{Y}_i}[Y_{iT+1}]\right)^2\right]. \tag{2.3.8}$$

For illustrative purposes, Consider the basic dynamic panel data model (2.2.1). For this model $\mathbb{E}^{\lambda_i,U_{iT+1}}_{\theta,\mathcal{Y}_i}[Y_{iT+1}] = \mathbb{E}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i] + \rho Y_{iT}$. A natural class of predictors is given by $\widehat{Y}_{iT+1} = \widehat{\mathbb{E}}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i] + \hat{\rho} Y_{iT}$, where $\widehat{\mathbb{E}}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i]$ 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 $\sum_{i=1}^N (\hat{\rho}-\rho)^2 Y^2_{iT} = O_p(1)$. Thus, whether a predictor attains ratio optimality crucially depends on the rate at which the discrepancy between $\mathbb{E}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i]$ and $\widehat{\mathbb{E}}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i]$ vanishes.

The denominator of the ratio in Definition 2.3.2 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 $\widehat{\mathbb{E}}^{\lambda_i}_{\mathcal{Y}_i}[\lambda_i]$ 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 (2.2.1) would be $\widehat{Y}^{FD}_{iT+1}(\rho) = Y_{iT} + \rho(Y_{iT} - Y_{iT-1})$, which is different from $\widehat{Y}^{opt}_{iT+1}$ in (2.3.5). Thus, we can immediately deduce from Theorem 2.3.1 that $\widehat{Y}^{FD}_{iT+1}(\rho)$ is not an optimal forecast. The quasi-differencing of Y_{it} introduces a predictable moving-average error term that is ignored

by the predictor $\widehat{Y}^{FD}_{iT+1}(\rho)$.

2.4 Implementation of the Optimal Forecast

We will construct a consistent approximation of the posterior mean $\mathbb{E}^{\lambda_i, U_{iT+1}}_{\theta, \mathcal{Y}^i}[\lambda_i]$ 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 2.4.1. In Section 2.4.2 we discuss the parametric estimation of the correction term and in Section 2.4.3 we consider a nonparametric kernel-based estimation. The QMLE and Generalized Method-of-Moments (GMM) estimation of the parameter θ are discussed in Sections 2.4.4 and 2.4.5.

2.4.1 Tweedie's Formula

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

Assumption 2.4.1. *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 2.4.1 we will now further manipulate the density $p(y_i, x_{2,i}, \lambda_i | h_i, \theta)$ in (2.2.2).[5]

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

To simplify the notation we will drop the i subscript. Define

$$\tilde{y}_t(\theta) = y_t - \rho' x_{t-1} - \alpha' z_{t-1}, \quad \Sigma(\theta) = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2), \tag{2.4.1}$$

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. Replacing $\varphi(v)$ in (2.2.2) with a Gaussian density function we obtain:

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

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

$$\hat{\lambda} = (w' \Sigma^{-1} w)^{-1} w' \Sigma^{-1} \tilde{y} \tag{2.4.2}$$

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

$$p(\hat{\lambda} | \lambda, h, \theta) = (2\pi)^{-k_w/2} |w' \Sigma^{-1} w|^{1/2} \exp\left\{-\frac{1}{2}(\hat{\lambda} - \lambda)' w' \Sigma^{-1} w (\hat{\lambda} - \lambda)\right\}. \tag{2.4.3}$$

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

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

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

Theorem 2.4.2. *Suppose that Assumptions 2.2.1 and 2.4.1 hold. The posterior mean of λ_i has the representation*

$$\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i]=\hat{\lambda}_i(\theta)+\left(W_i^{0:T-1\prime}\Sigma^{-1}(\theta)W_i^{0:T-1}\right)^{-1}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\theta)|H_i,\theta). \tag{2.4.4}$$

The optimal forecast is given by

$$\begin{aligned} \widehat{Y}^{opt}_{iT+1}(\theta) &= \left(\hat{\lambda}_i(\theta)+\left(W_i^{0:T-1\prime}\Sigma^{-1}(\theta)W_i^{0:T-1}\right)^{-1}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\theta)|H_i,\theta)\right)'W_{T+1} \\ &\quad +\rho'X_{iT}+\alpha'Z_{iT}. \end{aligned} \tag{2.4.5}$$

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 2.4.2 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 (2.4.4) 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.

2.4.2 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

$$\lambda_i|(H_i, \theta) \sim N\big(\Phi H_i, \underline{\Omega}\big). \tag{2.4.6}$$

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

$$\begin{aligned} p\big(\hat{\lambda}(\theta)\big|h, \theta, \xi\big) &= \int p\big(\hat{\lambda}(\theta)|\lambda, h, \theta\big)\pi(\lambda|h, \xi)d\lambda \\ &= (2\pi)^{-k_w/2}\big|\underline{\Omega}^{-1}\big|^{1/2}\big|w'\Sigma^{-1}w\big|^{1/2}\big|\bar{\Omega}\big|^{1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\big(\hat{\lambda}'w'\Sigma^{-1}w\hat{\lambda} + h'\Phi'\underline{\Omega}^{-1}\Phi h - \bar{\lambda}'\bar{\Omega}^{-1}\bar{\lambda}\big)\right\}. \end{aligned} \tag{2.4.7}$$

Here, we used the likelihood of $\hat{\lambda}$ in (2.4.3), the density associated with the Gaussian prior in (2.4.6), 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

$$\hat{\xi}(\theta) = \text{argmax}_{\xi} \prod_{i=1}^{N} p(\hat{\lambda}_i(\theta)|h_i, \theta, \xi) \tag{2.4.8}$$

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.

2.4.3 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.,

$$\begin{aligned} &\hat{p}(\hat{\lambda}_i(\theta), h_i|\theta) \qquad (2.4.9)\\ &= \frac{1}{N}\sum_{j=1}^{N}\Bigg[(2\pi)^{-k_w/2}|B_N|^{-k_w}|V_{\hat{\lambda}}|^{-1/2}\exp\left\{-\frac{1}{2B_N^2}\big(\hat{\lambda}_i(\theta)-\hat{\lambda}_j(\theta)\big)'V_{\hat{\lambda}}^{-1}\big(\hat{\lambda}_i(\theta)-\hat{\lambda}_j(\theta)\big)\right\}\\ &\quad\times(2\pi)^{-k_h/2}|B_N|^{-k_h}|V_h|^{-1/2}\exp\left\{-\frac{1}{2B_N^2}\big(h_i-h_j\big)'V_h^{-1}\big(h_i-h_j\big)\right\}\Bigg], \end{aligned}$$

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 (2.4.2) 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 2.5 we provide a detailed analysis of the Gaussian kernel estimator in the context of the basic dynamic panel data model in (2.2.1) with time-homoskedastic innovations.

2.4.4 QMLE Estimation of θ

Notice that under Assumption 2.4.1, $\hat{\lambda}_i(\theta)$ in (2.4.2) 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)

$$
\begin{aligned}
& p(y,x_2|h,\theta,\xi) \qquad (2.4.10)\\
&= \int p(y,x_2|h,\theta,\lambda)\pi_\lambda(\lambda|h,\hat{\xi}(\theta))d\lambda \\
&\propto |\Sigma(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}\big(\tilde{y}(\theta)-w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\theta)\big(\tilde{y}(\theta)-w\hat{\lambda}(\theta)\big)\right\} \\
&\quad \times \int \exp\left\{-\frac{1}{2}\big(\hat{\lambda}(\theta)-\lambda\big)'w'\Sigma^{-1}(\theta)w\big(\hat{\lambda}(\theta)-\lambda\big)\right\}\pi_\lambda\big(\lambda(\theta)|h,\hat{\xi}(\theta)\big)d\lambda \\
&\propto |\Sigma(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}\big(\tilde{y}(\theta)-w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\theta)\big(\tilde{y}(\theta)-w\hat{\lambda}(\theta)\big)\right\} \\
&\quad \times\big|w'\Sigma^{-1}w\big|^{-1/2}p(\hat{\lambda}(\theta)|h,\theta,\xi).
\end{aligned}
$$

Here, we used the definition of $\tilde{y}(\theta)$ in (2.4.1) and the product of Gaussian likelihood and prior in (2.4.2). Note that the term $p(\hat{\lambda}(\theta)|h,\theta,\xi)$ in the last line of (2.4.10) is identical to the objective function for ξ used in (2.4.8). Thus, we can now jointly determine θ and ξ by maximizing the integrated likelihood as a function:

$$
\big(\hat{\theta}_{QMLE},\hat{\xi}_{QMLE}\big) = \text{argmax}_{\theta,\xi}\prod_{i=1}^{N} p(y_i,x_{2i}|h_i,\theta,\xi). \qquad (2.4.11)
$$

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

2.4.5 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 2.2. For $t = 1, \ldots, T - k_w$, define

$$Y_{it}^* = Y_{it} - \left(\sum_{s=t+1}^{T} Y_{is} W_{is-1}' \right) \left(\sum_{s=t+1}^{T} W_{is-1} W_{is-1}' \right)^{-1} W_{it-1}. \tag{2.4.12}$$

Moreover, define X_{it-1}^* and Z_{it-1}^* by replacing $Y_{i\cdot}$ in (2.4.12) with $X_{i\cdot}$ and $Z_{i\cdot}$, respectively, and let

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

The continuous-updating GMM estimator of ρ and α solves

$$(\hat{\rho}_{GMM}, \hat{\alpha}_{GMM}) = \underset{\rho, \alpha}{\operatorname{argmin}} \left(\sum_{i=1}^{N} g_i(\rho, \alpha) \right)' \left(\sum_{i=1}^{N} g_i(\rho, \alpha) g_i(\rho, \alpha)' \right)^{-1} \left(\sum_{i=1}^{N} g_i(\rho, \alpha) \right) \tag{2.4.13}$$

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 2.6) and the empirical application (Section 2.7).[6]

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

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

where $\hat{\rho}$ and $\hat{\alpha}$ could be the estimators in (2.4.13). We use the sample analogue to a set of moment condition implied by a generalization of (2.2.3):

$$\begin{aligned} \hat{\gamma}_{GMM} = \operatorname{argmin}_\gamma \frac{1}{N} \sum_{i=1}^{N} \Big\| B \operatorname{vec} \Big(& M_i(\gamma) \Sigma_i^{-1/2}(\gamma) \tilde{Y}_i(\hat{\rho}, \hat{\alpha}) \\ & \times \tilde{Y}_i'(\hat{\rho}, \hat{\alpha}) \Sigma_i^{-1/2}(\gamma) M_i(\gamma) - M_i(\gamma) \Big) \Big\|^2, \end{aligned} \tag{2.4.14}$$

[6] 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).

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.

2.4.6 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. 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

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

As in the case of the one-step-ahead forecasts, the posterior mean $\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i]$ can be replaced by an approximation based on Tweedie's formula and the ρ's can be replaced by consistent estimates. A model with additional covariates would require external multi-step forecasts of the covariates, or the specification in (2.1.1) would have to be modified such that all exogenous regressors appear with an h-period lag.

2.5 Ratio Optimality in the Basic Dynamic Panel Model

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

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \quad U_{it} \sim iidN(0, \sigma^2), \quad (\lambda_i, Y_{i0}) \sim \pi(\lambda, y_{i0}). \tag{2.5.1}$$

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 (2.5.1), the sufficient statistic is given by

$$\hat{\lambda}_i(\rho) = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \rho Y_{it-1}) \tag{2.5.2}$$

and the posterior mean of λ_i simplifies to

$$\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i] = \mu\big(\hat{\lambda}_i(\rho), \sigma^2/T, p(\hat{\lambda}_i, Y_{i0})\big) = \hat{\lambda}_i(\rho) + \frac{\sigma^2}{T}\frac{\partial}{\partial\hat{\lambda}_i(\theta)}\ln p(\hat{\lambda}_i(\rho), Y_{i0}). \tag{2.5.3}$$

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 (2.5.4) 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 (2.4.9) by a leave-one-out estimator of the form:

$$\hat{p}^{(-i)}(\hat{\lambda}_i(\rho), Y_{i0}) = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho) - \hat{\lambda}_i(\rho)}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0} - Y_{i0}}{B_N}\right), \tag{2.5.4}$$

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:

$$\begin{aligned}\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] &= \int \frac{1}{\sqrt{\sigma^2/T + B_N^2}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{\sigma^2/T + B_N^2}}\right) \\ &\quad \times \left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0} - \tilde{y}_{i0}}{B_N}\right) p(\tilde{y}_{i0}|\lambda_i) d\tilde{y}_{i0}\right] p(\lambda_i) d\lambda_i. \end{aligned} \tag{2.5.5}$$

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 (2.5.5). 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

$$\widehat{Y}_{iT+1} = \left[\mu\big(\hat{\lambda}_i(\hat{\rho}), \hat{\sigma}^2/T + B_N^2, \hat{p}^{-i}(\cdot)\big)\right]^{C_N} + \hat{\rho} Y_{iT}, \tag{2.5.6}$$

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

Assumption 2.5.1 (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 2.5.2 (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 2.5.3 (Conditional distribution of $Y_{i0}|\lambda_i$)**.** *Let $\mathcal{Y}^\pi_\lambda$ 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 $y \in \mathcal{Y}^\pi_\lambda$ 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* $\lambda \in \Lambda^\pi$*:*

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

Assumption 2.5.4 (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 2.5.1 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 2.5.2 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.[7]

Assumption 2.5.3 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 (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

[7] In a nutshell, we need to control the behavior of $\hat{p}(\hat{\lambda}_i, Y_{i0})$ and its derivative uniformly, which, in certain steps of the proof, requires us to consider bounds of the form M/B_N^2, where M is a generic constant. If the bandwidth shrinks too fast, the bounds diverge too quickly to ensure that it suffices to standardize the regret in Definition 2.3.2 by N^{ϵ_0} if the λ_i coefficients are identical for each cross-sectional unit.

verify in the Appendix that a $\pi(y|\lambda)$ that satisfies Assumption 2.5.3 is $\pi(y|\lambda) = \phi(y - \lambda)$, where $\phi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$. Finally, Assumption 2.5.4 postulates the existence of finite sample moments of the estimators of the common parameter. The main result is stated in the following theorem:

Theorem 2.5.5. *Suppose that Assumptions 2.2.1, 2.4.1, and 2.5.1 to 2.5.4. Then, for the basic dynamic panel model the predictor $\widehat{Y}_{iT+1}$ defined in (2.5.6) satisfies the ratio optimality in Definition 2.3.2.*

The result in Theorem 2.5.5 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 2.3.2) in cases in which the intercepts are identical across units and $\pi(\lambda)$ is a pointmass.

2.6 Monte Carlo Simulations

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

2.6.1 Experiment 1: Gaussian Random Effects Model

The first Monte Carlo experiment is based on the basic dynamic panel data model in (2.2.1). 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

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: $N_{sim} = 1,000$

units and the number of time periods is $T = 3$. Generally, the smaller T relative to number 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 (2.3.5).

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 (2.4.11). Tweedie's formula (see (2.5.3) for the simplified version) is evaluated based on $p\big(\hat{\lambda}_i(\hat{\theta}_{QMLE}) | y_{i0}, \hat{\theta}_{QMLE}, \hat{\xi}_{QMLE}\big)$.

Posterior Predictive Mean Approximation Based on GMM Estimator. We use the Arellano-Bover estimator described in Section 2.4.5. The estimator for ρ is given by (2.4.13) and the estimator for γ by (2.4.14). 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

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

The estimator $\hat{\xi}(\hat{\theta}_{GMM})$ is obtained from (2.4.8). Finally, Tweedie's formula is evaluated based on $p\big(\hat{\lambda}_i(\hat{\theta}_{GMM}) | y_{i0}, \hat{\theta}_{GMM}, \hat{\xi}(\hat{\theta}_{GMM})\big)$.

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:

$$\hat{\rho}_L = \text{argmin}_\rho \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(Y_{it} - \rho Y_{it-1} - \hat{\lambda}_i(\rho)\big)^2, \quad \hat{\lambda}_i(\rho) = \frac{1}{T} \sum_{t=1}^{T} Y_{it} - \rho Y_{it-1}. \tag{2.6.1}$$

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

$$(\hat{\rho}_P, \hat{\lambda}_P) = \text{argmin}_{\rho,\lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(Y_{it} - \rho Y_{it-1} - \lambda\big)^2. \tag{2.6.2}$$

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 $\widehat{Y}^{FD}_{iT+1}(\hat{\rho}_{GMM})$.

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 2.3.2 (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 $\widehat{e}_{iT+1|T} = Y_{iT+1} - \widehat{Y}_{iT+1}$ 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.

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

	All Units		Bottom Group		Middle Group		Top Group	
Estimator / Predictor	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 2.3.2.

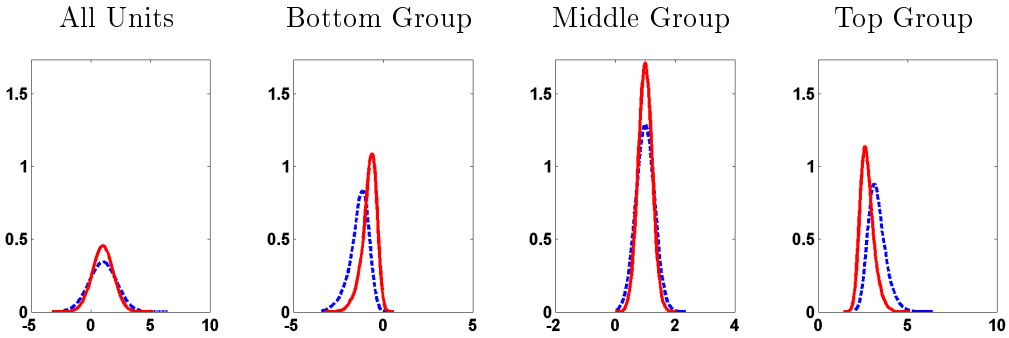
The density of $\hat{\lambda}_i$ is estimated parametrically, using a family of distributions that nests the true random effects distribution. Because it is based on a correctly specified likelihood function, the predictor based on $\hat{\theta}_{QMLE}$ performs slightly better than the predictor based 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 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

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.

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 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 $\widehat{\mathbb{E}}^{\lambda_i}_{\hat{\theta},\mathcal{Y}_i}[\lambda_i]$ 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.

2.6.2 Experiment 2: Non-Gaussian Correlated Random Effects Model

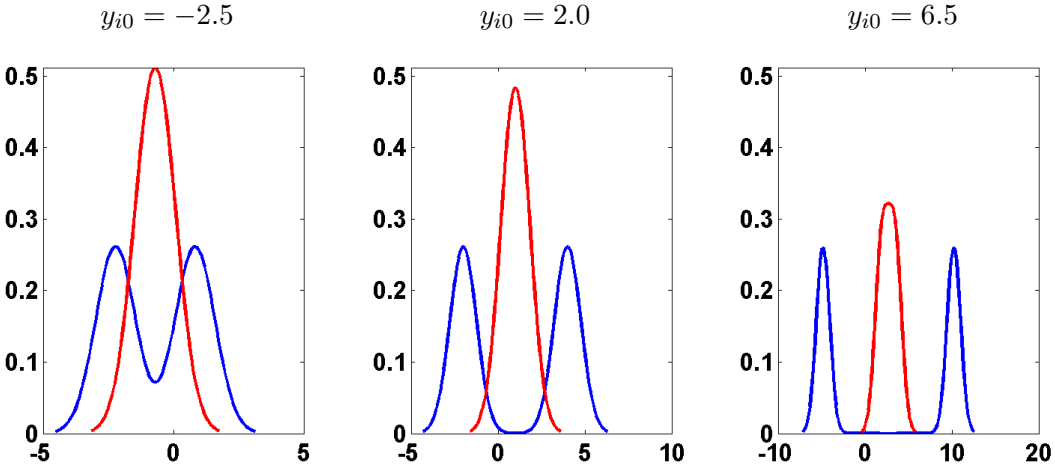
We now change the Monte Carlo design in two dimensions. First, we replace the Gaussian random effects specification with a non-Gaussian specification in which the heterogeneous coefficient λ_i is correlated with the initial condition Y_{i0}. Second, we consider a Tweedie correction based on a kernel density estimate of $p(\hat{\lambda}_i|Y_{i0})$ as discussed in Section 2.4.3.

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

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:
$\lambda_i \vert Y_{i0} \sim \begin{cases} N\big(\phi_+(Y_{i0}), \underline{\Omega}\big) & \text{with probability } p_\lambda \\ N\big(\phi_-(Y_{i0}), \underline{\Omega}\big) & \text{with probability } 1 - p_\lambda \end{cases}$,
$\phi_+(Y_{i0}) = \phi_0 + \delta + (\phi_1 + \delta) Y_{i0}$,
$\phi_-(Y_{i0}) = \phi_0 - \delta + (\phi_1 - \delta) Y_{i0}$,
$\underline{\Omega} = \left[\frac{1}{(1-\rho)^2} V_Y^{-1} + \underline{V}_\lambda^{-1}\right]^{-1}$, $\phi_0 = \underline{\Omega}\, \underline{V}_\lambda^{-1} \underline{\mu}_\lambda$, $\phi_1 = \frac{1}{1-\rho} \underline{\Omega} V_Y^{-1}$,
$p_\lambda = 1/2$, $\delta \in \{1/5,\ 1,\ 5\}$ $(\delta = 1/\sqrt{\kappa})$
Sample Size: $N = 1,000$, $T = 3$
Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

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.

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).

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 (2.4.9). The bandwidth is

chosen in accordance with the theory in Section 2.5. 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 $c \in \{1/2, 1, 2\}$.[8] Second, we use the adaptive estimator proposed by Botev *et al.* (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.[9] Unless otherwise noted, the subsequent results are based on the BGK estimator.

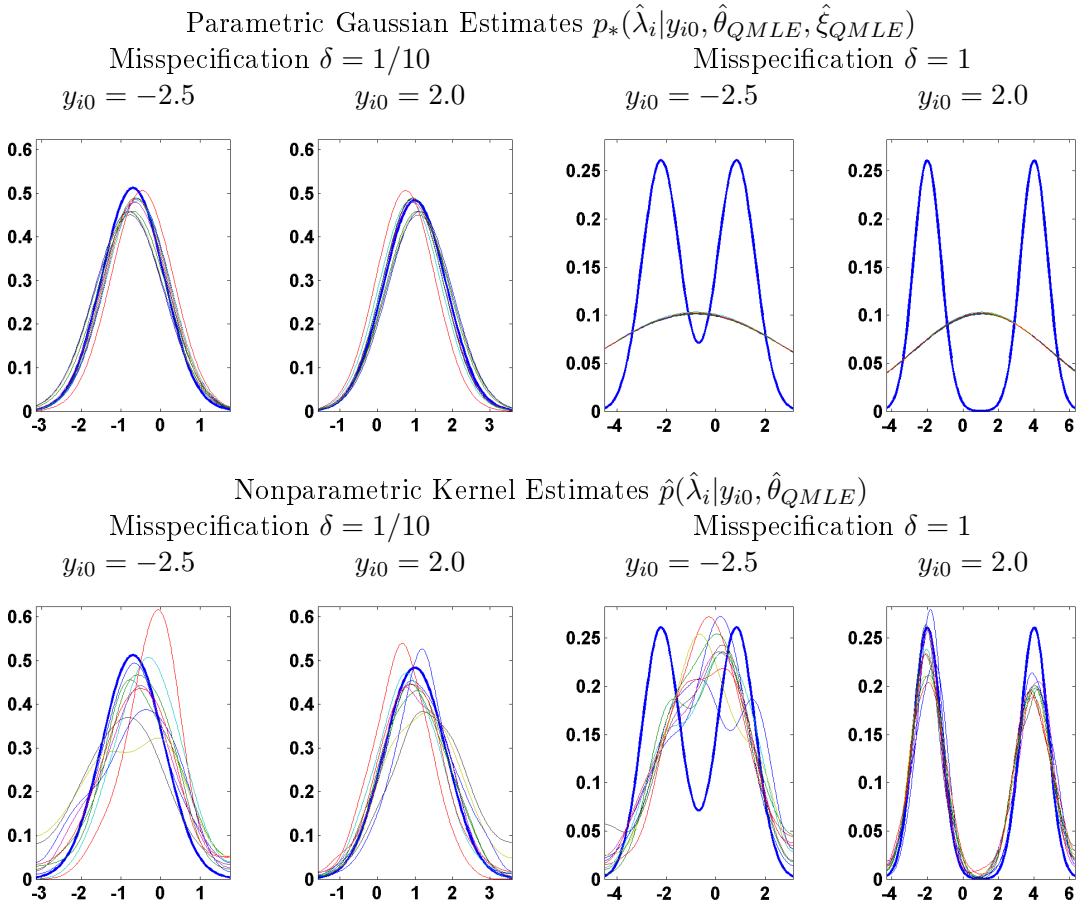
Figure 3 shows the "true" density $p(\hat{\lambda}_i|y_{i0}, \theta)$ as well as Gaussian and nonparametric approximations. Under the Gaussian correlated random effects distribution we can directly calculate the conditional distribution of $\hat{\lambda}_i$ given y_{i0}. The nonparametric approximation 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 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 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

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

[9] 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

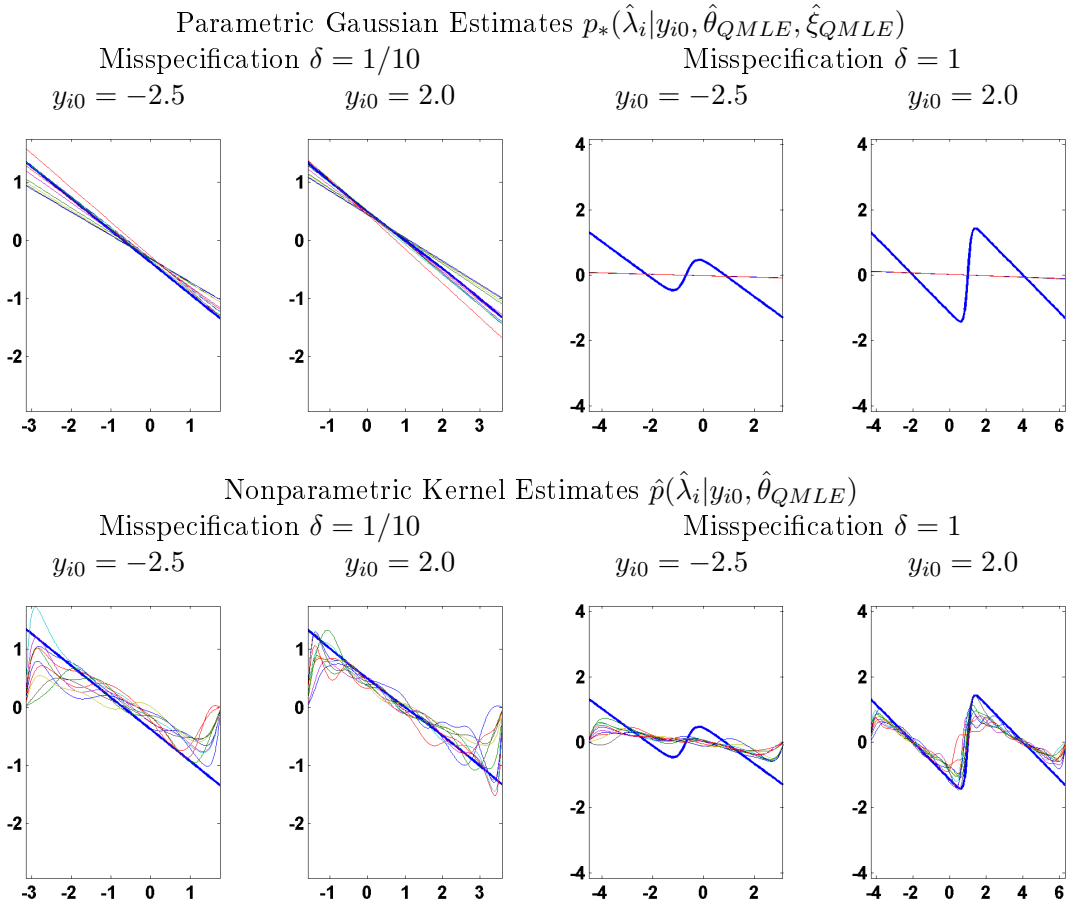


Notes: Solid (blue) lines depict "true" $p(\hat{\lambda}_i|y_{i0},\theta)$. Colored "hairs" depict 10 estimates from the Monte Carlo repetitions. The nonparametric estimates are based on the BGK kernel estimator. The Monte Carlo design is described in Table 3.

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 $\hat{\lambda}_i(\hat{\theta})$.

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



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.

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 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

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

	All Units		Bottom Group		Top Group	
Estimator / Predictor	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 2.3.2. The BGK estimator relies on a adaptive bandwidth choice. For the Gaussian kernel estimator in (2.4.9) we set $B_N = c/(\ln N)^{0.49}$.

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.

2.6.3 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.

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

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

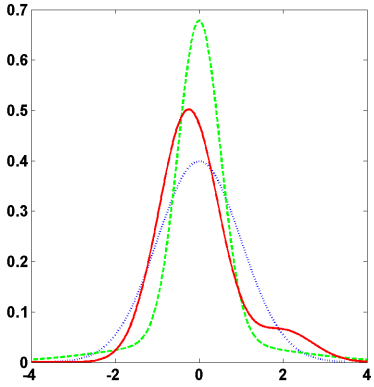
$\gamma^2 = 1 - p_u(\mu_u^+)^2 - (1 - p_u)(\mu_u^-)^2 = 1/2$

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).

We consider a scale mixture that generates excess kurtosis and a location mixture that generates skewness. The innovation distributions are normalized such that $\mathbb{E}[U_{it}] = 0$ and $\mathbb{V}[U_{it}] = 1$. 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:[10] 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.

[10] 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.

Table 6: Monte Carlo Experiment 3: Misspecified Likelihood Function

	All Units		Bottom Group		Top Group	
Estimator / Predictor	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 2.3.2.

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 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.

2.7 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 2.7.1 we use the basic dynamic panel data model to generate PPNR forecasts. In Section 2.7.2 we extend the model to include covariates and compare forecasts under the actual realization of the covariates and stressed scenarios in which we set the covariantes to

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.

counterfactual levels.

2.7.1 Results from the Basic Dynamic Panel Model

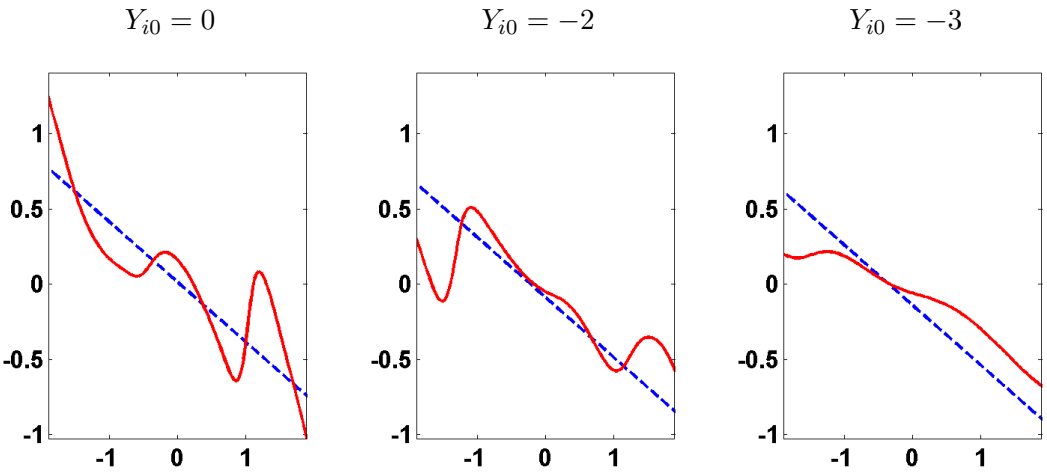
We begin by evaluating forecasts from the basic dynamic panel model in (2.5.1). The parametric Tweedie correction is based on $\lambda_i|(H_i, \theta) \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\omega}^2)$. The forecast evaluation criterion is the mean-squared error (MSE) computed across institutions and across time:

$$MSE = \frac{1}{M} \sum_{\tau=\tau_1}^{\tau_1+M-1} \left(\frac{\frac{1}{N_\tau} \sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau}) \big(Y_{i\tau+1} - \widehat{Y}_{i\tau+1}\big)^2}{\frac{1}{N_\tau} \sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau})} \right), \tag{2.7.1}$$

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, in particular for small

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 $\hat{\theta}_{QMLE}$.

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

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.

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 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 23 in the Appendix provides point estimates of the parameters of the basic dynamic panel model and the parametric correlated random effects distribution for $T = 5$

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

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 estimate $\hat{\phi}_1$ generates positive correlation between λ_i and Y_{i0}. The intercept of the correlated random effects distribution drops during the Great Recession[11], 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.

2.7.2 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

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

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 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 (UNRATE) 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.[12]

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

Table 10: MSE for Model with Unemployment 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)	0.88	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.

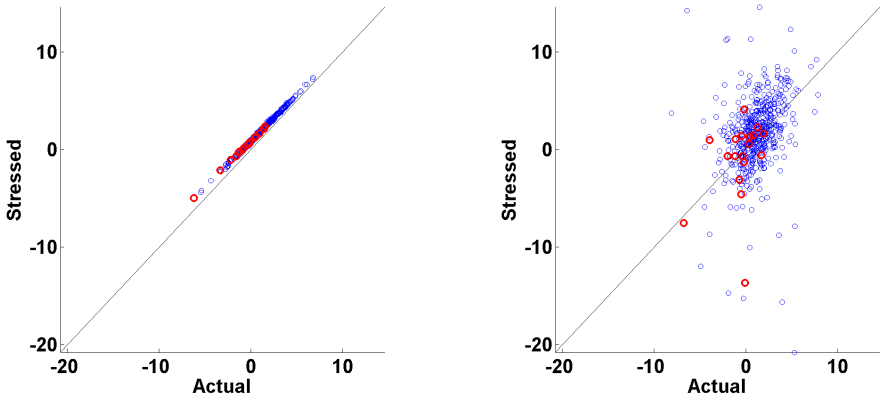
We now impose stress by increasing the unemployment rate by 5%. This corresponds to the unemployment movement in the *severely adverse* macroeconomic scenario in the Federal Reserve's CCAR 2016. In Figure 6 we are comparing one-year-ahead predictions for forecast origins $\tau = 2007$ and $\tau = 2012$ 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[13] 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.

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

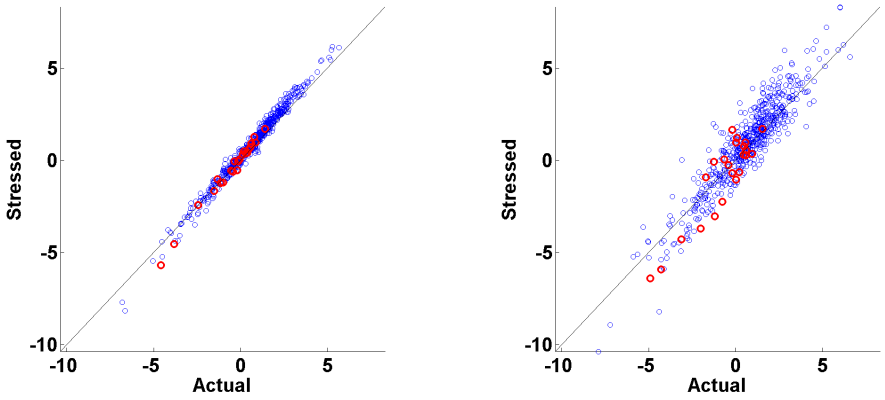
Figure 6: Predictions under Actual and Stressed Scenario for $T = 5$

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).

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: 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 $\tau = 2013$.

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.[14] 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.

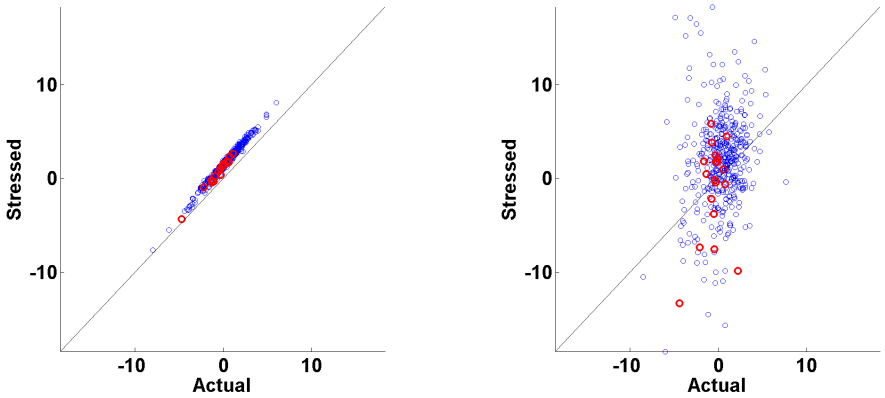
Figure 7 compares PPNR predictions under the actual macroeconomic conditions and a stressed macroeconomic scenario. The stressed scenario comprises an increase in the unemployment rate by 5% (as before) and an increase in nominal interest rates and spreads by 5%. This scenario could be interpreted as an aggressive monetary tightening that induced a sharp drop in macroeconomic activity. 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

[14] 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 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).

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.

2.8 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.

CHAPTER 3

Density Forecasts and Young Firm Dynamics[15]

3.1 Introduction

Panel data, such as a collection of firms or households observed repeatedly for a number of periods, are widely used in empirical studies and can be useful for forecasting individuals' future outcomes, which is interesting and important in many cases. For example, PSID can be used to analyze income dynamics (Hirano, 2002; Gu and Koenker, 2016b), and bank balance sheet data help conduct bank stress tests (Liu *et al.*, 2016). This paper constructs individual-specific density forecasts using a dynamic linear panel data model with common and heterogeneous parameters and cross-sectional heteroskedasticity.

In this paper, I consider young firm dynamics as the empirical application. For illustrative purposes, let us consider a simple dynamic panel data model as the baseline setup for this paper:

$$\underbrace{y_{it}}_{\text{performance}} = \beta y_{i,t-1} + \underbrace{\lambda_i}_{\text{skill}} + \underbrace{u_{it}}_{\text{shock}}, \quad u_{it} \sim N\left(0, \sigma^2\right), \tag{3.1.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. The y_{it} is the observed firm performance such as the log of employment,[16] λ_i is the unobserved skill of an individual firm, and u_{it} is an i.i.d. shock. Skill is independent of the shock, and the shock is independent across firms and times. β and σ^2 are common across firms, where β represents the persistence of the dynamic pattern, and σ^2 gives the size of the shocks. Because the number of observations

[15]This chapter builds on Liu (2016). I would like to acknowledge the Kauffman Foundation and the NORC Data Enclave for providing researcher support and access to the confidential microdata.

[16]Employment is a standard measure in the firm dynamics literature (Akcigit and Kerr, 2010; Zarutskie and Yang, 2015).

for each young firm is restricted by its age, the young firm panel is characterized by large cross-sectional dimension (N) but short time series (T).

Based on the observed panel, I am interested in forecasting the future performance of any specific firm, $y_{i,T+1}$, which is valuable to both investors and regulators. For investors, it is helpful to foresee which startups are more promising. For regulators, more accurate forecasts facilitate monitoring and regulation of bank-lending practices and entrepreneur funding.[17]

Considering that young firm dynamics involve sizeable uncertainties, a preferable forecast would provide a distribution that summarizes all kinds of uncertainties regarding firm i's future outcome. This is exactly the concept of density forecasts. Generally, forecasting can be done in point, interval, or density fashion, and density forecasts give the richest insight regarding future outcomes. A typical question that density forecasts could answer is: what is the chance that firm A will hire 5, 10, or 100 more people next year? Once the density forecasts are obtained, one can easily recover the point and interval forecasts.

In particular, for a panel data model as specified in equation (3.1.1), density forecasts capture uncertainties arising from both shocks u_{it}'s and heterogeneous skills λ_i's. The latter is due to the lack of time-series information available to infer individual λ_i. I assume that λ_i is drawn from the underlying skill distribution f, which serves as the key to characterize skill uncertainties and provide better density forecasts.

A benchmark for evaluating density forecasts is the posterior predictive distribution for $y_{i,T+1}$ under the assumption that the common parameters $\left(\beta, \sigma^2\right)$ and the distribution of the heterogeneous coefficients f are known. I refer to this predictive density as the (infeasible) oracle forecast. The role played by f can be more clearly appreciated in the following special case where the common parameters are set to be $\beta = 0$ and $\sigma^2 = 1$. It is straightforward to construct the oracle predictor for firm i, which combines firm i's shock uncertainty and

[17] The aggregate-level forecasts can be obtained by summing firm-specific forecasts over different subgroups.

skill uncertainty.

$$f_{i,T+1}^{oracle}(y) = \int \underbrace{\phi(y-\lambda_i)}_{\text{shock uncertainty}} \cdot \underbrace{p(\lambda_i | f_0, y_{i,1:T})}_{\text{skill uncertainty}} \cdot d\lambda_i.$$

Firm i's skill uncertainty can be interpreted as a posterior distribution with the prior belief being the common skill distribution f_0 and updated with firm i's data.

$$p(\lambda_i | f_0, y_{i,1:T}) = \frac{p(y_{i,1:T} | \lambda_i) f_0(\lambda_i)}{\int p(y_{i,1:T} | \lambda_i) f_0(\lambda_i) d\lambda_i}.$$

Therefore, the common skill distribution f_0 helps in formulating firm i's skill uncertainty and contributes to firm i's density forecasts through the channel of skill uncertainty.

In practice, however, the skill distribution f is unknown and unobservable, thus introducing another source of uncertainty. Now the oracle predictor becomes as an infeasible optimum. A good feasible predictor should be as close to the oracle as possible, which calls for a good estimate of the underlying skill distribution f. In this sense, the challenge is how we can model f more carefully and flexibly. The parametric Gaussian density misses many common features in the real world data, such as asymmetricity, heavy tails, or multiple peaks. Here I model f nonparametrically where the prior is constructed from a mixture model and allows for correlation between λ_i and y_{i0} (i.e. a correlated random effects model). Then, I pool the cross-sectional information to make inferences about f. The proposed semiparametric Bayesian procedure achieves better estimates of the underlying skill distribution f than parametric approaches, hence more accurate density forecasts of the future outcomes.

The contributions of this paper are threefold. First, I develop a posterior sampling algorithm specifically addressing nonparametric density estimation of the unobserved λ_i. For a random effects model, which is a special case with zero correlation between λ_i and y_{i0}, the f part becomes a relatively simple unconditional density estimation problem. I impose a Dirichlet Process Mixture (DPM) prior on f and construct a posterior sampler building on the blocked Gibbs sampler proposed by Ishwaran and James (2001, 2002). For a correlated random

effects model, I further adapt the proposed algorithm to the much harder conditional density estimation problem using a probit stick breaking process prior suggested by Pati *et al.* (2013).

Second, I establish the theoretical properties of the proposed semiparametric Bayesian predictor when the cross-sectional dimension N tends to infinity. Firstly, I provide conditions for identifying both the parametric component (β, σ^2) and the nonparametric component f. Then, I prove that both the estimated common parameters and the estimated distribution of the heterogeneous coefficients achieve posterior consistency, which is an essential building block for bounding the discrepancy between the proposed predictor and the oracle. Compared to previous literature on posterior consistency, there are several challenges in the current setting: (1) disentangling unobserved individual effects λ_i's and shocks u_{it}'s, (2) incorporating unknown shock size σ^2, (3) adding lagged dependent variables as covariates, and (4) addressing correlated random effects from a conditional density estimation point of view. Finally, I show that the density forecasts asymptotically converge to the oracle forecast in weak topology, which is new to the nonparametric Bayesian literature and specifically designed for density forecasts.

To accommodate many important features of real-world empirical studies, I extend the simple model (3.1.1) to a more general specification. First, a realistic application also incorporates other observables with common effects ($\beta' x_{i,t-1}$), where $x_{i,t-1}$ can include lagged y_{it}. Second, it is helpful to consider observables with heterogeneous effects ($\lambda_i' w_{i,t-1}$), i.e. a correlated random coefficients model. Finally, beyond heterogeneity in coefficients (λ_i), it is desirable to take into account heterogeneity in shock sizes (σ_i^2) as well.[18] All numerical methods and theoretical properties are further established for the general specification.

Third, Monte Carlo simulations demonstrate improvements in density forecasts relative to predictors with various parametric priors on f, evaluated by log predictive score. An application to young firm dynamics also shows that the proposed predictor provides more

[18] Here and below, the terminologies "random effects model" and "correlated random effects model" also apply to individual effects on σ_i^2, which are slightly different from the traditional definitions concentrated on λ_i.

accurate density predictions. The better forecasting performance is largely due to three key features (in order of importance): the nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients. The estimated model also helps shed light on the latent heterogeneity structure and how different factors (e.g. R&D, recession, etc.) contribute to the forecasts.

It is also worth mentioning that although I describe the econometric intuition using the young firm dynamics application as an example, the method is very general and can be applied to many economic and financial analyses that feature panel data with relatively large N and small T, such as microeconomic panel surveys (e.g. PSID, NLSY, and Consumer Expenditure Survey (CE)), macroeconomic sectoral and regional panel data (e.g. Industrial Production (IP), and State and Metro Area Employment, Hours, and Earnings (SAE)), and financial institution performance (e.g. Commercial Bank Data and Holding Company Data). Which T can be considered as a small T depends on the dimension of individual heterogeneity (d_w), the cross-sectional dimension (N), and size of the shocks (σ^2 or σ_i^2). There can still be a significant gain in density forecasts even when T exceeds 100. Roughly speaking, the proposed predictor would provide sizeable improvement as long as the time series for individual i is not informative enough to fully reveal its individual effects, λ_i and σ_i^2.

Related Literature First, this paper contributes to the literature on individual forecast in a panel data setup, and is closely related to Liu *et al.* (2016) and Gu and Koenker (2016a,b). Liu *et al.* (2016) focus on point forecasts. They utilize the idea of Tweedie's formula to steer away from the complicated deconvolution problem in estimating λ_i. Unfortunately, the Tweedie shortcut is not applicable to the inference of underlying λ_i distribution and therefore not suitable for density forecasts.

Gu and Koenker (2016b) address the density estimation problem. Their method is different from the one proposed in this paper in that this paper infers the underlying λ_i distribution via a full Bayesian approach (i.e. imposing a prior on the λ_i distribution and updating

the prior belief by the observed data), whereas they employ an empirical Bayes procedure (i.e. picking the λ_i distribution by maximizing the marginal likelihood of data). In principle, the full Bayesian approach is preferable for density forecasts as it captures all kinds of uncertainties, including estimation uncertainty of the underlying λ_i distribution, which has been omitted by the empirical Bayes procedure. In addition, this paper features correlated random effects allowing for both cross-sectional heterogeneities and cross-sectional heteroskedasticities interacting with the initial conditions, whereas the Gu and Koenker (2016b) approach focuses on random effects models without such interaction.

In their recent paper, Gu and Koenker (2016a) also compare their method with an alternative nonparametric Bayesian estimator featuring a Dirichlet Process (DP) prior under a set of fixed scale parameters. There are two major differences between their DP setup and the DPM prior used in this paper. First, the DPM prior provides continuous individual effect distributions, which is more reasonable in many empirical setups. Second, this paper incorporates a hyperprior for the scale parameter and updates it via the observed data, hence let the data choose the complexity of the mixture approximation, which can essentially be viewed as "automatic" model selection.[19]

There have also been empirical works on the DPM model with panel data, such as Hirano (2002), Burda and Harding (2013), Rossi (2014), and Jensen *et al.* (2015), but they focus on empirical studies rather than theoretical analysis. Hirano (2002) and Jensen *et al.* (2015) use linear panel models, while their setups are slightly different from this paper. Hirano (2002) considers flexibility in u_{it} distribution instead of λ_i distribution. Jensen *et al.* (2015) assume random effects instead of correlated random effects. Burda and Harding (2013) and Rossi (2014) implement nonlinear panel data models via either a probit model or a logit model, respectively.

Among others, Delaigle *et al.* (2008) have also studied the similar deconvolution problem

[19]Section 3.6 shows the simulation results comparing the DP prior vs the DPM prior. Both adopt a hyperprior for the scale parameter.

and estimated the λ_i distribution in a frequentist way, but the frequentist approach misses estimation uncertainty, which matters in density forecasts, as mentioned previously.

Second, in terms of asymptotic properties, this paper relates to the literature on posterior consistency of nonparametric Bayesian methods in density estimation problems. The pioneer work by Schwartz (1965) lays out two high-level sufficient conditions in a general density estimation context. Ghosal *et al.* (1999) bring Schwartz (1965)'s idea into the analysis of density estimation with DPM priors. Amewou-Atisso *et al.* (2003) extend the discussion to linear regression problems with an unknown error distribution. Tokdar (2006) further generalizes the results to cases in which the true density has heavy tails. For a more thorough review and discussion on posterior consistency in Bayesian nonparametric problems, please refer to the handbooks, Ghosh and Ramamoorthi (2003) and Hjort *et al.* (2010) (especially Chapters 1 and 2). To handle conditional density estimation, similar mixture structure can be implemented, where the mixing probabilities can be characterized by a multinomial choice model (Norets, 2010; Norets and Pelenis, 2012), a kernel stick break process (Norets and Pelenis, 2014; Pelenis, 2014), or a probit stick breaking process (Pati *et al.*, 2013). I adopt the Pati *et al.* (2013) approach to offer a more coherent nonparametric framework that is totally flexible in the conditional measure. This paper builds on the previous literature and establishes the posterior consistency result for panel data models. Furthermore, this paper obtains the convergence of the semiparametric Bayesian predictor to the oracle predictor, which is new to the literature and specific to density forecasts.

Third, the algorithms constructed in this paper build on the literature on the posterior sampling schemes for DPM models. The vast Markov chain Monte Carlo (MCMC) algorithms can be divided into two general categories. One is the Pólya urn style samplers that marginalize over the unknown distribution G (Escobar and West, 1995; Neal, 2000).[20] The other resorts to the stick breaking process (Sethuraman, 1994) and directly incorporates G into the sampling procedure. This paper utilizes a sampler from the second category, Ish-

[20]For the definition of G, see equation (3.2.5).

waran and James (2001, 2002)'s blocked Gibbs sampler, as a building block for the proposed algorithm. Basically, it incorporates truncation approximation and augments the data with auxiliary component probabilities, which helps break down the complex posterior structure and thus enhance mixing properties as well as reduce computation time.[21] I further adapt the proposed algorithm to the conditional density estimation for correlated random effects using the probit stick breaking process prior suggested by Pati *et al.* (2013).

Last but not least, the empirical application in this paper also links to the young firm dynamics literature. Akcigit and Kerr (2010) document the fact that R&D intensive firms grow faster, and such boosting effects are more prominent for smaller firms. Robb and Seamans (2014) examine the role of R&D in capital structure and performance of young firms. Zarutskie and Yang (2015) present some empirical evidence that young firms experienced sizable setbacks during the recent recession, which may partly account for the current slow and jobless recovery. For a thorough review on young firm innovation, please refer to the handbook by Hall and Rosenberg (2010). The empirical analysis of this paper builds on these previous findings. Besides providing more accurate density forecasts, we can also use the estimated model to analyze the latent heterogeneity structure and understand the effects of different factors (e.g. R&D, recession, etc.) on the forecasts.

The rest of the paper is organized as follows. Section 3.2 introduces the baseline panel data model as well as the oracle predictor and the feasible semiparametric Bayesian predictor. Section 3.3 proposes the posterior sampling algorithms. Section 3.4 characterizes identification conditions and large sample properties. Section 3.5 presents various extensions of the baseline model. Section 3.6 compares the performance of the semiparametric Bayesian predictor using simulated data, and Section 3.7 applies the proposed predictor to the confidential microdata from the Kauffman Firm Survey and analyzes the empirical findings on young firm dynamics. Finally, Section 3.8 concludes and sketches future research directions.

[21] Robustness checks have been conducted with the more sophisticated slice-retrospective sampler (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015), which does not involve hard truncation but is more complicated to implement. Results from the slice-retrospective sampler are comparable with the simpler truncation sampler.

Notations, proofs, as well as additional algorithms and results can be found in the Appendix.

3.2 Model

3.2.1 Baseline Panel Data Model

The baseline dynamic panel data model is specified in equation (3.1.1),

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T + h$. The y_{it} is the observed individual outcome, such as young firm performance. The main goal of this paper is to estimate the model using the sample from period 1 to period T and forecast the future distribution of $y_{i,T+h}$. In the remainder of the paper, I focus on the case where $h = 1$ (i.e. one-period-ahead forecasts) for notation simplicity, but the discussion can be extended to multi-period-ahead forecasts via either a direct or an iterated approach (Marcellino *et al.*, 2006).

In this baseline model, there are only three terms on the right hand side. $\beta y_{i,t-1}$ is the AR(1) term on lagged outcome, which captures the persistence pattern. λ_i is the unobserved individual heterogeneity modeled as individual-specific intercept, which implies that different firms may have different skill levels. u_{it} is the shock with zero mean and variance σ^2. To emphasize the basic idea, the baseline model assumes cross-sectional homoskedasticity, which means that the shock size σ^2 is the same across all firms.

As stressed in the motivation, the underlying skill distribution f is the key for better density forecasts. There can be two kinds of assumptions imposed on f. One is the random effects (RE) model, where the skill λ_i is independent of the initial performance y_{i0}. The other is the correlated random effects (CRE) model, where the skill λ_i and the initial performance y_{i0} can be potentially correlated with each other. This paper considers both RE and CRE models while focusing on the latter, as the CRE model is more realistic for young firm dynamics as well as many other empirical setups, and RE can be viewed as a special case

of CRE with zero correlation.

3.2.2 Oracle and Feasible Predictors

This subsection formally defines the infeasible optimal oracle predictor and the feasible semiparametric Bayesian predictor proposed in this paper. The kernel of both definitions relies on the conditional predictor,

$$f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)=\int\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right)d\lambda_i, \tag{3.2.1}$$

which provides the density forecasts of $y_{i,T+1}$ conditional on the common parameters (β,σ^2), underlying λ_i distribution (f), and firm i's data $(y_{i,0:T})$. The term $\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)$ captures firm i's shock uncertainty, and $p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right)$ characterizes firm i's skill uncertainty. Note that once conditioned on f, firms' performances are independent across i, and only firm i's data are needed for its density forecasts.

The infeasible oracle predictor is defined as if we knew all the elements that can be consistently estimated. Specifically, the oracle knows the common parameters (β_0,σ_0^2) and the underlying λ_i distribution (f_0), but not the skill of any individual firm λ_i. Then, the oracle predictor is formulated by plugging the true values $\left(\beta_0,\sigma_0^2,f_0\right)$ into the conditional predictor in equation (3.2.1),

$$f_{i,T+1}^{oracle}\left(y\right)=f_{i,T+1}^{cond}\left(y|\beta_0,\sigma_0^2,f_0,y_{i,0:T}\right). \tag{3.2.2}$$

In practice, $\left(\beta,\sigma^2,f\right)$ are all unknown but can be estimated via the Bayesian approach. First, I adopt the conjugate normal-inverse-gamma prior for the common parameters $\left(\beta,\sigma^2\right)$,

$$\left(\beta,\sigma^2\right)\sim N\left(m_0^\beta,\Sigma_0^\beta\right)\mathrm{IG}\left(\sigma^2;\ a_0^{\sigma^2},b_0^{\sigma^2}\right),$$

in order to stay close to the linear Gaussian regression framework. To flexibly model the

underlying skill distribution f, I resort to the nonparametric Bayesian prior, which is specified in detail in the next subsection. Then, I update the prior belief using the observations from the whole panel and obtain the posterior. The semiparametric Bayesian predictor is constructed by integrating the conditional predictor over the posterior distribution of (β, σ^2, f),

$$f_{i,T+1}^{sp}(y) = \int f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) d\Pi\left(\beta, \sigma^2, f\,|y_{1:N,0:T}\right) d\beta d\sigma^2 df. \tag{3.2.3}$$

3.2.3 Nonparametric Bayesian Priors

A prior on the skill distribution f can be viewed as a distribution over a set of distributions. Among other options, I choose mixture models for the nonparametric Bayesian prior, because according to the literature, mixture models can effectively approximate a general class of distributions (see Section 3.4) while being relatively easy to implement (see Section 3.3). Moreover, the choice of the nonparametric Bayesian prior also depends on whether f is characterized by a random effects model or a correlated random effects model. The correlated random effects setup is more involved but can be crucial in some empirical studies, such as the young firm dynamics application in this paper.

DPM Prior for Random Effects Model

In the random effects model, the skill λ_i is assumed to be independent of the initial performance y_{i0}, so the inference of the underlying skill distribution f can be considered as an unconditional density estimation problem. The DPM model is a typical nonparametric Bayesian prior designed for unconditional density estimation.

Dirichlet Process (DP) The key building block for the DPM model is the DP, which casts a distribution over a set of discrete distributions. A DP has two parameters: the base distribution G_0 characterizing the center of the DP, and the scale parameter α representing

the precision (inverse-variance) of the DP. Denote

$$G \sim DP\left(\alpha, G_0\right),$$

if for any partition $(A_1, \cdots, A_K)$,

$$\left(G\left(A_1\right), \cdots, G\left(A_K\right)\right) \sim \mathrm{Dir}\left(\alpha G_0\left(A_1\right), \cdots, \alpha G_0\left(A_K\right)\right).$$

$\mathrm{Dir}\left(\cdot\right)$ stands for the Dirichlet distribution with probability distribution function (pdf) being

$$f_{\mathrm{Dir}}\left(x_1, \cdots, x_K;\ \eta_1, \cdots, \eta_K\right) = \frac{\Gamma\left(\sum_{k=1}^{K} \eta_k\right)}{\prod_{k=1}^{K} \Gamma(\eta_k)} \prod_{k=1}^{K} x_k^{\eta_k - 1},$$

which is a multivariate generalization of the Beta distribution.

An alternative view of DP is given by the stick breaking process,

$$\begin{aligned} G &= \sum_{k=1}^{\infty} p_k \mathbf{1}\left(\theta = \theta_k\right), \\ \theta_k &\sim G_0, \quad k = 1, 2, \cdots, \\ p_k &= \begin{cases} \zeta_1, & k = 1, \\ \prod_{j=1}^{k-1}\left(1 - \zeta_j\right)\zeta_k, & k = 2, 3, \cdots, \end{cases} \\ &\text{where } \zeta_k \sim \mathrm{Beta}\left(1,\ \alpha\right), \quad k = 1, 2, \cdots. \end{aligned} \tag{3.2.4}$$

The stick breaking process distinguishes the roles of G_0 and α in that the former governs component value θ_k while the latter guides the choice of component probability p_k. From now on, for a concise exposition, I denote the p_k part in equation (3.2.4) as

$$p_k \sim \mathrm{SB}\left(1, \alpha\right),\ k = 1, 2, \cdots,$$

where the function name "SB" is the acronym for "stick breaking", and the two arguments

are passed from the parameters of the Beta distribution for "stick length" ζ_k.

Dirichlet Process Mixture (DPM) Prior By definition, a draw from DP is a discrete distribution. In this sense, imposing a DP prior on the skill distribution f amounts to restricting firms' skills to some discrete levels, which may not be very appealing for young firm dynamics as well as some other empirical applications. A natural remedy is to assume λ follows a continuous parametric distribution $f(\lambda;\theta)$ where θ are the parameters, and adopt a DP prior for the distribution of θ. Then, the parameters θ are discrete while the skill λ enjoys a continuous distribution. This additional layer of mixture lead to the idea of the DPM model. For variables supported on the whole real line, like the skill λ here, a typical choice of the kernel of $f(\lambda;\theta)$ is a normal distribution with $\theta = \left(\mu, \omega^2\right)$ being the mean and variance of the normal.

$$\begin{aligned} \lambda_i &\sim N\left(\lambda_i;\ \mu_i, \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) &\overset{iid}{\sim} G, \\ G &\sim DP\left(\alpha, G_0\right). \end{aligned} \tag{3.2.5}$$

Equivalently, with component label k, component probability p_k, and component parameters $\left(\mu_k, \omega_k^2\right)$, one draw from the DPM prior can be rewritten as an infinite mixture of normals,

$$\lambda_i \sim \sum_{k=1}^{\infty} p_k N\left(\lambda_i;\ \mu_k, \omega_k^2\right). \tag{3.2.6}$$

Different draws from the DPM prior are characterized by different combinations of $\{p_k, \mu_k, \omega_k^2\}$, and different combinations of $\left\{p_k, \mu_k, \omega_k^2\right\}$ lead to different shapes of f. That is why the DPM prior is flexible enough to approximate many distributions. The component parameters $\left(\mu_k, \omega_k^2\right)$ are directly drawn from the DP base distribution G_0, which is chosen to be the conjugate normal-inverse-gamma distribution. The component probability p_k is constructed

via the stick breaking process governed by the DP scale parameter α.

$$\begin{aligned}\left(\mu_k, \omega_k^2\right) &\sim G_0, \\ p_k &\sim \mathrm{SB}\left(1, \alpha\right),\ k = 1, 2, \cdots .\end{aligned}$$

Comparing the above two sets of expressions in equations (3.2.5) and (3.2.6), the first set links the flexible structure in λ to the flexible structure in $\left(\mu, \omega^2\right)$, and serves as a more convenient setup for the theoretical derivation of asymptotic properties as in Subsection 3.4.3; at the same time, the second set separates the channels regarding component parameters and component probabilities, and therefore is more suitable for the numerical implementation as in Section 3.3.

One virtue of the nonparametric Bayesian framework is to flexibly elicit the tuning parameter from the data. Namely, we can set up an additional hyperprior for the DP scale parameter α,

$$\alpha \sim \mathrm{Ga}\left(\alpha;\ a_0^\alpha, b_0^\alpha\right),$$

and update it based on the observations. Roughly speaking, the DP scale parameter α is linked to the number of unique components in the mixture density and thus determines and reflects the flexibility of the mixture density. Let K^* denote the number of unique components. As derived in Antoniak (1974), we have

$$\begin{aligned}E\left[K^*|\alpha\right] &\approx \alpha \log\left(\frac{\alpha + N}{\alpha}\right), \\ Var\left[K^*|\alpha\right] &\approx \alpha\left[\log\left(\frac{\alpha + N}{\alpha}\right) - 1\right].\end{aligned}$$

$\mathbf{MGLR_x}$ Prior for Correlated Random Effects Model

To accommodate the correlated random effects model where the skill λ_i can be potentially correlated with the initial performance y_{i0}, it is necessary to consider a nonparametric

Bayesian prior that is compatible with the much harder conditional density estimation problem. One issue is associated with the uncountable collection of conditional densities, and Pati *et al.* (2013) circumvent it by linking the properties of the conditional densities to the corresponding ones of the joint densities. As suggested in Pati *et al.* (2013), I utilize the Mixtures of Gaussian Linear Regressions (MGLR_x) prior, a generalization of the Gaussian-mixture prior for conditional density estimation. Conditioning on y_{i0},

$$\begin{aligned} \lambda_i | y_{i0} &\sim N\left(\lambda_i;\ \mu_i \left[1, y_{i0}\right]',\ \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) \equiv \theta_i &\overset{iid}{\sim} G\left(\cdot;\ y_{i0}\right), \\ G\left(\cdot;\ y_{i0}\right) &= \sum_{k=1}^{\infty} p_k\left(y_{i0}\right) \delta_{\theta_k}. \end{aligned} \tag{3.2.7}$$

In the baseline setup, both individual heterogeneity λ_i and conditioning set y_{i0} are scalars, so μ_i is a two-element row vector and ω_i^2 is a scalar. Similar to the DPM prior, the component parameters can be directly drawn from the base distribution, which is again specified as the conjugate normal-inverse-gamma distribution,

$$\theta_k \sim G_0, \quad k = 1, 2, \cdots. \tag{3.2.8}$$

Now the mixture probabilities are characterized by the probit stick breaking process

$$p_k\left(y_{i0}\right) = \Phi\left(\zeta_k\left(y_{i0}\right)\right) \prod_{j<k} \left(1 - \Phi\left(\zeta_j\left(y_{i0}\right)\right)\right), \tag{3.2.9}$$

where stochastic function ζ_k is drawn from the Gaussian process $\zeta_k \sim GP\left(0, V_k\right)$ for $k = 1, 2, \cdots$.[22]

Expression (3.2.7) can be perceived as a conditional counterpart of expression (3.2.5) for the purpose of theoretical derivation. The following expression (3.2.10) corresponds to expres-

[22] For a generic variable c which can be multi-dimensional, the Gaussian process $\zeta(c) \sim GP\left(m(c), V(c, \tilde{c})\right)$ is defined as follows: for any finite set of $\{c_1, c_2, \cdots, c_n\}$, $\left[\zeta(c_1), \zeta(c_2), \cdots, \zeta(c_n)\right]'$ has a joint Gaussian distribution with the mean vector being $\left[m(c_1), m(c_2), \cdots, m(c_n)\right]'$ and the i,j-th entry of covariance matrix being $V(c_i, c_j)$, $i, j = 1, \cdots, N$.

sion (3.2.6), which is in line with the numerical implementation in Section 3.3:

$$\lambda_i|y_{i0} \sim \sum_{k=1}^{\infty} p_k\left(y_{i0}\right) N\left(\mu_k\left[1, y_{i0}\right]',\ \omega_k^2\right), \tag{3.2.10}$$

where the component parameters and component probabilities are specified in equations (3.2.8) and (3.2.9), respectively.

This setup has three key features: (1) component means are linear in y_{i0}; (2) component variances are independent of y_{i0}; and (3) mixture probabilities are flexible functions of y_{i0}. This framework is general enough to accommodate many conditional distributions. Intuitively, by Bayes' theorem,

$$f\left(\lambda|y_0\right) = \frac{f\left(\lambda, y_0\right)}{f\left(y_0\right)}.$$

The joint distribution in the numerator can be approximated by a mixture of normals

$$f\left(\lambda, y_0\right) \approx \sum_{k=1}^{\infty} \tilde{p}_k \phi\left(\left[\lambda, y_0\right]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right),$$

where $\tilde{\mu}_k$ is a two-element column vector, and $\tilde{\Omega}_k$ is a 2×2 covariance matrix. Applying Bayes' theorem again to the normal kernel for each component k,

$$\phi\left(\left[\lambda, y_0\right]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right) = \phi\left(y_0;\ \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22}\right) \phi\left(\lambda;\ \mu_k\left[1, y_0\right]', \omega_k^2\right),$$

where $\mu_k = \left[\tilde{\mu}_{k,1} - \frac{\tilde{\Omega}_{k,12}}{\tilde{\Omega}_{k,22}}\tilde{\mu}_{k,2},\ \frac{\tilde{\Omega}_{k,12}}{\tilde{\Omega}_{k,22}}\right]$, $\omega_k^2 = \tilde{\Omega}_{k,11} - \frac{\left(\tilde{\Omega}_{k,12}\right)^2}{\tilde{\Omega}_{k,22}}$. Combining all the steps above, the conditional distribution can be approximated as

$$\begin{aligned} f\left(\lambda|y_0\right) &\approx \sum_{k=1}^{\infty} \frac{\tilde{p}_k \phi\left(y_0;\ \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22}\right) \phi\left(\lambda;\ \mu_k\left[1, y_0\right]', \omega_k^2\right)}{f\left(y_0\right)} \\ &= \sum_{k=1}^{\infty} p_k\left(y_0\right) \phi\left(\lambda;\ \mu_k\left[1, y_0\right]', \omega_k^2\right), \end{aligned}$$

The last line is given by collecting marginals of y_{i0} into $p_k(y_0) = \frac{\tilde{p}_k \phi(y_0; \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22})}{f(y_0)}$. In summary, the current setup is similar to approximating the conditional density via Bayes' theorem, but does not explicitly model the distribution of the conditioning variable y_{i0}, and thus allows for more relaxed assumptions on it.

3.3 Numerical Implementation

In this section, I propose a posterior sampling procedure for the baseline panel data model introduced in Subsection 3.2.1 together with the nonparametric Bayesian prior specified in Subsection 3.2.3 that enjoys desirable theoretical properties as discussed in Section 3.4.

Recall the baseline model,

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

and the conjugate normal-inverse-gamma prior for the common parameters $\left(\beta, \sigma^2\right)$,

$$\left(\beta, \sigma^2\right) \sim N\left(m_0^\beta, \psi_0^\beta \sigma^2\right) \mathrm{IG}\left(\sigma^2;\ a_0^{\sigma^2}, b_0^{\sigma^2}\right).$$

The hyperparameters are chosen in a relatively ignorant sense without inferring too much from the data except aligning the scale according to the variance of the data (see Appendix B.2.1 for details). The skill λ_i is drawn from the underlying skill distribution f, which can be characterized by either the random effects model or the correlated random effects model. Subsection 3.3.1 describes the posterior sampler for the former, and Subsection 3.3.2 delineates the posterior sampler for the latter.

3.3.1 Random Effects Model

For the random effects model, I impose the Gaussian-mixture DPM prior on f. The posterior sampling algorithm builds on the blocked Gibbs sampler proposed by Ishwaran and James (2001, 2002). They truncate the number of components by a large K, and prove that as long

as K is large enough, the truncated prior is "virtually indistinguishable" from the original one. Once truncation is conducted, it is possible to augment the data with latent component probabilities, which boosts numerical convergence and leads to faster code.

To check the robustness regarding the truncation, I also implement the more sophisticated yet complicated slice-retrospective sampler (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015) which does not truncate the number of components at a predetermined K. The full algorithm for the general model (3.5.1) can be found as Algorithm B.2.4 in the Appendix. The estimates and forecasts for the two samplers are comparable, so I will only show the results generated from the simpler truncation sampler in this paper.

Suppose the number of components is truncated at K. Then, the Gaussian-mixture DPM prior can be expressed as[23]

$$\lambda_i \sim \sum_{k=1}^{K} p_k N\left(\mu_k, \omega_k^2\right), \quad i = 1, \cdots, N.$$

The parameters for each component can be viewed as directly drawn from the DP base distribution G_0. A typical choice of G_0 is the normal-inverse-gamma prior, which respects the conjugacy when the DPM kernel is also normal (see Appendix B.2.1 for details of hyperparameter choices).

$$G_0\left(\mu_k, \omega_k^2\right) = N\left(\mu_k;\ m_0^\lambda, \psi_0^\lambda \omega_k^2\right) \mathrm{IG}\left(\omega_k^2;\ a_0^\lambda, b_0^\lambda\right).$$

The component probabilities are constructed via a truncated stick breaking process governed

[23] In this section, the nonparametric Bayesian priors are formulated as in equations (3.2.6) and (3.2.10). Such expressions explicitly separate the channels regarding component parameters and component probabilities, and hence facilitate the construction of the posterior samplers.

by the DP scale parameter α.

$$p_k = \begin{cases} \zeta_1, & k = 1, \\ \prod_{j=1}^{k-1} (1-\zeta_j)\,\zeta_k, & k = 2, \cdots, K-1, \\ 1 - \sum_{j=1}^{K-1} p_j, & k = K, \end{cases}$$
$$\text{where } \zeta_k \sim \text{Beta}\,(1,\ \alpha)\,, \quad k = 1, \cdots, K-1.$$

Note that due to the truncation approximation, the probability for component K is different from its infinite mixture counterpart in equation (3.2.4). Resembling the infinite mixture case, I denote the above truncated sticking process as

$$p_k \sim \text{TSB}\,(1, \alpha, K)\,,\ k = 1, \cdots K,$$

where "TSB" is for "truncated stick breaking", the first two arguments are passed from the parameters of the Beta distribution, and the last argument is the truncated number of components.

Let γ_i be firm i's component affiliation, which can take values $\{1, \cdots, K\}$, J_k be the set of firms in component k, i.e. $J_k = \{i :\ \gamma_i = k\}$, and n_k be the number of individuals in component k, i.e. $n_k = \#J_k$. Then, the (data-augmented) joint posterior for the model parameters is given by

$$\begin{aligned} & p\left(\alpha, \left\{p_k, \mu_k, \omega_k^2\right\}, \left\{\gamma_i, \lambda_i\right\}, \beta, \sigma^2 \middle|\, y_{1:N,0:T}\right) \\ & = \prod_{i,t} p\left(y_{it} \left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) \cdot \prod_i p\left(\lambda_i \left|\mu_{\gamma_i}, \omega_{\gamma_i}^2\right.\right) p\left(\gamma_i \left|\{p_k\}\right.\right) \\ & \quad \cdot \prod_k p\left(\mu_k, \omega_k^2\right) p\left(p_k|\alpha\right) \cdot p\left(\alpha\right) \cdot p\left(\beta, \sigma^2\right), \end{aligned} \tag{3.3.1}$$

where $k = 1, \cdots, K$, $i = 1, \cdots N$, and $t = 1, \cdots, T$.

The first block $\prod_{i,t} p\left(y_{it} \left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right)$ links observations to model parameters $\{\lambda_i\}, \beta$,

and σ^2. The second block $\prod_i p\left(\lambda_i \left| \mu_{\gamma_i}, \omega_{\gamma_i}^2 \right.\right) p\left(\gamma_i \left| \{p_k\} \right.\right)$ links the skill λ_i to the underlying skill distribution f. The last block $\prod_k p\left(\mu_k, \omega_k^2\right) p\left(p_k|\alpha\right) \cdot p\left(\alpha\right) \cdot p\left(\beta, \sigma^2\right)$ formulates the prior belief on $\left(\beta, \sigma^2, f\right)$.

The following Gibbs sampler cycles over the following blocks of parameters (in order): (1) component probabilities, $\alpha, \{p_k\}$; (2) component parameters, $\left\{\mu_k, \omega_k^2\right\}$; (3) component memberships, $\{\gamma_i\}$; (4) individual effects, $\{\lambda_i\}$; (5) common parameters, β, σ^2. A sequence of draws from this algorithm forms a Markov chain with the sampling distribution converging to the posterior density.

Note that if the skill λ_i were known, only step (5) would be sufficient to recover the common parameters. If the mixture structure of f were known (i.e. $\left(p_k, \mu_k, \omega_k^2\right)$ for all components were known), steps (3)-(5) would be needed to first assign firms to components and then infer firm i's skill based on the specific component that it has been assigned to. In reality, neither skill λ_i nor its distribution f is known, so I incorporate two more steps (1)-(2) to model the underlying skill distribution f.

Below, I present the formulas for the key nonparametric Bayesian steps, and leave the details of standard posterior sampling procedures, such as drawing from a normal-inverse-gamma distribution or a linear regression, to Appendix B.2.3.

Algorithm 3.3.1. *(Baseline Model: Random Effects)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities:*
   (a) *Draw* $\alpha^{(s)}$ *from a gamma distribution* $p\left(\alpha^{(s)} \middle| p_K^{(s-1)}\right)$*:*

$$\alpha^{(s)} \sim Ga\left(\alpha^{(s)};\ a_0^\alpha + K - 1,\ b_0^\alpha - \log p_K^{(s-1)}\right).$$

   (b) *For* $k = 1, \cdots, K$*, draw* $p_k^{(s)}$ *from the truncated stick breaking process*

$p\left(\left\{p_k^{(s)}\right\}\middle|\alpha^{(s)},\left\{n_k^{(s-1)}\right\}\right)$:

$$p_k^{(s)} \sim TSB\left(1+n_k^{(s-1)},\ \alpha^{(s)}+\sum_{j=k+1}^{K} n_j^{(s-1)},\ K\right),\ k=1,\cdots K.$$

2. *Component parameters: For* $k=1,\cdots,K$, *draw* $\left(\mu_k^{(s)},\omega_k^{2(s)}\right)$ *from a normal-inverse-gamma distribution* $p\left(\mu_k^{(s)},\omega_k^{2(s)}\middle|\left\{\lambda_i^{(s-1)}\right\}_{i\in J_k^{(s-1)}}\right)$.
3. *Component memberships: For* $i=1,\cdots N$, *draw* $\gamma_i^{(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{(s)}\right\}\middle|\left\{p_k^{(s)},\mu_k^{(s)},\omega_k^{2(s)}\right\},\ \lambda_i^{(s-1)}\right)$:

$$\gamma_i^{(s)}=k,\ \textit{with probability}\ p_{ik},\ k=1,\cdots,K,$$
$$p_{ik}\propto p_k^{(s)}\phi\left(\lambda_i^{(s-1)};\ \mu_k^{(s)},\omega_k^{2(s)}\right),\quad \sum_{k=1}^{K}p_{ik}=1.$$

4. *Individual effects: For* $i=1,\cdots,N$, *draw* $\lambda_i^{(s)}$ *from a normal distribution* $p\left(\lambda_i^{(s)}\middle|\mu_{\gamma_i^{(s)}}^{(s)},\omega_{\gamma_i^{(s)}}^{2(s)},\beta^{(s-1)},\sigma^{2(s-1)},y_{i,0:T}\right)$.
5. *Common parameters: Draw* $\left(\beta^{(s)},\sigma^{2(s)}\right)$ *from a linear regression model* $p\left(\beta^{(s)},\sigma^{2(s)}\middle|\left\{\lambda_i^{(s)}\right\},\ y_{1:N,0:T}\right)$.

3.3.2 Correlated Random Effects Model

To account for the conditional structure in the correlated random effects model, I implement the MGLR_x prior as specified in Subsection 3.2.3, which can be viewed as the conditional counterpart of the Gaussian-mixture prior. In the baseline setup, the conditioning set is a singleton with y_{i0} being the only element.

The major computational difference from the random effects model in the previous subsection is that now the component probabilities become flexible functions of y_{i0}. As suggested in Pati *et al.* (2013), I adopt the following priors and auxiliary variables in order to take advantage of conjugacy as much as possible. First, the covariance function for Gaussian

process $V_k(c, \tilde{c})$ is specified as

$$V_k(c, \tilde{c}) = \exp\left(-A_k |c - \tilde{c}|^2\right),$$

where $k = 1, 2, \cdots$. An exponential prior is imposed on A_k, i.e.

$$p(A_k) \propto \exp(-A_k),$$

so $p(A_k)$ has full support on $\mathbb{R}^+$ and satisfies Pati *et al.* (2013) Remark 5.2.

Furthermore, it is helpful to introduce a set of auxiliary stochastic functions $\xi_k(y_{i0})$, $k = 1, 2, \cdots$, such that

$$\begin{aligned}
&\xi_k(y_{i0}) \sim N(\zeta_k(y_{i0}), 1), \\
&p_k(y_{i0}) = \text{Prob}(\xi_k(y_{i0}) \geq 0, \text{ and } \xi_j(y_{i0}) < 0 \text{ for all } j < k).
\end{aligned}$$

Note that the probit stick breaking process defined in equation (3.2.9) can be recovered by marginalizing over $\xi_k(y_{i0})$'s.

Finally, I blend the MGLR_x prior with Ishwaran and James (2001, 2002) truncation approximation to simplify the numerical procedure while still retaining reliable results.

Denote $N \times 1$ vectors

$$\begin{aligned}
\boldsymbol{\zeta}_k &= [\zeta_k(y_{10}), \zeta_k(y_{20}), \cdots, \zeta_k(y_{N0})]', \\
\boldsymbol{\xi}_k &= [\xi_k(y_{10}), \xi_k(y_{20}), \cdots, \xi_k(y_{N0})]',
\end{aligned}$$

as well as an $N \times N$ matrix $\boldsymbol{V}_k$ with the ij-th element being

$$(\boldsymbol{V}_k)_{ij} = \exp\left(-A_k |y_{i0} - y_{j0}|^2\right).$$

The next algorithm extends Algorithm 3.3.1 to the correlated random effects scenario. Step 1 for component probabilities has been changed, while the rest of the steps are in line with those in Algorithm 3.3.1.

Algorithm 3.3.2. *(Baseline Model: Correlated Random Effects)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities:*

   (a) *For* $k = 1, \cdots, K-1$, *draw* $A_k^{(s)}$ *via the random-walk Metropolis-Hastings approach,*

   $$p\left(A_k^{(s)} \middle| \boldsymbol{\zeta}_k^{(s-1)}, \{y_{i0}\}\right) \propto \exp\left(-A_k^{(s)}\right) \phi\left(\boldsymbol{\zeta}_k^{(s-1)};\ 0,\ \exp\left(-A_k^{(s)} \left|y_{i0} - y_{j0}\right|^2\right)\right).$$

   Then, calculate $\boldsymbol{V}_k^{(s)}$ *such that*

   $$\left(\boldsymbol{V}_k^{(s)}\right)_{ij} = \exp\left(-A_k^{(s)} \left|y_{i0} - y_{j0}\right|^2\right).$$

   (b) *For* $k = 1, \cdots, K-1$, *and* $i = 1, \cdots, N$, *draw* $\xi_k^{(s)}(y_{i0})$ *from a truncated normal distribution* $p\left(\xi_k^{(s)}(y_{i0}) \middle| \zeta_k^{(s-1)}(y_{i0}), \gamma_i^{(s-1)}\right)$:

   $$\xi_k^{(s)}(y_{i0}) \begin{cases} \propto N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right) \mathbf{1}\left(\xi_k^{(s)}(y_{i0}) < 0\right), & \text{if } k < \gamma_i^{(s-1)}, \\ \propto N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right) \mathbf{1}\left(\xi_k^{(s)}(y_{i0}) \geq 0\right), & \text{if } k = \gamma_i^{(s-1)}, \\ \sim N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right), & \text{if } k > \gamma_i^{(s-1)}, \end{cases}$$

   .

   (c) *For* $k = 1, \cdots, K-1$, *draw* $\boldsymbol{\zeta}_k^{(s)}$ *from a multivariate normal distribution*

$p\left(\boldsymbol{\zeta}_k^{(s)} \middle| \boldsymbol{V}_k^{(s)}, \boldsymbol{\xi}_k^{(s)}\right)$:

$$\begin{aligned}
\boldsymbol{\zeta}_k^{(s)} &\sim N\left(m_k^{\zeta},\ \Sigma_k^{\zeta}\right), \\
\Sigma_k^{\zeta} &= \left[\left(\boldsymbol{V}_k^{(s)}\right)^{-1} + I_N\right]^{-1}, \\
m_k^{\zeta} &= \Sigma_k^{\zeta} \boldsymbol{\xi}_k^{(s)}.
\end{aligned}$$

(d) For $k = 1, \cdots, K$, and $i = 1, \cdots, N$, the component probabilities $p_k^{(s)}(y_{i0})$ are fully determined by $\boldsymbol{\zeta}_k^{(s)}$:

$$p_k^{(s)}(y_{i0}) = \begin{cases} \Phi\left(\zeta_1^{(s)}(y_{i0})\right), & \text{if } k = 1, \\ \Phi\left(\zeta_k^{(s)}(y_{i0})\right) \prod_{j<k}\left(1 - \Phi\left(\zeta_j^{(s)}(y_{i0})\right)\right), & \text{if } k = 2, \cdots, K-1, \\ 1 - \sum_{j=1}^{K-1} p_k^{(s)}(y_{i0}), & \text{if } k = K. \end{cases}$$

2. *Component parameters: For $k = 1, \cdots, K$, draw $\left(\mu_k^{(s)}, \omega_k^{2(s)}\right)$ from a linear regression model $p\left(\mu_k^{(s)}, \omega_k^{2(s)} \middle| \left\{\lambda_i^{(s-1)}, y_{i0}\right\}_{i \in J_k^{(s-1)}}\right)$.*
3. *Component memberships: For $i = 1, \cdots N$, draw $\gamma_i^{(s)}$ from a multinomial distribution $p\left(\left\{\gamma_i^{(s)}\right\} \middle| \left\{p_k^{(s)}, \mu_k^{(s)}, \omega_k^{2(s)}\right\},\ \lambda_i^{(s-1)}, y_{i0}\right)$:*

$$\begin{aligned}
&\gamma_i^{(s)} = k, \text{ with probability } p_{ik},\ k = 1, \cdots, K, \\
&p_{ik} \propto p_k^{(s)}(y_{i0})\, \phi\left(\lambda_i^{(s-1)};\ \mu_k^{(s)}[1, y_{i0}]', \omega_k^{2(s)}\right), \quad \sum_{k=1}^{K} p_{ik} = 1.
\end{aligned}$$

4. *Individual effects: For $i = 1, \cdots, N$, draw $\lambda_i^{(s)}$ from a normal distribution $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^{(s)}}^{(s)}, \omega_{\gamma_i^{(s)}}^{2(s)}, \beta^{(s-1)}, \sigma^{2(s-1)}, y_{i,0:T}\right)$.*
5. *Common parameters: Draw $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ from a linear regression model $p\left(\beta^{(s)}, \sigma^{2(s)} \middle| \left\{\lambda_i^{(s)}\right\},\ y_{1:N,0:T}\right)$.*

Remark 3.3.3. With the above prior specification, all steps enjoy closed-form conditional posterior distributions except step 1-a for A_k, which does not exhibit a well-known density

form. Hence, I resort to the random-walk Metropolis-Hastings (RWMH) algorithm to sample A_k. In addition, I also incorporate an adaptive procedure based on Atchadé and Rosenthal (2005) and Griffin (2016), which adaptively adjusts the random walk step size and keep acceptance rates around 30%. Intuitively, when the acceptance rate for the current iteration is too high (low), the adaptive algorithm increases (decreases) the step size in the next iteration, and thus potentially raises (lowers) the acceptance rate in the next round. The change in step size decreases with the number of iterations completed, and the step size converges to the optimal value. Please refer to the detailed description in Algorithm B.2.1 in the Appendix.

3.4 Theoretical Properties

3.4.1 Background

Generally speaking, Bayesian analysis starts with a prior belief and updates it with data. It is desirable to ensure that the prior belief does not dominate the posterior inference asymptotically. Namely, as more and more data have been observed, one would have weighed more on the data and less on prior, and the effect from the prior would have ultimately been washed out. For pure Bayesians who have different prior beliefs, the asymptotic properties make sure that they will eventually agree on similar predictive distributions (Blackwell and Dubins, 1962; Diaconis and Freedman, 1986). For frequentists who perceive that there is an unknown true data generating process, the asymptotic properties act as frequentist justification for the Bayesian analysis—as the sample size increases, the updated posterior recovers the unknown truth. Moreover, the conditions for posterior consistency provide guidance in choosing better-behaved priors.

In the context of infinite dimensional analysis such as density estimation, posterior consistency cannot be taken as given. On the one hand, Doob's theorem (Doob, 1949) indicates that Bayesian posterior will achieve consistency almost surely under the prior measure. On the other hand, the null set for the prior can be topologically large, and hence the true

model can easily fall beyond the scope of the prior, especially in nonparametric analysis. Freedman (1963) gives a simple counter-example in the nonparametric setup, and Freedman (1965) further examines the combinations of the prior and the true parameters that yield a consistent posterior, and proves that such combinations are meager in the joint space of the prior and the true parameters. Therefore, for problems involving density estimation, it is crucial to find reasonable conditions on the joint behavior of the prior and the true density to establish the posterior consistency argument.

In this section, I show the asymptotic properties of the proposed semiparametric Bayesian predictor when the cross-sectional dimension N tends to infinity. Basically, under reasonably general conditions, the joint posterior of the common parameters and the individual effect distribution concentrates in an arbitrarily small region around the true underlying model, and the density forecasts concentrate in an arbitrarily small region around the oracle. Subsection 3.4.2 provides the conditions for identification, which lays the foundation for posterior consistent analysis. Subsection 3.4.3 proves the posterior consistency of the estimator, which is an essential building block for bounding the discrepancy between the proposed predictor and the oracle. Finally, Subsection 3.4.4 establishes the main Bayesian asymptotic argument for density forecasts.

3.4.2 Identification

To establish the posterior consistency argument, we first need to ensure identification for both the common parameters and the (conditional) distribution of individual effects. Here, I present the identification result in terms of the correlated random effects model, with the random effects model being a special case. In the baseline setup, the identification argument directly follows Assumptions 2.1-2.2 and Theorem 2.3 in Liu *et al.* (2016), which is in turn based on early works, such as Arellano and Bover (1995) and Arellano and Bonhomme (2012b), so below I only state the assumption and the proposition without extensive discussion. Please refer to Subsection 3.5.3 for more general results addressing correlated random coefficients, cross-sectional heteroskedasticities, and unbalanced panels.

Assumption 3.4.1. *(Baseline Model: Identification)*

1. $\{y_{i0},\ \lambda_i\}$ *are i.i.d. across* i.
2. u_{it} *is i.i.d. across* i *and* t, *and independent of* λ_i.
3. *The characteristic function for* $\lambda_i|y_{i0}$ *is non-vanishing almost everywhere.*
4. $T \geq 2$.

The first condition characterizes the correlated random effects model, where there can be potential correlation between skill λ_i and initial performance y_{i0}. For the random effects case, this condition can be altered to "λ_i is independent of y_{i0} and i.i.d. across i". The second condition ensures that skill is independent of shock, and that shock is independent across firms and times, so skill and shock are intrinsically different and distinguishable. The third condition facilitates the deconvolution between the signal (skill) and the noise (shock) via Fourier transformation. The last condition guarantees that the time span is long enough to distinguish persistence ($\beta y_{i,t-1}$) and individual effects (λ_i). Then, the identification statement is established as follows.

Proposition 3.4.2. *(Baseline Model: Identification)*

Under Assumption 3.4.1, the common parameters $\left(\beta, \sigma^2\right)$ *and the conditional distribution of individual effects* $f(\lambda_i|y_{i0})$ *are all identified.*

3.4.3 Posterior Consistency

In this subsection, I establish the posterior consistency of the estimated common parameters $\left(\beta, \sigma^2\right)$ and the estimated (conditional) distribution of individual effects f in the baseline setup. Subsections 3.4.3 and 3.4.3 examine the random effects model and the correlated random effects model, respectively. Further discussion of the general model can be found in Subsection 3.5.4.

Random Effects Model

First, let us consider the random effects model with f being an unconditional distribution. Let $\Theta = \mathbb{R} \times \mathbb{R}^+$ be the space for the parametric component $\vartheta = \left(\beta, \sigma^2\right)$, and let $\mathcal{F}$ be the set of densities on $\mathbb{R}$ (with respect to Lebesgue measure) as the space for the nonparametric component f. The true data generating process is characterized by (ϑ_0, f_0). The posterior consistency results are established with respect to the weak topology, which is generated by a neighborhood basis constituted of the weak neighborhoods defined below.

Definition 3.4.3. A *weak neighborhood* of f_0 is defined as

$$U_{\epsilon,\Phi}(f_0) = \left\{ f \in \mathcal{F} : \left| \int \varphi_j f - \int \varphi_j f_0 \right| < \epsilon \right\}$$

where $\epsilon > 0$ and $\Phi = \{\varphi_j\}_{j=1}^J$ are bounded, continuous functions.

Let $\Pi(\cdot,\cdot)$ be a joint prior distribution on $\Theta \times \mathcal{F}$ with marginal priors being $\Pi^{\vartheta}(\cdot)$ and $\Pi^f(\cdot)$. The corresponding joint posterior distribution is denoted as $\Pi(\cdot,\cdot|y_{1:N,0:T})$ with the marginal posteriors defined similarly as above.

Definition 3.4.4. The posterior achieves *weak consistency* at (ϑ_0, f_0) if for any $U_{\epsilon,\Phi}(f_0)$ and any $\delta > 0$, as $N \to \infty$,

$$\Pi\left((\vartheta, f) : \ \|\vartheta - \vartheta_0\| < \delta, \ f \in U_{\epsilon,\Phi}(f_0) | \, y_{1:N,0:T}\right) \to 1, \ a.s.$$

As stated in the original Schwartz (1965) theorem (Lemma 3.4.6), weak consistency is closely related to the Kullback-Leibler (KL) divergence. For any two distributions f_0 and f, the *KL divergence* of f from f_0 is defined as

$$d_{KL}(f_0, f) = \int f_0 \log \frac{f_0}{f}.$$

The *KL property* is characterized based on KL divergence as follows.

Definition 3.4.5. If for all $\epsilon > 0$, $\Pi^f\left(f \in \mathcal{F}:\ d_{KL}\left(f_0, f\right) < \epsilon\right) > 0$, we say f_0 is *in the KL support of* Π^f, or $f_0 \in KL\left(\Pi^f\right)$.

Preliminary: Schwartz (1965) Theorem The following lemma restates the Schwartz (1965) theorem of weak posterior consistency. It is established in a simpler scenario where we observe λ_i (not y_i) and wants to infer its distribution.

Lemma 3.4.6. *(Schwartz, 1965)*

The posterior is weakly consistent at f_0 under two sufficient conditions:

1. *Kullback-Leibler property: f_0 is in the KL support of Π, or $f_0 \in KL\left(\Pi\right)$.*
2. *Uniformly exponentially consistent tests: For any $U = U_{\epsilon,\Phi}\left(f_0\right)$, there exists $\gamma > 0$ and a sequence of tests $\boldsymbol{\varphi_N}\left(\lambda_1, \cdots, \lambda_N\right)$ testing*[24]

$$H_0:\ f = f_0 \quad against \quad H_1:\ f \in U^c$$

such that[25]

$$\mathbb{E}_{f_0}\left(\boldsymbol{\varphi_N}\right) < \exp\left(-\gamma N\right) \quad and \quad \sup_{f \in U^c} \mathbb{E}_f\left(1 - \boldsymbol{\varphi_N}\right) < \exp\left(-\gamma N\right) \tag{3.4.1}$$

for all $N > N_0$, where N_0 is a positive integer.

The following sketch of proof gives the intuition behind the two sufficient conditions. Note that the posterior probability of U^c is given by

$$\begin{aligned} \Pi\left(U^c | \lambda_{1:N}\right) &= \frac{\int_{U^c} \prod_{i=1}^N \frac{f(\lambda_i)}{f_0(\lambda_i)} d\Pi\left(f\right)}{\int_{\mathcal{F}} \prod_{i=1}^N \frac{f(\lambda_i)}{f_0(\lambda_i)} d\Pi\left(f\right)} \equiv \frac{\text{numer}_N}{\text{denom}_N} \\ &\leq \boldsymbol{\varphi_N} + \frac{\left(1 - \boldsymbol{\varphi_N}\right) \text{numer}_N}{\text{denom}_N}, \end{aligned} \tag{3.4.2}$$

and we want it to be arbitrarily small.

[24] $\boldsymbol{\varphi_N} = 0$ favors the null hypothesis H_0, whereas $\boldsymbol{\varphi_N} = 1$ favors the alternative hypothesis H_1.
[25] $\mathbb{E}_{f_0}\left(\boldsymbol{\varphi_N}\right)$ and $\sup_{f \in U^c} \mathbb{E}_f\left(1 - \boldsymbol{\varphi_N}\right)$ can be interpreted as type-I and type-II errors, respectively.

First, based on the Borel-Cantelli lemma, the condition on the type-I error suggests that the first term $\boldsymbol{\varphi_N} \to 0$ almost surely.

Second, for the numerator of the second term, the condition on the type-II error implies that

$$\begin{aligned}
\mathbb{E}_{f_0}\left(\left(1-\boldsymbol{\varphi_N}\right)\text{numer}_N\right) &= \int\left(1-\boldsymbol{\varphi_N}\right)\cdot\int_{U^c}\prod_{i=1}^{N}\frac{f\left(\lambda_i\right)}{f_0\left(\lambda_i\right)}d\Pi\left(f\right)\cdot\prod_{i=1}^{N}f_0\left(\lambda_i\right)d\lambda_i \\
&= \int_{U^c}\int\left(1-\boldsymbol{\varphi_N}\right)\prod_{i=1}^{N}f\left(\lambda_i\right)d\lambda_i\cdot d\Pi\left(f\right) \\
&\leq \sup_{f\in U^c}\mathbb{E}_f\left(\left(1-\boldsymbol{\varphi_N}\right)\right) \\
&< \exp\left(-\gamma N\right).
\end{aligned}$$

Hence, $\exp\left(\frac{\gamma N}{2}\right)\left(1-\boldsymbol{\varphi_N}\right)\text{numer}_N \to 0$ almost surely.

Third, for the denominator of the second term, as $N \to 0$,

$$\text{denom}_N = \int_{\mathcal{F}}\exp\left(-\sum_{i=1}^{N}\log\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}\right)d\Pi\left(f\right) \to \int_{\mathcal{F}}\exp\left(-N\cdot d_{KL}\left(f_0,f\right)\right)d\Pi\left(f\right).$$

Combine it with the KL property $f_0 \in KL\left(\Pi\right)$, then

$$\liminf_{N\to\infty} e^{\tilde{\gamma}N}\cdot\text{denom}_N = \infty, \text{ for all } \tilde{\gamma} > 0.$$

Hence, $\exp\left(\frac{\gamma N}{4}\right)\text{denom}_N \to \infty$ almost surely.

Therefore, the posterior probability of U^c

$$\Pi\left(U^c|\lambda_{1:N}\right) \to 0, \ a.s.$$

Schwartz (1965) Theorem guarantees posterior consistency in a general density estimation context. However, as mentioned in the introduction, there are a number of challenges in

adapting these two conditions even to the baseline setup with random effects. The first challenge is that, because we observe y_{it} rather than λ_i, we need to disentangle the uncertainties generated from unknown cross-sectional heterogeneities λ_i's and from independent shocks u_{it}'s. Second is to incorporate unknown shock size σ^2. Third is to take care of the lagged dependent variables as covariates.

In all these scenarios, note that:

(1) The KL requirement ensures that the prior puts positive weight on the true distribution. To satisfy the KL requirement, we need some joint assumptions on the true distribution f_0 and the prior Π. Compared to general nonparametric Bayesian modeling, the DPM structure (and the MGLR_x structure for the correlated random effects model) offers more regularities on the prior Π and thus weaker assumptions on the true distribution f_0 (see Lemma 3.4.8 and Assumption 3.4.14).

(2) Uniformly exponentially consistent tests guarantee that the data is informative enough to differentiate the true distribution from the alternatives. These tests are not specific to the DPM setup but closely related to the definition of the weak neighborhood, hence linked to the identification argument as well.

In the following discussion, I will tackle the aforementioned three challenges one by one.

Disentangle Skills and Shocks Now let us consider a simple cross-sectional case where $\beta = 0$, $\sigma^2 = 1$, and $T = 1$. Since there is only one period, the t subscript is omitted.

$$y_i = \lambda_i + u_i, \quad u_i \sim N(0, 1), \tag{3.4.3}$$

The only twist here is to distinguish the uncertainties originating from unknown individual effects λ_i's and from independent shocks u_i's. Note that unlike previous studies that estimate distributions of observables,[26] here the target λ_i intertwines with u_i and cannot be easily

[26]Some studies (Amewou-Atisso *et al.*, 2003; Tokdar, 2006) estimate distributions of quantities that can be inferred from observables given common coefficients. For example, in the linear regression problems with

inferred from the observed y_i.

Proposition 3.4.7. *(Baseline Model: Skills vs Shocks)*

In setup (3.4.3) with the random effects version of Assumption 3.4.1 (1-3), if $f_0 \in KL\left(\Pi^f\right)$, *the posterior is weakly consistent at* f_0.

At the first glance, Proposition 3.4.7 looks similar to the classical Schwartz (1965) theorem. However, here both the KL requirement and the uniformly exponentially consistent tests are constructed on the observed y_i whereas the weak consistency result is established on the unobserved λ_i. There is a gap between the two, as previously mentioned.

The KL requirement is achieved through the convexity of the KL divergence. In terms of the tests, intuitively, if we obtain enough data and know the distribution of the shocks, it is possible to separate the signal λ_i from the noise u_i even in the cross-sectional setting. The exact argument is delivered via proof by contradiction that utilizes characteristic functions to uncouple the effects from λ_i and u_i. Please refer to Appendix B.3.1 for the detailed proof.

Previous studies have proposed many sets of conditions to ensure that f_0 is in the KL support of Π^f. Based on Wu and Ghosal (2008) Theorem 5, the next lemma gives one set of conditions for f_0 together with the Gaussian-mixture DPM prior,[27]

$$\begin{aligned} \lambda_i &\sim N\left(\mu_i, \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) &\overset{iid}{\sim} G, \\ G &\sim DP\left(\alpha, G_0\right). \end{aligned}$$

Lemma 3.4.8. *(Wu and Ghosal, 2008: Gaussian)*

If f_0 *and its prior* G_0 *satisfy the following conditions:*

---

an unknown error distribution, i.e. $y_i = \beta' x_i + u_i$, conditional on the regression coefficients β, $u_i = y_i - \beta' x_i$ is inferable from the data.

[27] In this section, the nonparametric Bayesian priors are in the form of equations (3.2.5) and (3.2.7), which are more suitable for the posterior consistency analysis.

1. $f_0(\lambda)$ *is a continuous density on* $\mathbb{R}$.
2. *For some* $0 < M < \infty$, $0 < f_0(\lambda) \leq M$ *for all* λ.
3. $\left|\int f_0(\lambda) \log f_0(\lambda)\, d\lambda\right| < \infty$.
4. *For some* $\delta > 0$, $\int f_0(\lambda) \log \frac{f_0(\lambda)}{\varphi_\delta(\lambda)} d\lambda < \infty$, *where* $\varphi_\delta(\lambda) = \inf_{\|\lambda' - \lambda\| < \delta} f_0(\lambda')$.
5. *For some* $\eta > 0$, $\int |\lambda|^{2(1+\eta)} f_0(\lambda)\, d\lambda < \infty$.
6. G_0 *has full support on* $\mathbb{R} \times \mathbb{R}^+$.

then $f_0 \in KL\left(\Pi^f\right)$.

Conditions 1-5 ensure that the true distribution f_0 is well-behaved, and condition 6 further guarantees that the DPM prior is general enough to contain the true distribution.

If the true distribution f_0 has heavy tails, we can resort to Lemma B.5.1 following Tokdar (2006) Theorem 3.3. Lemma B.5.1 ensures the posterior consistency of Cauchy f_0 when G_0 is the standard conjugate normal-inverse-gamma distribution.

Unknown Shock Size Most of the time in practice, we do not know the shock variances in advance. In this part, I consider cross-sectionally homoskedastic shocks with unknown variance as in the baseline model. The cross-sectional heteroskedasticity scenario can be found in Subsection 3.5.4. Now consider a panel setting $(T > 1)$[28] with $\beta = 0$:

$$y_{it} = \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right), \tag{3.4.4}$$

where σ^2 is unknown with the true value being σ_0^2. The joint posterior consistency for $\left(\sigma^2, f\right)$ is stated in the following proposition.

Proposition 3.4.9. *(Baseline Model: Unknown Shock Size)*

In setup (3.4.4) with the random effects version of Assumption 3.4.1, if $f_0 \in KL\left(\Pi^f\right)$ *and* $\sigma_0^2 \in supp\left(\Pi^{\sigma^2}\right)$, *the posterior is weakly consistent at* $\left(\sigma_0^2, f_0\right)$.

[28] Note that when λ_i and u_{it} are both Gaussian with unknown variances, we cannot separately identify the variances in the cross-sectional setting $(T = 1)$. This is no longer a problem if either of the distributions is non-Gaussian or if we work with panel data.

Paralleling the previous subsection, we can refer to Lemma 3.4.8 for conditions that ensure $f_0 \in KL\left(\Pi^f\right)$.

Appendix B.3.1 provides the complete proof. The KL requirement is satisfied based on the dominated convergence theorem. The intuition behind the tests is to split the alternative region of $\left(\sigma^2, f\right)$ into two parts. First, when a candidate σ^2 is far from the true σ_0^2, we can employ orthogonal forward differencing to get rid of λ_i (see Appendix B.4.1), and then use the residues to construct a sequence of tests which distinguish Gaussian distributions with different variances. Second, when σ^2 is close to σ_0^2 but f is far from f_0, we need to make sure that the deviation generated from σ^2 is small enough so that it cannot offset the difference in f.

Lagged Dependent Variables Lagged dependent variables are essential for predictions, as persistence is usually an important feature of economic data. Now let us add a one-period lag of y_{it} to the right hand side of equation (3.5.4), which gives exactly the baseline model (3.1.1):

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

where $\vartheta = \left(\beta, \sigma^2\right)$ are unknown with the true value being $\vartheta_0 = \left(\beta_0, \sigma_0^2\right)$. The following assumption ensures the existence of the required tests in the presence of a linear regressor.

Assumption 3.4.10. *(Initial Conditions)*

y_{i0} is compactly supported.

Proposition 3.4.11. *(Baseline Model: Random Effects)*

In the baseline setup (3.1.1) with random effects, suppose we have:

1. *The random effects version of Assumption 3.4.1.*
2. *y_{i0} satisfies Assumption 3.4.10.*
3. *f and G satisfy Lemma 3.4.8.*
4. *$\vartheta_0 \in \operatorname{supp}\left(\Pi^\vartheta\right)$.*

Then, the posterior is weakly consistent at (ϑ_0, f_0).

The proof can be found in Appendix B.3.1. The KL requirement is established as in previous cases. The uniformly exponentially consistent tests are constructed by dividing the alternative region into two parts: the tests on β and σ^2 are achieved via orthogonal forward differencing followed by a linear regression, while the tests on f are crafted to address the non-i.i.d. observables due to the AR(1) term.

Once again, we can refer to Tokdar (2006) Theorem 3.3 in order to account for heavy tails in the true unknown distributions. For further details, please see Proposition B.5.3 regarding the general model (3.5.1).

Correlated Random Effects Model

In the young firm example, the correlated random effects model can be interpreted as that a young firm's initial performance may reflect its underlying skill, which is a more sensible assumption.

For the correlated random effects model, the definitions and notations are parallel with the random effects ones with slight adjustment considering that now f is a conditional distribution. In the baseline setup, the conditioning set $c_i = y_{i0}$. As in Pati *et al.* (2013), it is helpful to link the properties of the conditional densities to the corresponding ones of the joint densities, which circumvents the difficulty associated with an uncountable set of conditional densities. Let $\mathcal{C}$ be a compact subset of $\mathbb{R}$ for the conditioning variable $c_i = y_{i0}$, $\mathcal{H}$ be the set of joint densities on $\mathbb{R} \times \mathcal{C}$ (with respect to Lebesgue measure), and $\mathcal{F}$ be the set of conditional densities on $\mathbb{R}$ given conditioning variable $c \in \mathcal{C}$.

Let h, f, and q be the joint, conditional, and marginal densities, respectively. Denote

$$h_0(\lambda, c) = f_0(\lambda|c) \cdot q_0(c), \quad h(\lambda, c) = f(\lambda|c) \cdot q_0(c).$$

where $h,\ h_0 \in \mathcal{H}$, and $f,\ f_0 \in \mathcal{F}$. h_0, f_0, and q_0 are the true densities. Note that h and h_0 share the same marginal density q_0, but different conditional densities f and f_0. This setup does not require estimating q_0 and thus relaxes the assumption on the initial conditions.

The definitions of weak neighborhood and KL property rely on the joint density characterization. Note that in both definitions, the conditioning variable c is integrated out with respect to the true q_0.

Definition 3.4.12. A *weak neighborhood* of f_0 is defined as

$$U_{\epsilon,\Phi}(f_0) = \left\{ f \in \mathcal{F} : \left| \int \varphi_j h - \int \varphi_j h_0 \right| < \epsilon \right\}$$

where $\epsilon > 0$ and $\Phi = \{\varphi_j\}_{j=1}^{J}$ are bounded, continuous functions of (λ, c).

Definition 3.4.13. If for all $\epsilon > 0$, $\Pi^f(f \in \mathcal{F} : d_{KL}(h_0, h) < \epsilon) > 0$, we say f_0 is *in the KL support of* Π^f, or $f_0 \in KL\left(\Pi^f\right)$.

As described in Subsection 3.2.3, the MGLR_x prior is a conditional version of the nonparametric Bayesian prior. It can be specified as follows, with the conditioning set simply being a scalar, y_{i0}.

$$\begin{aligned}
\lambda_i | y_{i0} &\sim N\left(\lambda_i;\ \mu_i [1, y_{i0}]',\ \omega_i^2\right), \\
\left(\mu_i, \omega_i^2\right) \equiv \theta_i &\overset{iid}{\sim} G(\cdot;\ y_{i0}), \\
G(\cdot;\ y_{i0}) &= \sum_{k=1}^{\infty} p_k(y_{i0})\, \delta_{\theta_k}.
\end{aligned}$$

where for components $k = 1, 2, \cdots$

$$\begin{aligned}
\theta_k &\sim G_0, \\
p_k(y_{i0}) &= \Phi(\zeta_k(y_{i0})) \prod_{j<k} (1 - \Phi(\zeta_j(y_{i0}))), \\
\zeta_k &\sim GP(0, V_k).
\end{aligned}$$

The induced prior on the mixing measures $G(\theta_i;\ y_{i0})$ is denoted as $\tilde{\Pi}$.

Assumption 3.4.14. *(Baseline Model: Correlated Random Effects)*

1. *Conditions on f_0:*
   (a) *For some $0 < M < \infty$, $0 < f_0(\lambda|y_0) \leq M$ for all (λ, y_0).*
   (b) $\left|\int \left[\int f_0(\lambda|y_0) \log f_0(\lambda|y_0)\, d\lambda\right] q_0(y_0)\, dy_0\right| < \infty$.
   (c) $\left|\int \left[\int f_0(\lambda|y_0) \log \frac{f_0(\lambda|y_0)}{\varphi_\delta(\lambda|y_0)} d\lambda\right] q_0(y_0)\, dy_0\right| < \infty$,
   where $\varphi_\delta(\lambda|y_0) = \inf_{|\lambda'-\lambda|<\delta} f_0(\lambda|y_0)$, for some $\delta > 0$.
   (d) *For some $\eta > 0$, $\int \left[\int |\lambda|^{2(1+\eta)} f_0(\lambda|y_0)\, d\lambda\right] q_0(y_0)\, dy_0 < \infty$.*
   (e) *$f_0(\cdot|\cdot)$ is jointly continuous in (λ, y_0).*
   (f) *$q_0(y_0) > 0$ for all $y_0 \in \mathcal{C}$.*
2. *Conditions on $\tilde{\Pi}$:*
   (a) *For $k = 1, 2, \cdots$, V_k is chosen such that $\zeta_k \sim GP(0, V_k)$ has continuous path realizations.*
   (b) *For $k = 1, 2, \cdots$, for any continuous $g(\cdot)$, and any $\epsilon > 0$, $\tilde{\Pi}(\sup_{y_0 \in \mathcal{C}} |\zeta_k(y_0) - g(y_0)| < \epsilon) > 0$.*
   (c) *G_0 is absolutely continuous.*

These conditions follow Assumptions A1-A5 and S1-S3 in Pati *et al.* (2013) for posterior consistency under the conditional density topology. The first group of conditions can be viewed as conditional density analogs of the conditions in Lemma 3.4.8. These requirements are satisfied for flexible classes of models, i.e. generalized stick-breaking process mixtures with the stick-breaking lengths being monotone differentiable functions of a continuous stochastic process.

Proposition 3.4.15. *(Baseline Model: Correlated Random Effects)*

In the baseline setup (3.1.1) with correlated random effects, suppose we have:

1. *Assumption 3.4.1.*
2. *y_{i0} satisfies Assumption 3.4.10.*

3. *f and G satisfy Assumption 3.4.14.*

4. $\vartheta_0 \in \text{supp}\left(\Pi^{\vartheta}\right)$.

Then, the posterior is weakly consistent at (ϑ_0, f_0).

The proof in Appendix B.3.2 is similar to the random effects case except that now the KL property and the uniformly exponentially consistent tests are on the joint distribution of (λ_i, y_{i0}).

3.4.4 Density forecasts

Once the posterior consistency results are obtained, we can bound the discrepancy between the proposed predictor and the oracle by the estimation uncertainties in β, σ^2, and f, and then show the asymptotical convergence of the density forecasts to the oracle forecast (see Appendix B.3.3 for the detailed proof).

Proposition 3.4.16. *(Baseline Model: Density Forecasts)*

In the baseline setup (3.1.1), suppose we have:

1. *For the random effects model, conditions in Proposition 3.4.11.*
2. *For the correlated random effects model,*
   (a) *conditions in Proposition 3.4.15,*
   (b) *$q_0(y_0)$ is continuous, and there exists $\underline{q} > 0$ such that $|q_0(y_0)| > \underline{q}$ for all $y_0 \in \mathcal{C}$.*

Then, the density forecasts converge to the oracle predictor in the following two ways:

1. *Convergence of $f^{cond}_{i,T+1}$ in weak topology: for any i and any $U_{\epsilon,\Phi}\left(f^{oracle}_{i,T+1}\right)$, as $N \to \infty$,*

$$\mathbb{P}\left(f^{cond}_{i,T+1} \in U_{\epsilon,\Phi}\left(f^{oracle}_{i,T+1}\right) \middle| y_{1:N,0:T}\right) \to 1, \ a.s.$$

2. *"Pointwise" convergence of $f^{sp}_{i,T+1}$: for any i, any y, and any $\epsilon > 0$, as $N \to \infty$,*

$$\left|f^{sp}_{i,T+1}(y) - f^{oracle}_{i,T+1}(y)\right| < \epsilon, \ a.s.$$

The first result focuses on the conditional predictor (3.2.1) and is more coherent with the weak topology for posterior consistency in the previous subsection. The second result is established for the semiparametric Bayesian predictor (3.2.3), which is the posterior mean of the conditional predictor. In addition, the asymptotic convergence of aggregate-level density forecasts can be derived by summing individual-specific forecasts over different subcategories.

3.5 Extensions

3.5.1 General Panel Data Model

The general panel data model with correlated random coefficients can be specified as

$$y_{it} = \beta' x_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right) \tag{3.5.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. Similar to the baseline setup in Subsection 3.2.1, the y_{it} is the observed individual outcomes, and I am interested in providing density forecasts of $y_{i,T+1}$ for any individual i.

The $w_{i,t-1}$ is a vector of observed covariates that have heterogeneous effects on the outcomes, with λ_i being the unobserved individual heterogeneities. $w_{i,t-1}$ is strictly exogenous and captures the key sources of individual heterogeneities. The simplest choice would be $w_{i,t-1} = 1$ where λ_i can be interpreted as an individual-specific intercept, i.e. firm i's skill level in the baseline model (3.1.1). Moreover, it is also helpful to include other key covariates of interest whose effects are more diverse cross-sectionally, such as observables that characterize innovation activities. Furthermore, the current setup can also take into account deterministic or stochastic aggregate effects, such as time dummies for the recent recession. For notation clarity, I decompose $w_{i,t-1} = \left(w_{t-1}^{A\prime}, w_{i,t-1}^{I\prime}\right)'$, where w_{t-1}^{A} stands for a vector of aggregate variables, and $w_{i,t-1}^{I}$ is composed of individual-specific variables. In the simple individual-specific-intercept case, we have $w_{t-1}^{A} = 1$ for all t, and the corresponding scalar λ_i's give the values for the heterogeneous intercepts.

The $x_{i,t-1}$ is a vector of observed covariates that have homogeneous effects on the outcomes, and β is the corresponding vector of common parameters. $x_{i,t-1}$ can be either strictly exogenous or predetermined, which can be further denoted as $x_{i,t-1} = \left(x_{i,t-1}^{O\prime}, x_{i,t-1}^{P\prime}\right)'$, where $x_{i,t-1}^{O}$ is the strictly exogenous part while $x_{i,t-1}^{P}$ is the predetermined part. The one-period-lagged outcome $y_{i,t-1}$ is a typical candidate for $x_{i,t-1}^{P}$ in the dynamic panel data literature, which captures the persistence structure. In addition, both $x_{i,t-1}^{O}$ and $x_{i,t-1}^{P}$ can incorporate other general control variables, such as firm characters as well as local and national economic conditions. The notation $x_{i,t-1}^{P*}$ indicates the subgroup of $x_{i,t-1}^{P}$ excluding lagged outcomes. Here, the distinction between homogeneous effects ($\beta' x_{i,t-1}$) versus heterogeneous effects ($\lambda_i' w_{i,t-1}$) allows us to enjoy the best of both worlds—revealing the latent nonstandard structures for the key effects while avoiding the curse-of-dimensionality problem, which shares the same idea as Burda *et al.* (2012).

The u_{it} is an individual-time-specific shock characterized by zero mean and cross-sectional heteroskedasticity, σ_i^2. The normality assumption is not very restrictive due to the flexibility in σ_i^2 distribution. Table 1 in Fernandez and Steel (2000) demonstrates that scale mixture of normals can capture "a rich class of continuous, symmetric, and unimodal distributions" (p. 81), including Cauchy, Laplace, Logistic, etc. More rigorously, as proved by Kelker (1970), this class is composed of marginal distributions of higher-dimensional spherical distributions.

In the correlated random coefficients model, λ_i can depend on some of the covariates and initial conditions. Specifically, I define the conditioning set at period t to be

$$c_{i,t-1} = \left\{y_{i,0:t-1},\ x_{i,0:t-1}^{P*},\ x_{i,0:T}^{O},\ w_{i,0:T}\right\} \tag{3.5.2}$$

and allow the distribution of λ_i to be a function of c_{i0}. Note that as lagged y_{it} and $x_{i,t-1}^{P*}$ are predetermined variables, the sequences of $x_{i,t-1}^{P*}$ in the conditioning set $c_{i,t-1}$ start from period 0 to period $t-1$; while $x_{i,t-1}^{O}$ and $w_{i,t-1}$ are both strictly exogenous, so the conditioning set $c_{i,t-1}$ contains their entire sequences. For future use, I also define the part of

$c_{i,t-1}$ that is composed of individual-specific variables as

$$c^*_{i,t-1} = \left\{ y_{i,0:t-1},\ x^{P*}_{i,0:t-1},\ x^{O}_{i,0:T},\ w^{I}_{i,0:T} \right\}.$$

3.5.2 Posterior Samplers

Random Coefficients Model

Compared to Subsection 3.3.1 for the baseline setup, the major change here is to account for cross-sectional heteroskedasticity via another flexible prior on the distribution of σ_i^2. Define $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$ where $\underline{\sigma}^2$ is some small positive number. Then, the support of $f_0^{\sigma^2}$ is bounded below by $\underline{\sigma}^2$ and thus satisfies the requirement for the asymptotic convergence of the density forecasts in Proposition 3.5.12.[29] The log transformation ensures an unbounded support for l_i so that Algorithm 3.3.1 with Gaussian-mixture DPM prior can be directly employed. Beyond cross-sectional heteroskedasticity, there is a minor alternation due to the (potentially) multivariate λ_i. In this scenario, the component mean μ_k is a vector and component variance Ω_k is a positive definite matrix.

The following algorithm parallels Algorithm 3.3.1. Both algorithms are based on truncation approximation, which is relatively easy to implement and enjoys good mixing properties. For the slice-retrospective sampler, please refer to Algorithm B.2.4 in the Appendix.

Denote $D = \{\{D_i\}, D_A\}$ as a shorthand for the data sample used for estimation, where $D_i = c^*_{i,T}$ contains the observed data for individual i, and $D_A = w^A_{0:T}$ is composed of the aggregate regressors with heterogeneous effects. Note that because λ_i and σ_i^2 are independent with respect to each other, their mixture structures are completely separate. As their mixture structures are almost identical, I define a generic variable z which can represent either λ or l, and then include z as a superscript to indicate whether a specific parameter

[29] Note that only Proposition 3.5.12 for density forecasts needs a positive lower bound on the distribution of σ_i^2. The propositions for identification and posterior consistency of the estimates are not restricted to but can accommodate such requirement.

belongs to the λ part or the l part. Most of the conditional posteriors are either similar to Algorithm B.2.4 or standard for posterior sampling (see Appendix B.2.3), except for the additional term $\left(\sigma_i^2 - \underline{\sigma}^2\right)^{-1}$ in step 4-b, which takes care of the change of variables from $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$ to σ_i^2.

Algorithm 3.5.1. *(General Model: Random Coefficients)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities: For* $z = \lambda,\ l$,
   (a) *Draw* $\alpha^{z(s)}$ *from a gamma distribution* $p\left(\alpha^{z(s)} \middle| p_{K^z}^{z(s-1)}\right)$.
   (b) *For* $k^z = 1, \cdots, K^z$, *draw* $p_{k^z}^{z(s)}$ *from the truncated stick breaking process* $p\left(\left\{p_{k^z}^{z(s)}\right\} \middle| \alpha^{z(s)}, \left\{n_{k^z}^{z(s-1)}\right\}\right)$.
2. *Component parameters: For* $z = \lambda,\ l$, *for* $k^z = 1, \cdots, K^z$, *draw* $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ *from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if* z *is a scalar)* $p\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
3. *Component memberships: For* $z = \lambda,\ l$, *for* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \middle| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\},\ z_i^{(s-1)}\right)$.
4. *Individual-specific parameters:*
   (a) *For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a multivariate-normal distribution (or a normal distribution if* λ *is a scalar)* $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A\right)$.
   (b) *For* $i = 1, \cdots, N$, *draw* $\left(\sigma_i^2\right)^{(s)}$ *via the random-walk Metropolis-Hastings approach*

$$
\begin{aligned}
& p\left(\left(\sigma_i^2\right)^{(s)} \middle| \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A\right) \\
\propto & \left(\left(\sigma_i^2\right)^{(s)} - \underline{\sigma}^2\right)^{-1} \phi\left(\log\left(\left(\sigma_i^2\right)^{(s)} - \underline{\sigma}^2\right);\ \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}\right) \\
& \cdot \prod_{t=1}^{T} \phi\left(y_{it};\ \lambda_i^{(s)\prime} w_{i,t-1} + \beta^{(s-1)\prime} x_{i,t-1}, \left(\sigma_i^2\right)^{(s)}\right).
\end{aligned}
$$

5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model* $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\},\ D\right)$.

Correlated Random Coefficients Model

Regarding conditional density estimation, I impose the MGLR_{x} prior on both λ_i and l_i. Compared to Algorithm 3.3.2 for the baseline setup, the algorithm here makes the following changes: (1) generic variable $z = \lambda,\ l$, (2) $\left(\sigma_i^2 - \underline{\sigma}^2\right)^{-1}$ in step 4-b, (3) vector λ_i, and (4) vector conditioning set c_{i0}. The conditioning set c_{i0} is characterized by equation (3.5.2) for balanced panels or equation (3.5.3) for unbalanced panels. In practice, it is more computationally efficient to incorporate a subset of c_{i0} or a function of c_{i0} guided by the specific problem at hand.

Algorithm 3.5.2. *(General Model: Correlated Random Coefficients)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities: For* $z = \lambda,\ l$,
   (a) *For* $k^z = 1, \cdots, K^z - 1$, *draw* $A_{k^z}^{z(s)}$ *via the random-walk Metropolis-Hastings approach,* $p\left(A_{k^z}^{z(s)} \middle| \boldsymbol{\zeta}_{k^z}^{z(s-1)}, \{c_{i0}\}\right)$ *and then calculate* $\boldsymbol{V}_k^{(s)}$.
   (b) *For* $k^z = 1, \cdots, K^z - 1$, *and* $i = 1, \cdots, N$, *draw* $\xi_{k^z}^{z(s)}(c_{i0})$ *from a truncated normal distribution* $p\left(\xi_{k^z}^{z(s)}(c_{i0}) \middle| \zeta_{k^z}^{z(s-1)}(c_{i0}), \gamma_i^{z(s-1)}\right)$.
   (c) *For* $k^z = 1, \cdots, K^z - 1$, $\boldsymbol{\zeta}_{k^z}^{z(s)}$ *from a multivariate normal distribution* $p\left(\boldsymbol{\zeta}_{k^z}^{z(s)} \middle| \boldsymbol{V}_{k^z}^{z(s)}, \boldsymbol{\xi}_{k^z}^{z(s)}\right)$.
   (d) *For* $k^z = 1, \cdots, K^z - 1$, *and* $i = 1, \cdots, N$, *the component probabilities* $p_{k^z}^{z(s)}(c_{i0})$ *are fully determined by* $\boldsymbol{\zeta}_{k^z}^{z(s)}$.
2. *Component parameters: For* $z = \lambda,\ l$, *for* $k^z = 1, \cdots, K^z$,
   (a) *Draw* $\mu_{k^z}^{z(s)}$ *from a matricvariate-normal distribution (or a multivariate-normal distribution if* z *is a scalar)* $p\left(\mu_{k^z}^{z(s)} \middle| \Omega_{k^z}^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
   (b) *Draw* $\Omega_{k^z}^{z(s)}$ *from an inverse-Wishart distribution (or an inverse-gamma distribution if* z *is a scalar)* $p\left(\Omega_{k^z}^{z(s)} \middle| \mu_{k^z}^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
3. *Component memberships: For* $z = \lambda,\ l$, *for* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \middle| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\},\ z_i^{(s-1)}, c_{i0}\right)$.

4. *Individual-specific parameters:*

   (a) *For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a multivariate-normal distribution (or a normal distribution if* λ *is a scalar)* $p\left(\lambda_i^{(s)} \left| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A \right.\right)$.

   (b) *For* $i = 1, \cdots, N$, *draw* $\left(\sigma_i^2\right)^{(s)}$ *via the random-walk Metropolis-Hastings approach* $p\left(\left(\sigma_i^2\right)^{(s)} \left| \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A \right.\right)$.

5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model* $p\left(\beta^{(s)} \left| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D \right.\right)$.

3.5.3 Identification

Assumption 3.5.3. *(General Model: Setup)*

1. *Conditional on* $w_{0:T}^A$, $\left\{c_{i0}^*, \lambda_i, \sigma_i^2\right\}$ *are i.i.d. across* i.
2. *For all* t, *conditional on* $\{y_{it}, c_{i,t-1}\}$, x_{it}^{P*} *is independent of* $\left\{\lambda_i, \sigma_i^2\right\}$ *and* β.
3. $\left\{x_{i,0:T}^O, w_{i,0:T}\right\}$ *are independent of* $\left\{\lambda_i, \sigma_i^2\right\}$ *and* β.
4. *Let* $u_{it} = \sigma_i v_{it}$. v_{it} *is i.i.d. across* i *and* t *and independent of* $c_{i,t-1}$.

Remark 3.5.4. (i) For the random effects case, the first condition can be altered to "$\left\{\lambda_i, \sigma_i^2\right\}$ are independent of c_{i0} and i.i.d. across i".

(ii) For the distribution of the shock u_{it}, a general class of shock distributions can be accommodated by the scale mixture of normals generated from the flexible distribution of σ_i^2 (Kelker, 1970; Fernandez and Steel, 2000). It is possible to allow some additional flexibility in the distribution of u_{it}. For example, the identification argument still holds as long as (1) v_{it} is i.i.d. across i and independent over t, and (2) the distributions of v_{it}, $f_t^v(v_{it})$, have known functional forms, such that $\mathbb{E}[v_{it}] = 0$, $\mathbb{V}[v_{it}] = 1$. Nevertheless, as this paper studies panels with short time spans, time-varying shock distribution may not play a significant role. I will keep the normality assumption in the rest of this paper to streamline the arguments.

Assumption 3.5.5. *(General Model: Identification) For all* i,

1. *The common parameter vector* β *is identifiable.*[30]

[30]The identification of common parameters in panel data models is standard in the literature. For

2. *$w_{i,0:T-1}$ has full rank d_w.*

3. *Conditioning on c_{i0}, λ_i and σ_i^2 are independent of each other.*

4. *The characteristic functions for $\lambda_i|c_{i0}$ and $\sigma_i^2|c_{i0}$ are non-vanishing almost everywhere.*

Proposition 3.5.6. *(General Model: Identification)*

Under Assumptions 3.5.3 and 3.5.5, the common parameters β and the conditional distribution of individual effects, $f^{\lambda}(\lambda_i|c_{i0})$ and $f^{\sigma^2}(\sigma_i^2|c_{i0})$, are all identified.

Please refer to Appendix B.4.1 for the proof. Assumption 3.5.3-3.5.5 and Proposition 3.5.6 are similar to Assumption 2.1-2.2 and Theorem 2.3 in Liu *et al.* (2016) except for the treatment of heteroskedasticity. First, this paper supports unobserved cross-sectional heteroskedasticity whereas Liu *et al.* (2016) incorporate cross-sectional heteroskedasticity as a parametric function of observables. Second, Liu *et al.* (2016) allow for time-varying heteroskedasticity whereas the identification restriction in this paper can only permit time-varying distribution for v_{it} (see Remark 3.5.4 (ii)) while keeping zero mean and unit variance. However, considering that this paper focuses on the scenarios with short time dimension, lack of time-varying heteroskedasticity would not be a major concern.

Furthermore, the above identification results can be extended to unbalanced panels. Let T_i denote the longest chain for individual i that has complete observations, from t_{0i} to t_{1i}. That is, $\{y_{it},\ w_{i,t-1},\ x_{i,t-1}\}$ are observed for all $t = t_{0i}, \cdots, t_{1i}$. Then, I discard the unobserved periods and redefine the conditioning set at time $t = 1, t_{0i}, \cdots, t_{1i}, T+1$ to be

$$c_{i,t-1} = \left\{ y_{i,\tau_{i,t-1}^P},\ x_{i,\tau_{i,t-1}^P}^{P*},\ x_{i,\tau_{iT}^P}^{O},\ w_{i,\tau_{iT}^P} \right\}, \tag{3.5.3}$$

where the set for time periods $\tau_{i,t-1}^P = \{0, t_{0i}-1, \cdots, t_{1i}-1, T\} \cap \{0, \cdots, t-1\}$. Note that t_{i0} can be 1, and t_{i1} can be T, so this structure is also able to accommodate balanced panels.

---

example, there have been various ways to difference data across t to remove the individual effects λ_i (e.g. orthogonal forward differencing, see Appendix B.4.1), and we can construct moment conditions based on the transformed data to identify the common parameters β. Here I follow Liu *et al.* (2016) and state a high-level identification assumption.

Accordingly, the individual-specific component of $c_{i,t-1}$ is

$$c^*_{i,t-1} = \left\{ y_{i,\tau^P_{i,t-1}},\ x^{P*}_{i,\tau^P_{i,t-1}},\ x^O_{i,\tau^P_{iT}},\ w^I_{i,\tau^P_{iT}} \right\}.$$

Assumption 3.5.7. *(Unbalanced Panels) For all i,*

1. *c_{i0} is observed.*
2. *x_{iT} and w_{iT} are observed.*
3. *The common parameter vector β is identifiable.*
4. *$w_{i,(t_{0i}-1):(t_{1i}-1)}$ has full rank d_w.*

The first condition guarantees the existence of the initial conditioning set for the correlated random coefficients model. In practice, it is not necessary to incorporate all initial values of the predetermined variables and the whole series of the strictly exogenous variables. It is more feasible to only take into account a subset of c_{i0} or a function of c_{i0} that is relevant for the specific analysis. The second condition ensures that the covariates in the forecast equation are available in order to make predictions. The third condition is the same as Assumption 3.5.5 (1) that makes a high-level assumption on the identification of common parameters. The fourth condition is the unbalanced panel counterpart of Assumption 3.5.5 (2). It guarantees that the observed chain is long and informative enough to distinguish different aspects of individual effects. Now we can state similar identification results for unbalanced panels.

Proposition 3.5.8. *(Identification: Unbalanced Panels)*

For unbalanced panels, under Assumptions 3.5.3, 3.5.5 (3-4), and 3.5.7, the common parameter vector β and the conditional distributions of individual effects, $f^{\lambda}(\lambda_i|c_{i0})$ and $f^{\sigma^2}(\sigma_i^2|c_{i0})$, are all identified.

3.5.4 Asymptotic Properties

In Subsection 3.5.4, I address posterior consistency of f^{σ^2} with unknown individual-specific heteroskedasticity σ_i^2. In Subsection 3.5.4, I proceed with the general setup (3.5.1) by considering (correlated) random coefficients, adding other strictly exogenous and predetermined covariates into x_{it}, and accounting for unbalanced panels, then the posterior consistency can be obtained with respect to the common parameters vector β and the (conditional) distributions of individual effects, f^{λ} and f^{σ}. In Subsection 3.5.4, I establish the asymptotic properties of the density forecasts.

Let d_z be the dimension of z_{it}, where z is a generic variable which can be w or x. Then, $\Theta = \mathbb{R}^{d_x}$, $\mathcal{F}^{\lambda}$ is a set of (conditional) densities on $\mathbb{R}^{d_w}$, and $\mathcal{F}^{\sigma^2}$ is a set of (conditional) densities on $\mathbb{R}^{+}$. The data sample used for estimation is $D = \{\{D_i\}, D_A\}$ defined in Subsection 3.5.1, which constitutes the conditioning set for posterior inference.

Cross-sectional Heteroskedasticity

In many empirical applications, such as the young firm analysis in Section 3.7, risk may largely vary over the cross-section. Therefore, it is more realistic to address cross-sectional heteroskedasticity, which also contributes considerably to density forecasts. Now let us adapt the simple panel model in equation (3.4.4) to incorporate cross-sectional heteroskedastic shocks.

$$y_{it} = \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right), \tag{3.5.4}$$

where $\beta = 0$, and λ_i is independent of σ_i^2. Their distributions, $f^{\lambda}(\lambda_i)$ and $f^{\sigma^2}\left(\sigma_i^2\right)$, are unknown, with the true distributions being $f_0^{\lambda}(\lambda_i)$ and $f_0^{\sigma^2}\left(\sigma_i^2\right)$, respectively. Their posteriors are consistently estimated as established in the following proposition.

Proposition 3.5.9. *(Cross-sectional Heteroskedasticity)*

In setup (3.5.4) with the random effects version of Assumption 3.5.3 (1 and 4) and Assumption 3.5.5 (3-4), if $f_0^{\lambda} \in KL\left(\Pi^{f^{\lambda}}\right)$ *and* $f_0^{\sigma^2} \in KL\left(\Pi^{f^{\sigma^2}}\right)$*, the posterior is weakly consistent*

$at\ \left(f_0^{\lambda},\ f_0^{\sigma^2}\right)$.

Please refer to Appendix B.4.2 for the complete proof. The KL requirement is again given by the convexity of KL divergence. The intuition of the tests is again to break down the alternatives into two circumstances. First, when a candidate f^{σ^2} and the true $f_0^{\sigma^2}$ are not identical, we can once again rely on orthogonal forward differencing (see Appendix B.4.1) to distinguish variance distributions. Note that the Fourier transformation (i.e. characteristic functions) is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004) instead. The second circumstance comes when the variance distributions are close to each other, but f^{λ} is far from f_0^{λ}. Here I apply the argument for Proposition 3.4.7 with slight adaption.

$f_0^{\lambda} \in KL\left(\Pi^{f^{\lambda}}\right)$ is guaranteed by conditions in Lemma 3.4.8 (or Lemma B.5.1 for true distribution with heavy tails). Concerning $f_0^{\sigma^2}$, I impose a Gaussian-mixture DPM prior on $l = \log\left(\sigma^2 - \underline{\sigma}^2\right)$, and similar sufficient conditions apply to the distribution of l as well.

General Setup

In this subsection, I generalize the setup to the full panel data model in equation (3.5.1) with regard to the following three aspects. The proofs are along the same lines of the baseline model plus cross-sectionally heteroskedasticity.

First, in practice, it is more desirable to consider a vector of λ_i interacting with observed w_{it}. In the young firm example, different young firms may respond differently to the financial crisis, and R&D activities may benefit the young firms in different magnitudes. A (correlated) random coefficient model can capture such heterogeneities and facilitate predictions.

The uniformly exponentially consistent tests for multivariate λ_i are constructed in a similar way as Proposition 3.4.7 outlined in the "disentangle skills and shocks" part of Subsection 3.4.3. Note that for each $l = 1, \cdots, d_w$, we can implement orthogonal forward differencing with respect to all other $\{\lambda_{im}\}_{m \neq l}$ and reduce the problem to λ_{il} versus shocks as in equation

(3.4.3). The same logic still holds when we add lagged dependent variables and other predictors. Furthermore, a multi-dimensional version of Lemma 3.4.8 or Assumption 3.4.14 guarantees the KL property of multivariate λ_i .

Second, additional strictly exogenous ($x^O_{i,t-1}$) and predetermined ($x^{P*}_{i,t-1}$) predictors help control for other sources of variation and gain more accurate forecasts. We can reproduce the analysis for Proposition 3.4.15 by allowing the conditioning set c_{i0} to include the initial values of the predetermined variables and the whole series of the strictly exogenous variables.

Third, it is constructive to account for unbalanced panels with missing observations, which incorporates more data into the estimation and elicits more information for the prediction. The posterior consistency argument is still valid in like manner given Assumption 3.5.7.

Combining above discussions all together, we achieve the posterior consistency result for the general panel data model. The random coefficients model is relatively more straightforward regarding posterior consistency, as the random coefficients setup together with Assumption 3.5.5 (3) implies that $\left(\lambda_i, \sigma_i^2, c_{i0}\right)$ are independent among one another. The theorem for the random coefficients model is stated as follows.

Proposition 3.5.10. *(General Model: Random Coefficients)*

Suppose we have:

1. *Assumptions 3.5.3, 3.5.5 (3-4), 3.5.7, and 3.4.10.*
2. *Lemma 3.4.8 on λ and l.*
3. $\beta_0 \in \text{supp}\left(\Pi^\beta\right)$.

Then, the posterior is weakly consistent at $\left(\beta_0, f_0^\lambda, f_0^{\sigma^2}\right)$.

For heavy tails in the true unknown distributions, Lemma B.5.2 generalizes Lemma B.5.1 to the multivariate scenario, and Proposition B.5.3 gives a parallel posterior consistency result.

In the world of correlated random coefficients, λ_i is independent of σ_i^2 conditional on c_{i0}. In

other words, λ_i and σ_i^2 can potentially depend on the initial condition c_{i0}, and therefore can potentially relate to each other through c_{i0}. For example, a young firm's initial performance may reveal its underlying ability and risk. The following proposition is established for the correlated random coefficients model.

Proposition 3.5.11. *(General Model: Correlated Random Coefficients)*

Under Assumptions 3.5.3, 3.5.5 (3-4), 3.5.7, 3.4.10, and 3.4.14, if $\beta_0 \in supp\left(\Pi^{\beta}\right)$, *the posterior is weakly consistent at* $\left(\beta_0, f_0^{\lambda}, f_0^{\sigma^2}\right)$.

Note that Propositions 3.5.10 and 3.5.11 are parallel with each other, as the first group of conditions in Assumption 3.4.14 is the conditional analog of Lemma 3.4.8 conditions.

Density Forecasts

In the sequel, the next proposition shows convergence of density forecasts in the general model.

Proposition 3.5.12. *(General Model: Density Forecasts)*

In the general model (3.5.1), suppose we have:

1. *For the random coefficients model,*
   (a) *conditions in Proposition 3.5.10,*
   (b) $supp\left(f_0^{\sigma^2}\right)$ *is bounded below by some* $\underline{\sigma}^2 > 0$.
2. *For the correlated random coefficients model,*
   (a) *conditions in Proposition 3.5.11,*
   (b) $q_0(y_0)$ *is continuous, and there exists* $\underline{q} > 0$ *such that* $|q_0(y_0)| > \underline{q}$ *for all* $y_0 \in \mathcal{C}$,
   (c) $supp\left(f_0^{\sigma^2}\right)$ *is bounded below by some* $\underline{\sigma}^2 > 0$.

Then the density forecasts converge to the oracle predictor in the following two ways:

1. *Convergence of* $f_{i,T+1}^{cond}$ *in weak topology: for any* i *and any* $U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right)$, *as* $N \to \infty$,

$$\mathbb{P}\left(f_{i,T+1}^{cond} \in U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right) \middle| y_{1:N,0:T}\right) \to 1, \ a.s.$$

2. *"Pointwise" convergence of* $f_{i,T+1}^{sp}$: *for any* i, *any* y, *and any* $\epsilon > 0$, *as* $N \to \infty$,

$$\left|f_{i,T+1}^{sp}(y) - f_{i,T+1}^{oracle}(y)\right| < \epsilon, \ a.s.$$

The additional requirement that the support of $f_0^{\sigma^2}$ is bounded below ensures that the likelihood would not explode. Then, the proof is in the same vein as the baseline setup.

3.6 Simulation

In this section, I have conducted extensive Monte Carlo simulation experiments to examine the numerical performance of the proposed semiparametric Bayesian predictor. Subsection 3.6.1 describes the evaluation criteria for point forecasts and density forecasts. Subsection 3.6.2 introduces other alternative predictors. Subsection 3.6.3 considers the baseline setup with random effects. Subsection 3.6.4 extends to the general setup incorporating cross-sectional heterogeneity and correlated random coefficients.

3.6.1 Forecast Evaluation Methods

As mentioned in the model setup in Subsection 3.2.1, this paper focuses on one-step-ahead forecasts, but a similar framework can be applied to multi-period-ahead forecasts. The forecasting performance is evaluated along both the point and density forecast dimensions, with particular attention to the latter.

Point forecasts are evaluated via the Mean Square Error (MSE), which resonates with the quadratic loss function. Let $\hat{y}_{i,T+1}$ denote the forecast made by the model,

$$\hat{y}_{i,T+1} = \hat{\beta}' x_{iT} + \hat{\lambda}_i' w_{iT},$$

where $\hat{\lambda}_i$ and $\hat{\beta}$ stand for the estimated parameter values. Then, the forecast error is defined as

$$\hat{e}_{i,T+1} = y_{i,T+1} - \hat{y}_{i,T+1},$$

with $y_{i,T+1}$ being the realized value at time $T+1$. The formula for the MSE is provided in the following equation,

$$MSE = \frac{1}{N}\sum_i \hat{e}^2_{i,T+1}.$$

The Diebold and Mariano (1995) test is further implemented to assess whether or not the difference in the MSE is significant.

The accuracy of the density forecasts is measured by the log predictive score (LPS) as suggested in Geweke and Amisano (2010),

$$LPS = \frac{1}{N}\sum_i \log \hat{p}\left(y_{i,T+1}|D\right),$$

where $y_{i,T+1}$ is the realization at $T+1$, and $\hat{p}\left(y_{i,T+1}|D\right)$ represents the predictive likelihood with respect to the estimated model conditional on the observed data D. I also perform the Amisano and Giacomini (2007) test to examine the significance in the LPS difference.

3.6.2 Alternative Predictors

In the simulation experiments, I compare the proposed semiparametric Bayesian predictor with other alternatives, including Bayesian estimators with the prior of λ_i being a homogeneous prior, a flat prior, a parametric prior, and a DP prior (more rigorously, the DP prior is on f rather than λ_i).

The homogeneous prior is defined as $\lambda_i \sim \delta_{\lambda^*}$, where δ_{λ^*} is the Dirac delta function representing a degenerate distribution $\mathbb{P}\left(\lambda_i = \lambda^*\right) = 1$. Intuitively, this prior believes that all firms share the same level of skill λ^*. Because λ^* is unknown beforehand, it becomes another common parameter, similar to β. Hence I adopt a multivariate-normal-inverse-gamma prior on $\left([\beta,\ \lambda^*]',\ \sigma^2\right)$, which can be viewed as a Bayesian counterpart of the pooled OLS

estimator.

The flat prior is specified as $p(\lambda_i) \propto 1$, an uninformative prior with the posterior mode being the MLE estimate. Roughly speaking, the flat prior infers firm i's skill λ_i only using firm i's history.

The parametric prior is given by $\lambda_i \sim N\left(\mu_i, \omega_i^2\right)$, and a normal-inverse-gamma hyperprior is further imposed on $\left(\mu_i, \omega_i^2\right)$. It can be considered as a special case of the DPM prior when the scale parameter $\alpha \to \infty$, so there is only one component, and $\left(\mu_i, \omega_i^2\right)$ are directly drawn from the base distribution G_0. This choice of hyperprior follows the suggestion by Basu and Chib (2003) to match the Gaussian model with the DPM model such that "the predictive (or marginal) distribution of a single observation is identical under the two models" (pp. 226-227).

This paper focuses on the scenario in which f is continuous and approximated by a mixture model, as a continuous distribution may be more sensible for the skill of young firms as well as other similar empirical studies. To examine how much can be gained or lost from the continuity assumption, I also implement a DP prior where λ_i follows a flexible nonparametric distribution but on a discrete support.

These priors are denoted as "Homog", "Flat", "Param", and "NP-disc", respectively, in the graphs and tables below. In addition, "NP-R" denotes the proposed nonparametric prior for random effects/coefficients models, and "NP-C" for correlated random effects/coefficients models.

3.6.3 Baseline Model

Let us first consider the baseline model with random effects. The specifications are summarized in Table 12.

β_0 is set to be 0.8 as economic data usually exhibit some degree of persistence. σ_0^2 equals 1/4, so the rough magnitude of signal-noise ratio is $\sigma_0^2/\mathbb{V}(\lambda_i) = 1/4$. The initial conditions y_{i0} is

Table 12: Simulation Setup: Baseline Model

(a) Dynamic Panel Data Model

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it},\ u_{it} \sim N\left(0, \sigma^2\right)$
Common parameters	$\beta_0 = 0.8,\ \sigma_0^2 = 1$
Initial conditions	$y_{i0} \sim TN\left(0, 1, -5, 5\right)$
Sample size	$N = 1000,\ T = 6$

(b) Random Effects

Degenerate	$\lambda_i = 0$
Skewed	$\lambda_i \sim \frac{1}{9}N\left(2, \frac{1}{2}\right) + \frac{8}{9}N\left(-\frac{1}{4}, \frac{1}{2}\right)$
Fat tail	$\lambda_i \sim \frac{1}{5}N\left(0, 4\right) + \frac{4}{5}N\left(0, \frac{1}{4}\right)$
Bimodal	$\lambda_i \sim 0.35N\left(0, 1\right) + 0.65N\left(10, 1\right)$, normalized to $Var\left(\lambda_i\right) = 1$

drawn from a truncated normal distribution where I take the standard normal as the base distribution and truncate it at $|y_{i0}| < 5$. This truncation setup complies with Assumption 3.4.10 such that y_{i0} is compactly supported. Choices of N and T are comparable with the young firm dynamics application.

There are four parameterizations of the true distribution of λ_i, $f_0\left(\cdot\right)$. As this subsection focuses on the simplest baseline model with random effects, all the four parameterizations are independent of y_{i0}. The degenerate λ_i distribution suggests that all firms enjoy the same skill level. Note that it does not satisfy the first condition in Lemma 3.4.8, which requires the true λ_i distribution to be continuous. The purpose of this distribution is to learn how bad things can go under the misspecification that the true λ_i distribution is completely off the prior support. The functional forms of the skewed and fat tail distributions are borrowed from Monte Carlo design 2 in Liu *et al.* (2016). These two specifications reflect more realistic scenarios in empirical studies. The last setup portrays a bimodal distribution with asymmetric weights put on the two components.

I simulated 1,000 panel datasets for each setup and report the average statistics of these 1,000 exercises. Forecasting performance, especially the relative rankings and magnitudes, is highly stable across simulations. In each simulation exercise, I generated 40,000 MCMC

draws with the first 20,000 being discarded as burn-in. Based on graphical and statistical tests, the MCMC draws seem to converge to a stationary distribution. Both the Brook-Draper diagnostic and the Raftery-Lewis diagnostic yield desirable MCMC accuracy. For trace plots, prior/posterior distributions, rolling means, and autocorrelation graphs of β, σ^2, α, and λ_1, please refer to Figures 15 to 18.

Table 13 shows the forecasting comparison among alternative priors. The point forecasts are evaluated by MSE together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the LPS and the Amisano and Giacomini (2007) test. For the oracle predictor, the table reports the exact values of MSE and LPS (multiplied by the cross-sectional dimension N). For other predictors, the table reports the percentage deviations from the oracle MSE and difference with respect to the oracle LPS*N. The tests are conducted with respect to NP-R, with significance levels indicated by *: 10%, **: 5%, and ***: 1%. The entries in bold indicate the best feasible predictor in each column.

For each λ_i distribution, point forecasts and density forecasts share comparable rankings. When the λ_i distribution is degenerate, "Homog" and "NP-disc" are the best, as expected. They are followed by "NP-R" and "Param", and "Flat" is considerably worse. When the λ_i distribution is non-degenerate, there is a substantial gain in both point forecasts and density forecasts from employing the "NP-R" predictor. In the bimodal case, the "NP-R" predictor exceeds all other competitors. In principle, the nonparametric prior constructed from mixtures of normals should perform the best when the true DGP is made up of distinct normal components. In the skewed and fat tailed cases, the "Flat" and "Param" predictors are second best, yet still significantly inferior to "NP-R". The "Homog" and "NP-disc" predictors yield the poorest forecasts, which suggests that their discrete supports are not able to approximate the continuous λ_i distribution, and even the nonparametric DP prior with countably infinite support ("NP-disc") is far from enough.

Therefore, when researchers believe that the underlying λ_i distribution is indeed discrete, the DP prior ("NP-disc") is a more sensible choice; on the other hand, when the underlying

λ_i distribution is actually continuous, the DPM prior (or the MGLR_x prior later for the correlated random effects model) promotes better forecasts. In the empirical application to young firm dynamics, it would be more reasonable to assume continuous distributions of individual heterogeneities in levels, reactions to R&D, and shock sizes, and results show that the continuous nonparametric prior outperforms the discrete DP prior in terms of density forecasts (see Table 19).

To investigate the sources of the gain in forecasts, Figure 8 demonstrates the posterior distribution of the λ_i distribution (i.e. a distribution over distributions) for experiments "Skewed", "Fat Tail", and "Bimodal". In each case, the graphs are constructed from the estimation results of one simulation exercise among the 1,000 simulation exercises. The left subgraph is given by the "Param" estimator, which is compared and contrasted with the right subgraph by "NP-R". In each subgraph, the black solid line represents the true λ_i distribution, f_0. The blue bands show the posterior distribution of f, $\Pi\left(f \mid y_{1:N,0:T}\right)$.

For the skewed λ_i distribution, the "NP-R" estimator better tracks the peak on the left and the tail on the right. For the λ_i distribution with fat tails, the "NP-R" estimator accommodates the slowly decaying tails, but is still not able to fully mimic the spiking peak. For the bimodal λ_i distribution, it is not surprising that the "NP-R" estimator captures the M-shape fairly nicely. In summary, the nonparametric prior flexibly approximates a vast set of distributions, which helps provide more precise estimates of the underlying λ_i distributions and consequently more accurate density forecasts. This observation confirms the connection between skill distribution estimation and density forecasts as stated in Propositions 3.4.11 and 3.4.16.

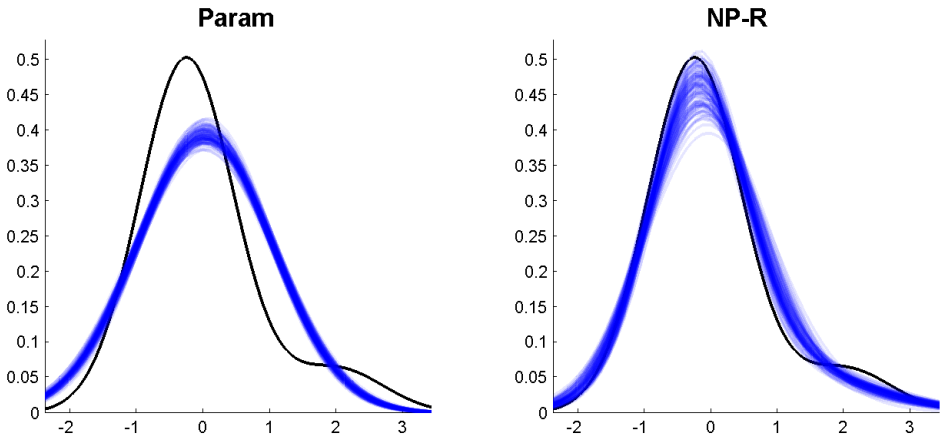
I have also considered various robustness checks. In terms of the setup, I have tried different cross-sectional dimensions $N = 100,\ 500,\ 1000,\ 10^5$, different time spans $T = 6,\ 10,\ 20,\ 50$, different persistences $\beta = 0.2,\ 0.5,\ 0.8,\ 0.95$, different sizes of the i.i.d. shocks $\sigma^2 = 1/4$ and 1, which govern the signal-to-noise ratio, and different underlying λ_i distributions including standard normal. In general, the "NP-R" predictor is the overall best for density forecasts

Table 13: Forecast Evaluation: Baseline Model

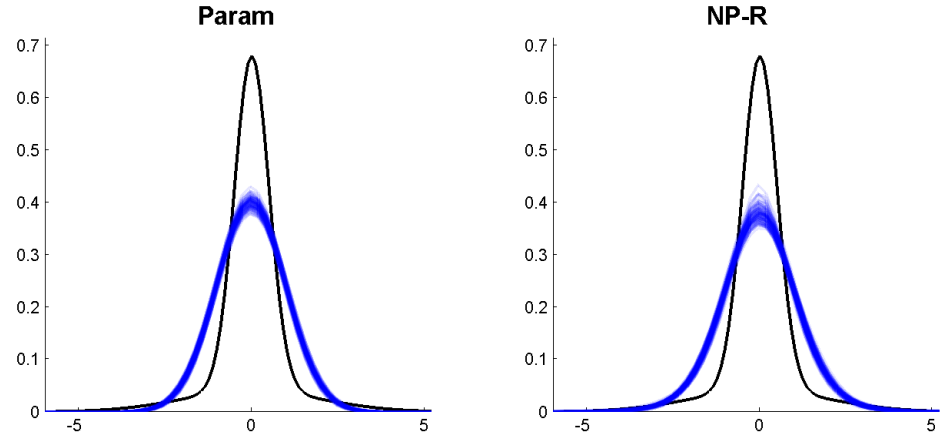
	Degenerate		Skewed		Fat Tail		Bimodal	
	MSE	LPS*N	MSE	LPS*N	MSE	LPS*N	MSE	LPS*N
Oracle	0.25	-725	0.29	-798	0.29	-804	0.27	-766
NP-R	0.8%	-4	**0.04%**	**-0.3**	**0.08%**	**-1**	**1.2%**	**-6**
Homog	**0.03%*****	**-0.2*****	32%***	-193***	29%***	-187***	126%***	-424***
Flat	21%***	-102***	1.4%***	-7***	0.3%***	-2***	8%***	-38***
Param	0.8%	-4	0.3%***	-1***	0.1%***	-1.5***	7%***	-34***
NP-disc	**0.03%*****	**-0.2*****	31%***	-206***	29%***	-205***	7%***	-40***

Figure 8: f_0 vs $\Pi(f \mid y_{1:N,0:T})$: Baseline Model

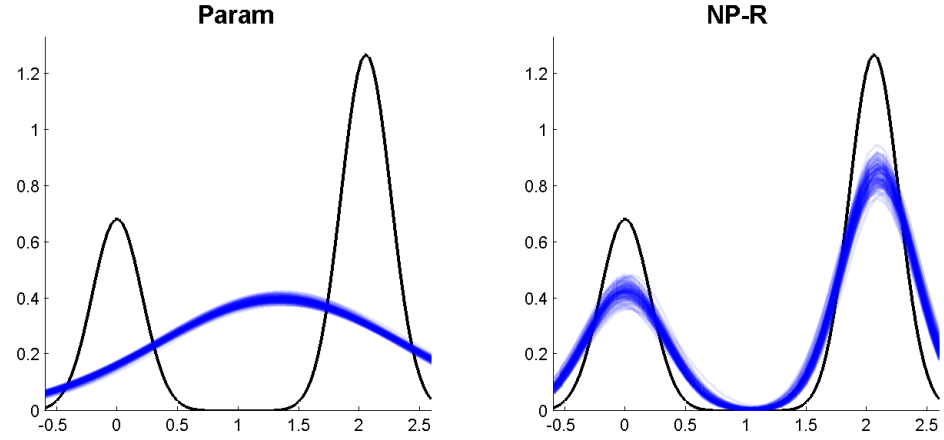
(a) Skewed



(b) Fat Tail



(c) Bimodal



except when the true λ_i comes from a degenerate distribution or a normal distribution. In the latter case, the parsimonious "Param" prior coincides with the underlying λ_i distribution and is not surprisingly but only marginally better than the "NP-R" predictor. Roughly speaking, the superiority of the "NP-R" predictor is more prominent when the time series for a specific "firm" i is not informative enough to reveal its "skill" but the whole panel can recover the skill distribution and hence "firm" i's "skill uncertainty". That is, "NP-R" works the best when N is not too small, T is not too long, σ^2 is not too large, and the λ_i distribution is relatively non-Gaussian. For instance, as the cross-sectional dimension N increases, the blue band in Figure 8 gets closer to the true f_0 and eventually completely overlaps it (see Figure 19), which resonates the posterior consistency statement.

In terms of estimators, I have also constructed the posterior sampler for more sophisticated priors, such as the Pitman-Yor process which allows power law tail for clustering behaviors, as well as DPM with skew normal components which better accommodates asymmetric data generating process. They provide some improvement in the corresponding situations, but call for extra computation efforts.

3.6.4 General Model

The general model accounts for three key features: (i) multidimensional individual heterogeneity, (ii) cross-sectional heteroskedasticity, and (iii) correlated random coefficients. The exact specification is characterized in Table 14.

In terms of multidimensional individual heterogeneity, now λ_i is a 3-by-1 vector, and the corresponding covariates are composed of the level, time-specific $w_{t-1}^{(2)}$, and individual-time-specific $w_{i,t-1}^{(3)}$.

In terms of correlated random coefficients, I adopt the conditional distribution following Dunson and Park (2008) and Norets and Pelenis (2014). They regard it as a challenging problem because such conditional distribution exhibits rapid changes in its shape which considerably restricts local sample size. The original conditional distribution in their papers

Table 14: Simulation Setup: General Model

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it},\ u_{it} \sim N\left(0, \sigma_i^2\right)$
Covariates	$w_{i,t-1} = [1,\ w_{t-1}^{(2)},\ w_{i,t-1}^{(3)}]'$, where $w_{t-1}^{(2)} \sim N(0,1)$ and $w_{i,t-1}^{(3)} \sim \text{Ga}(1,1)$
Common parameters	$\beta_0 = 0.8$
Initial conditions	$y_{i0} \sim U(0,1)$
Correlated random coef.	$\lambda_i \vert y_{i0} \sim e^{-2y_{i0}} N\left(y_{i0} v,\ 0.1^2 vv'\right) + \left(1 - e^{-2y_{i0}}\right) N\left(y_{i0}^4 v,\ 0.2^2 vv'\right)$, where $v = [1,\ 2,\ -1]'$
Cross-sectional heterosk.	$\sigma_i^2 \vert y_{i0} \sim 0.454\,(y_{i0} + 0.5)^2 \cdot (\text{IG}(51, 40) + 0.2)$
Sample size	$N = 1000,\ T = 6$

is one-dimensional, and I expand it to accommodate the three-dimensional λ_i via a linear transformation of the original. In Figure 9 panel (a), the left subgraph presents the joint distribution of λ_{i1} and y_{i0}, where λ_{i1} is the coefficient on $w_{i,t-1}^{(1)} = 1$ and can be interpreted as the heterogeneous intercept. It shows that the shape of the joint distribution is fairly complex, containing many local peaks and valleys. The right subgraph shows the conditional distribution of λ_{i1} given $y_{i0} = 0.25,\ 0.5, 0.75$. We can see that the conditional distribution is also irregular and evolves with y_{i0}.

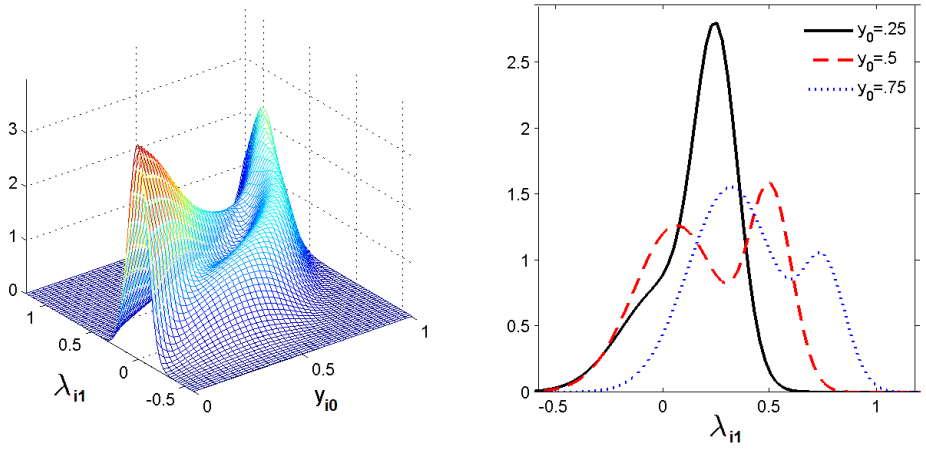
In addition, I also let the cross-sectional heteroskedasticity interact with the initial conditions, and the functional form is modified from Pelenis (2014) case 2. The modification guarantees the continuity of σ_i^2 distribution, bounds it above zero (see conditions for Propositions 3.5.10-3.5.12), and ensures that the signal-to-noise ratio is not far from 1. Their joint and conditional distributions are depicted in Figure 9 panel (b).

The rest of the setup is the same as the baseline scenario in the previous subsection.

Due to cross-sectional heteroskedasticity and correlated random coefficients, the prior structures become more complicated. Table 15 describes the prior setups of λ_i and l_i, with the predictor labels being consistent with the definitions in Subsection 3.6.2. Note that I further add the "Homosk-NP-C" predictor in order to examine whether it is practically relevant to model heteroskedasticity.

Figure 9: DGP: General Model

(a) $p\left(\lambda_{i1}|y_{i0}\right)$



(b) $p\left(\sigma_i^2|y_{i0}\right)$

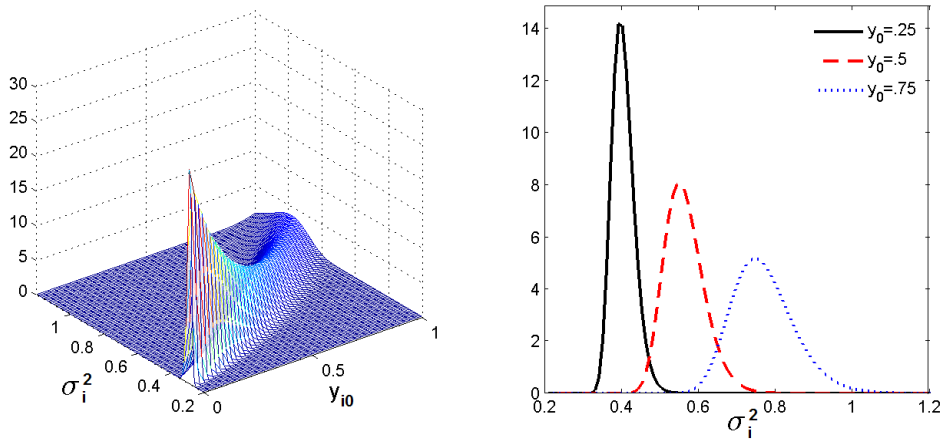


Table 15: Prior Structures

Predictor		λ_i prior	l_i prior
Heterosk	NP-C	MGLR_x	MGLR_x
Homog		Point mass	Point mass
Homosk	NP-C	MGLR_x	Point mass
Heterosk	Flat	Uninformative	Uninformative
	Param	N	IG
	NP-disc	DP	DP
	NP-R	DPM	DPM

Table 16 assesses the forecasting performance of these predictors. From the best to the worst, the point forecast ranking is "Heterosk-NP-R", "Heterosk-Param", "Heterosk-NP-disc", "Heterosk-NP-C", "Homosk-NP-C", "Homog", and "Heterosk-Flat". The first two constitute the first tier, the next two can be viewed as the second tier, the next one is the third tier, and the last two are markedly inferior. It is anticipated that more parsimonious estimators would outperform "Heterosk-NP-C" in terms of point forecasts, though "Heterosk-NP-C" is correctly specified while the parsimonious ones are not.

Nevertheless, the focus of this paper is density forecasting, where "Heterosk-NP-C" becomes the most accurate density predictor. Several lessons can be inferred from a more detailed comparison among predictors. First, based on the comparison between "Heterosk-NP-C" and "Homog"/"Homosk-NP-C", it is important to account for individual effects in both coefficients λ_i's and shock sizes σ_i^2's. Second, comparing "Heterosk-NP-C" with "Heterosk-Flat"/"Heterosk-Param", we see that the flexible nonparametric prior plays a significant role in enhancing density forecasts. Third, the difference between "Heterosk-NP-C" and "Heterosk-NP-disc" indicates that the discrete prior performs less satisfactorily when the underlying individual heterogeneity is continuous. Last, "Heterosk-NP-R" is less favorable than "Heterosk-NP-C", which necessitates a careful modeling of the correlated random coefficient structure.

3.7 Empirical Application: Young Firm Dynamics

3.7.1 Background and Data

To understand how the proposed predictor works in real world analysis, I applied it to provide density forecasts of young firm performance. Studies have documented that young firm performance is affected by R&D, recession, etc. and that different firms may react differently to these factors (Akcigit and Kerr, 2010; Robb and Seamans, 2014; Zarutskie and Yang, 2015). In this empirical application, I examine these channels from a density forecasting perspective.

Table 16: Forecast Evaluation: General Model

		MSE	LPS*N
Oracle		0.70	-1150
Heterosk	NP-C	13.68%	**-74**
Homog		89.28%***	-503***
Homosk	NP-C	20.84%***	-161***
Heterosk	Flat	151.60%***	-515***
	Param	11.30%*	-139***
	NP-disc	13.08%	-150***
	NP-R	**11.25%***	-93*

The point forecasts are evaluated by the Mean Square Error (MSE) together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the log predictive score (LPS) and the Amisano and Giacomini (2007) test. For the oracle predictor, the table reports the exact values of MSE and LPS. For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS. The tests are conducted with respect to Heterosk-NP-C, with significance levels indicated by *: 10%, **: 5%, ***: 1%. The entries in bold indicate the best feasible predictor in each column.

To analyze firm dynamics, traditional cross-sectional data are not sufficient whereas panel data are more suitable as they track the firms over time. In particular, it is desirable to work with a dataset that contains sufficient information on early firm financing[31] and innovation, and spreads over the recent recession. The restricted-access Kauffman Firm Survey (KFS) is the ideal candidate for such purpose, as it offers the largest panel of startups (4,928 firms founded in 2004, nationally representative sample) and longest time span (2004-2011, one baseline survey and seven follow-up annual surveys), together with detailed information on young firms. For further description of the survey design, please refer to Robb *et al.* (2009).[32]

3.7.2 Model Specification

I consider the general model with multidimensional individual heterogeneity in λ_i and cross-sectional heteroskedasticity in σ_i^2. Following the firm dynamics literature, such as Akcigit and Kerr (2010) and Zarutskie and Yang (2015), firm performance is measured by employ-

[31] In the current version of the empirical exercises, firm financing variables (e.g. capital structure) are not included as regressors because they overly restrict the cross-sectional dimension, but I intend to include them in future work in which I will explicitly model firm exit and thus allow for a larger cross-section.

[32] Here I do not impose weights on firms as the purpose of the current study is forecasting individual firm performance. Further extensions can easily incorporate weights into the estimation procedure.

ment. Specifically, here y_{it} is chosen to be the log of employment denoted as $\log \text{emp}_{it}$. I adopt the log of employment instead of employment growth rate since the latter significantly reduces the cross-sectional sample size. It is preferable to work with larger N according to the theoretical argument.

For the key variables with potential heterogeneous effects ($w_{i,t-1}$), I compare the forecasting performance of the following three setups:[33]

(i) $w_{i,t-1} = 1$, which specifies the baseline model with λ_i being the individual-specific intercept.

(ii) $w_{i,t-1} = [1,\ \text{rec}_{t-1}]'$. rec_t is an aggregate dummy variable indicating the recent recession. It is equal to 1 for 2008 and 2009, and is equal to 0 for other periods.

(iii) $w_{i,t-1} = [1,\ \text{R\&D}_{i,t-1}]'$. $\text{R\&D}_{it}$ is given by the ratio of a firm's R&D employment over its total employment, considering that R&D employment has more complete observations compared to other innovation intensity gauges.[34]

The panel used for estimation spans 2004 to 2010 with time-series dimension $T = 6$.[35] The data for 2011 is reserved for pseudo out-of-sample forecast evaluation. Sample selection is performed as follows:

(i) For any (i, t) combination where R&D employment is greater than the total employment, there is an incompatibility issue, so I set $\text{R\&D}_{it} = NA$, which only affects 0.68% of the observations.

(ii) I only keep firms with long enough observations according to Assumption 3.5.7, which ensures identification in unbalanced panels. This results in cross-sectional dimension $N =$

[33] I do not jointly incorporate recession and R&D because such specification largely restricts the cross-sectional sample size.

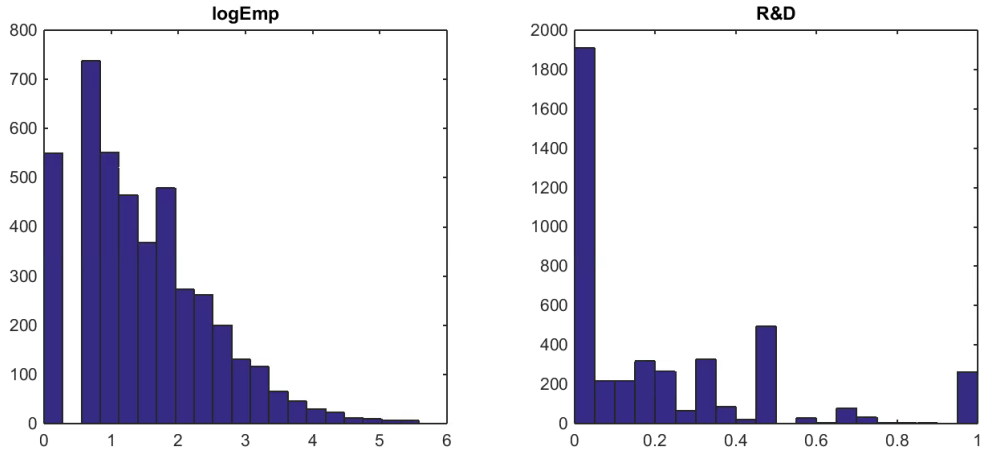
[34] I have also explored other measures of innovation activities (e.g. a binary variable on whether the firm spends any money on R&D, numbers of intellectual properties–patents, copyrights, or trademarks–owned or licensed by the firm). The estimated AR(1) coefficients and relative rankings of density forecasts are generally robust across measures.

[35] Note that the estimation panel starts from period 0 (i.e. 2004) and ends at period T (i.e. 2010) with $T + 1 = 7$ periods in total.

Table 17: Descriptive Statistics for Observable

	10%	mean	med	90%	std	skew	kurt
log emp	0.41	1.44	1.34	2.63	0.86	0.82	3.58
R&D	0.05	0.22	0.17	0.49	0.18	1.21	4.25

Figure 10: Histograms for Observables



859 for the baseline specification, $N = 794$ with recession, and $N = 677$ with R&D.

(iii) In order to compare forecasting performance across different setups, the sample is further restricted so that all three setups share exactly the same set of firms.

After all these data cleaning steps, we are left with $N = 654$ firms. The proportion of missing values are $(\#\text{missing obs}) / (NT) = 6.27\%$. The descriptive statistics for $\log \text{emp}_{it}$ and $\text{R\&D}_{it}$ are summarized in Table 17, and the corresponding histograms are plotted in Figure 10, where both distributions are right skewed and may have more than one peak.

3.7.3 Results

The alternative priors are similar to those in the Monte Carlo simulation except for one additional prior, "Heterosk-NP-C/R", which assumes that λ_i is correlated with y_{i0} while σ_i^2 is not, by imposing an MGLR_x prior on λ_i and a DPM prior on $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$. It is possible to craft other priors according to the specific heterogeneity structure of the empirical problem at hand. For example, let λ_{i1} correlate with y_{i0} while setting λ_{i2} independent of y_{i0}.

Table 18: Common Parameter β

		Baseline		Recession		R&D	
		mean	std	mean	std	mean	std
Heterosk	NP-C/R	0.48	0.01	0.46	0.02	0.52	0.01
Homog		0.85	0.02	0.85	0.02	0.89	0.02
Homosk	NP-C	0.37	0.02	0.88	0.02	0.51	0.03
Heterosk	Flat	0.19	0.02	0.25	0.00	0.50	0.00
	Param	0.48	0.03	0.26	0.03	0.56	0.03
	NP-disc	0.55	0.02	0.79	0.02	0.84	0.04
	NP-R	0.47	0.03	0.30	0.03	0.74	0.04
	NP-C	0.38	0.02	0.40	0.06	0.53	0.01

I will leave this to future exploration. The conditioning set is chosen to be standardized y_{i0}. The standardization ensures numerical stability in practice, as the conditioning variables enter exponentially into the covariance function for the Gaussian process.

Table 18 characterizes the posterior estimates of the common parameter β. In most of the cases except for "Homog" and "NP-disc", the posterior means are around $0.4 \sim 0.5$, which suggests that the young firm performance exhibits some degree of persistency, but not remarkably strong, which is reasonable as young firms generally experience more uncertainty. For "Homog" and "NP-disc", their posterior means of β are much larger. This may arise from the fact that homogeneous or discrete λ_i structure is not able to capture all individual effects, so these estimators may attribute the remaining individual effects to persistence and thus overestimate β. In all scenarios, the posterior standard deviations are relatively small, which indicates that the posterior distributions are very tight.

Table 19 compares the forecasting performance of the predictors across different model setups. The "Heterosk-NP-C/R" predictor is chosen to be the benchmark for all comparisons. For the benchmark predictor, the table reports the exact values of MSE and LPS (multiplied by the cross-sectional dimension N). For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS*N.

For density forecasts measured by LPS, the overall best is the "Heterosk-NP-C/R" predictor

in the R&D setup. Comparing setups, the one with recession yields the worst density forecasts (and point forecasts as well), so the recession dummy does not contribute much to forecasting and may even incur overfitting.

Comparing across predictors for the baseline and R&D setups, the main message is similar to the Monte Carlo simulation of the general model in Subsection 3.6.4. In summary, it is crucial to account for individual effects in both coefficients λ_i's and shock sizes σ_i^2's through a flexible nonparametric prior that acknowledges continuity and correlated random effects/coefficients when the underlying individual heterogeneity is likely to possess these features. Note that now both "NP-R" and "NP-C" are inferior to "NP-C/R" where the distribution of λ_i depends on the initial conditions but the distribution of σ_i^2 does not.[36]

In terms of point forecasts, most of the estimators are comparable according to MSE, with only "Flat" performing poorly in all three setups. Intuitively, shrinkage in general leads to better forecasting performance, especially for point forecasts, whereas the "Flat" prior does not introduce any shrinkage to individual effects $\left(\lambda_i, \sigma_i^2\right)$. Conditional on the common parameter β, the "Flat" estimator of $\left(\lambda_i, \sigma_i^2\right)$ is a Bayesian analog of individual-specific MLE/OLS that utilizes only the individual-specific observations, which is inadmissible under fixed T (Robbins, 1956; James and Stein, 1961; Efron, 2012).

Figure 11 provides the histograms of the probability integral transformation (PIT) in the R&D setup. While LPS characterizes the relative ranks of predictors, PIT supplements LPS and can be viewed as an absolute evaluation on how good the density forecasts coincide with the true (unobserved) conditional forecasting distributions with respect to the current information set. In this sense, under the null hypothesis that the density forecasts coincide with the truth, the probability integral transforms are i.i.d. $U(0,1)$ and the histogram is close to a flat line. For details of PIT, please refer to Diebold *et al.* (1998). In each subgraph, the two red lines indicate the confidence interval. We can see that, in "NP-C/R",

[36] This result cannot be directly compared to the Gibrat's law literature (Lee *et al.*, 1998; Santarelli *et al.*, 2006), as the dependent variable here is the log of employment instead of employment growth.

Table 19: Forecast Evaluation: Young Firm Dynamics

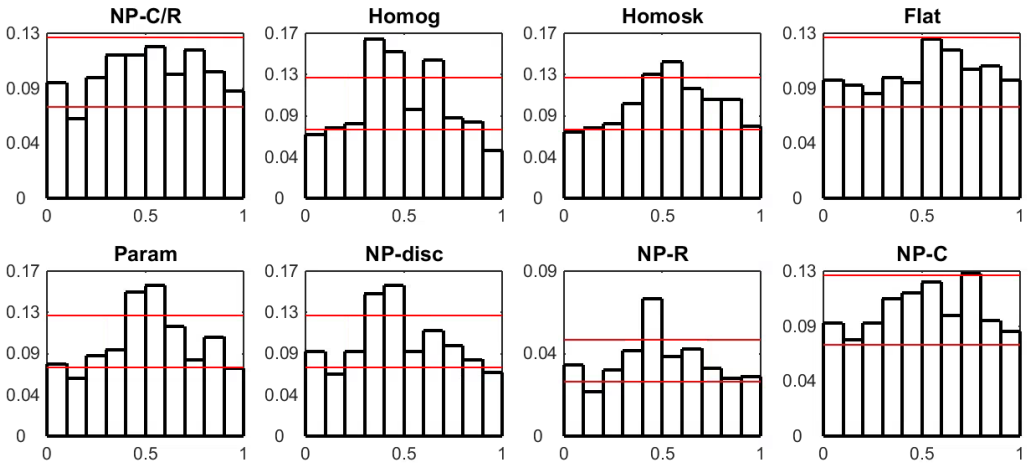
		Baseline		Recession		R&D	
		MSE	LPS*N	MSE	LPS*N	MSE	LPS*N
Heterosk	NP-C/R	**0.20**	**-230**	0.23	**-272**	**0.20**	**-228**
Homog		10%**	-81***	-2%	-41***	8%*	-74***
Homosk	NP-C	7%**	-66***	2%	-17**	9%	-52***
Heterosk	Flat	22%***	-42***	44%***	-701***	102%***	-309***
	Param	4%*	-60***	35%***	-135***	7%	-52***
	NP-disc	1%	-9**	**-7%**	-1	2%	-20***
	NP-R	1%	-5*	28%***	-63***	3%	-16***
	NP-C	3%*	-6*	3%	-5**	0.1%	-5**

The point forecasts are evaluated by the Mean Square Error (MSE) together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the log predictive score (LPS) and the Amisano and Giacomini (2007) test. For the benchmark predictor Heterosk-NP-C/R, the table reports the exact values of MSE and LPS. For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS. The tests are conducted with respect to the benchmark, with significance levels indicated by *: 10%, **: 5%, ***: 1%. The entries in bold indicate the best predictor in each column.

"NP-C" and "Flat", the histogram bars are mostly within the confidence band, while other predictors yield apparent inverse-U shapes. The reason might be that the other predictors do not take correlated random coefficients into account but instead attributes the subtlety of correlated random coefficients to the estimated variance, which leads to more diffused predictive distributions.

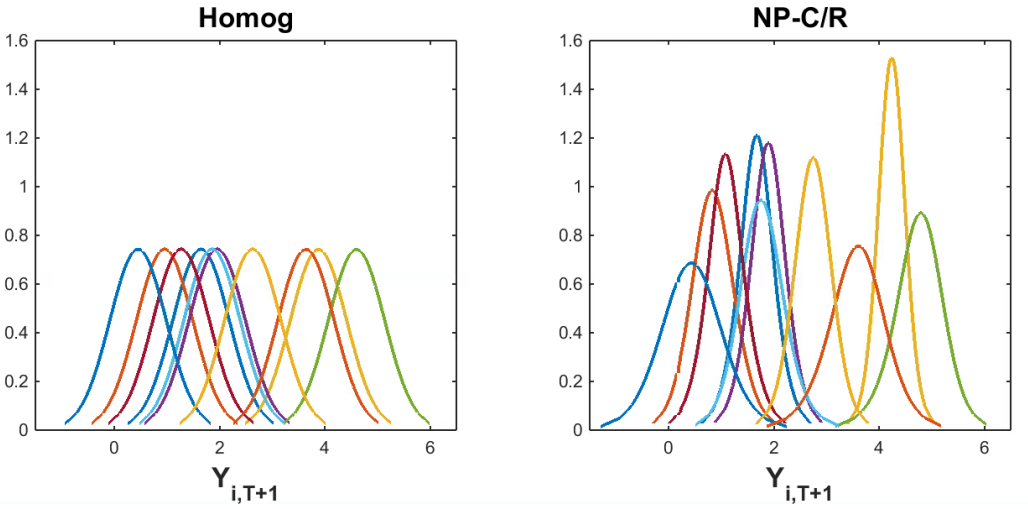
Figure 12 shows the predictive distributions of 10 randomly selected firms in the R&D setup. In terms of the "Homog" predictor, all predictive distributions share the same Gaussian shape paralleling with each other. On the contrary, in terms of the "NP-C/R" predictor, it is clear that the predictive distributions are fairly different in the center location, variance, and skewness. Figure 13 further aggregates the predictive distributions over sectors based on two-digit NAICS codes (Table 20). It plots the predictive distributions of the log of the average employment within each sector. Comparing "Homog" and "NP-C/R" across sectors, we can see the following several patterns. First, "NP-C/R" predictive distributions tend to be narrower and have longer right tails, whereas "Homog" ones are distributed in the standard bell shape. Second, there are substantial heterogeneities in density forecasts

Figure 11: PIT



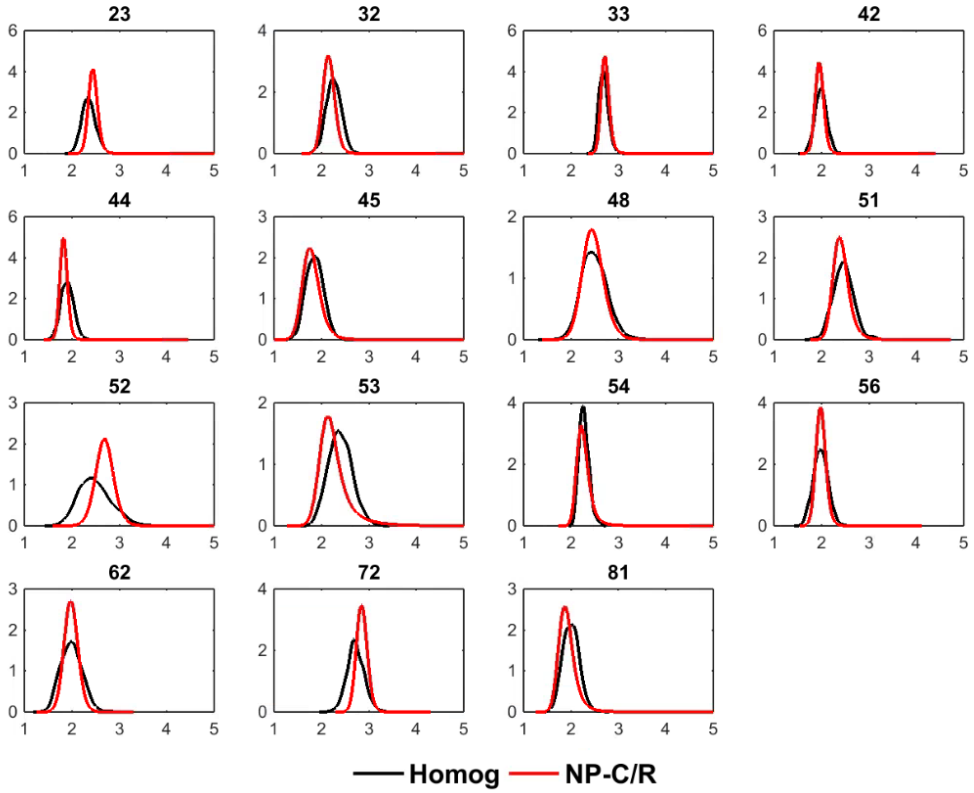
Red lines indicate the confidence interval.

Figure 12: Predictive Distributions: 10 Randomly Selected Firms



across sectors. For sectors with relatively large average employment, e.g. "construction" (sector 23), "Homog" pushes the forecasts down, hence systematically underpredicts their future employment, while "NP-C/R" respects this source of heterogeneity and significantly lessens the underprediction problem. On the other hand, for sectors with relatively small average employment, e.g. "Retail Trade" (sector 44), "Homog" introduces an upward bias into the forecasts, while "NP-C/R" reduces such bias by flexibly estimating the underlying distribution of firm-specific heterogeneities.

Figure 13: Predictive Distributions: Aggregated by Sectors



Subgraph titles are two-digit NAICS codes. Only sectors with more than 10 firms are shown.

Table 20: Two-digit NAICS Codes

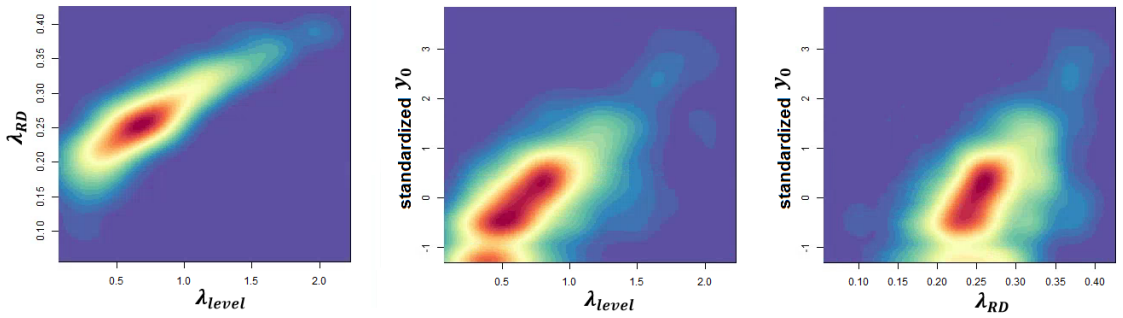
Code	Sector
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)

The latent heterogeneity structure is presented in Figure 14, which plots the joint distributions of the estimated individual effects and the conditional variable in the R&D setup. We can see that $\lambda_{i,\text{level}}$, $\lambda_{i,\text{RD}}$, and standardized y_{i0} are positively correlated with each other, which roughly indicates that larger firms respond more positively to R&D activities within the KFS young firm sample. In all the three subgraphs, the pairwise relationships among $\lambda_{i,\text{level}}$, $\lambda_{i,\text{RD}}$, and standardized y_{i0} are nonlinear and exhibit multiple components, which reassures the utilization of nonparametric prior with correlated random coefficients.

3.8 Concluding Remarks

This paper proposes a semiparametric Bayesian predictor which performs well in density forecasts of individuals in a panel data setup. Monte Carlo simulations and an empirical application to young firms dynamics show that the keys for better density forecasts are, in order of importance, nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients.

Figure 14: Joint Distributions of $\hat{\lambda}_i$ and Condition Variable



Moving forward, I plan to extend my research in the following several directions: Theoretically, I will continue the Bayesian asymptotic discussion with strong posterior consistency and rates of convergence. Methodologically, I will explore some variations of the current setup. First, some empirical studies may include a large number of covariates with potential heterogeneous effects (i.e. more variables included in $w_{i,t-1}$), so it is both theoretically and empirically desirable to investigate a variable selection scheme in a high-dimensional nonparametric Bayesian framework. Chung and Dunson (2012) and Liverani *et al.* (2015) employ variable selection via binary switches, which may be adaptable to the panel data setting. Another possible direction is to construct a Bayesian-Lasso-type estimator coherent with the current nonparametric Bayesian implementation. Second, I will consider panel VAR (Canova and Ciccarelli, 2013), a useful tool to incorporate several variables for each of the individuals and to jointly model the evolution of these variables, allowing me to take more information into account for forecasting purposes and offer richer insights into the latent heterogeneity structure. Meanwhile, it is also interesting to incorporate extra cross-variable restrictions and implement the Bayesian GMM method as proposed in Shin (2014). Third, I will experiment with nonlinear panel data models, such as the Tobit model that helps accommodate firms' endogenous exit choice. Such extension would be numerically feasible, but requires further theoretical work. A natural next step would be extending the theoretical discussion to the family of "generalized linear models".

APPENDIX A

Point Forecasts and Bank Stress Tests

A.1 Theoretical Derivations and Proofs

A.1.1 Proofs for Section 2.2

Lemma A.1.1. *Suppose that $T \geq k_w + 1 \geq 2$. Suppose that W is a $T \times k_w$ matrix with* $\mathrm{rank(W) = k_w}$. *Let Σ be a $T \times T$ matrix of rank T. Let $S = \Sigma W$. Then,* $\mathrm{rank(M_{S\otimes S}B) = T}$, *where $M_{S\otimes S}$ and B are defined in the proof of Theorem 2.2.3.*

Proof of Lemma A.1.1. 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, $\mathrm{rank}(S) = \mathrm{rank}(\Sigma W) = \mathrm{rank}(W) = k_w$. If $\mathrm{rank}(S) = k_w$, then $\mathrm{rank}(S \otimes S) = k_w^2$. Since the rank of the projection matrix is the same as its trace, we have $\mathrm{rank}(M_{S\otimes S}) = tr(M_{S\otimes S}) = T^2 - k_w^2$.

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, $\mathrm{rank}(M_{S\otimes S}B) = \mathrm{rank}(F\Lambda F'B) = \mathrm{rank}(\Lambda F'B)$. 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, $\mathrm{rank}(F'B) = T$. 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. □

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 2.2.3. (i) The parameters α and ρ are identifiable by Assumption 2.2.2.

(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

$$M_i(\tilde{\gamma})\big(\Sigma_i^{-1/2}(\tilde{\gamma})\Sigma_i(\gamma)\Sigma_i^{-1/2}(\tilde{\gamma}) - I\big)M_i'(\tilde{\gamma}) = 0. \tag{A.1.1}$$

for each h_i. Taking expectations with respect to H_i and using Assumption 2.2.2(ii), we deduce that

$$\mathbb{E}\big[M_i(\tilde{\gamma})\big(\Sigma_i^{-1/2}(\tilde{\gamma})\Sigma_i(\gamma)\Sigma_i^{-1/2}(\tilde{\gamma}) - I\big)M_i'(\tilde{\gamma})\big] = 0. \tag{A.1.2}$$

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

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

$$\tilde{Y}_i = \Sigma_i^{-1/2}(\gamma)(Y_i - X_i\rho - Z_i\alpha) = S_i(\gamma)\lambda_i + V_i. \tag{A.1.3}$$

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

$$\ln \Psi_{\tilde{Y}}(\tau|h) = \ln \Psi_{\lambda}(S'\tau|h) + \ln \Psi_V(\tau) \tag{A.1.4}$$

Taking the second derivative with respect to τ leads to

$$\begin{aligned}\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_{\tilde{Y}}(\tau|h) &= \frac{\partial^2}{\partial\tau\partial\tau'}\left(\ln\Psi_\lambda(S'\tau|h)\right)+\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau) \\ &= S\left(\frac{\partial^2}{\partial\xi\partial\xi'}\ln\Psi_\lambda(S'\tau|h)\right)S'+\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau).\end{aligned} \tag{A.1.5}$$

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,

$$\begin{aligned}\text{vec}\left(\frac{\partial^2}{\partial\tau\partial\tau'}\ln\Psi_V(\tau)\right) &= \text{vec}\left(\text{diag}\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)\right) \\ &= B\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)'\end{aligned} \tag{A.1.6}$$

for a suitably chosen matrix B. Let

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

Then,

$$M_{S\otimes S}\text{vec}(\ln\Psi_{\tilde{Y}}(\tau|h))=M_{S\otimes S}B\left(\frac{\partial^2}{\partial\tau_1^2}\ln\Psi_{V_1}(\tau_1),...,\frac{\partial^2}{\partial\tau_T^2}\ln\Psi_{V_T}(\tau_T)\right)'. \tag{A.1.7}$$

Because $\Sigma(\gamma)$ is of full rank T (Assumption 2.2.2(iii)) and W is of full rank of k_w (Assumption 2.2.2(iv)), $S(\gamma)$ has full rank k_w. Notice that $T\geq k_w+1$. Then, according to Lemma A.1.1, $M_{S\otimes S}B$ is also full rank. In turn, from $(A.1.7)$, 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.1.4) we obtain

$$\ln \Psi_{\tilde{Y}}(\tau|h) - \ln \Psi_V(\tau) = \ln \Psi_\lambda(S'\tau|h). \tag{A.1.8}$$

Taking second derivatives, we obtain

$$\frac{\partial^2}{\partial\tau\partial\tau'}\left(\ln \Psi_{\tilde{Y}}(\tau|h) - \sum_{t=1}^{T} \ln \Psi_V(\tau_t)\right) = S\left(\frac{\partial^2}{\partial\xi\partial\xi'} \ln \Psi_\lambda(S'\tau|h)\right) S'. \tag{A.1.9}$$

Because S is of full rank, we can identify

$$\frac{\partial^2}{\partial\xi\partial\xi'} \ln \Psi_\lambda(S'\tau|h) = (S'S)^{-1}S'\left[\frac{\partial^2}{\partial\tau\partial\tau'}\left(\ln \Psi_{\tilde{Y}}(\tau|h) - \sum_{t=1}^{T} \ln \Psi_V(\tau_t)\right)\right] S(S'S)^{-1}. \tag{A.1.10}$$

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

$$\hat{\lambda} = (S'S)^{-1}S'\tilde{Y} = \lambda + (S'S)^{-1}S'V. \tag{A.1.11}$$

Taking expectations yields

$$\mathbb{E}(\lambda|h) = \mathbb{E}[\hat{\lambda}|h], \tag{A.1.12}$$

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 2.2.2(i). We discuss an example of how to identify α and ρ based on moment conditions in the general model (2.1.1). Under the model (2.1.1) we can

remove the effect of λ_i with the following within projections:

$$\begin{aligned} Y^*_{it} &= Y_{it} - \left(\sum_{s=t+1}^{T} Y_{is} W'_{is-1}\right)\left(\sum_{s=t+1}^{T} W_{is-1} W'_{is-1}\right)^{-1} W_{it-1} \\ X^*_{it-1} &= X_{it-1} - \left(\sum_{s=t+1}^{T} X_{is-1} W'_{is-1}\right)\left(\sum_{s=t+1}^{T} W_{is-1} W'_{is-1}\right)^{-1} W_{it-1} \\ Z^*_{it-1} &= Z_{it-1} - \left(\sum_{s=t+1}^{T} Z_{is-1} W'_{is-1}\right)\left(\sum_{s=t+1}^{T} W_{is-1} W'_{is-1}\right)^{-1} W_{it-1} \end{aligned}$$

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

$$\mathbb{E}\left[\left(Y^*_{it} - \begin{bmatrix} \tilde{\rho}' & \tilde{\alpha}' \end{bmatrix} \begin{bmatrix} X^*_{it-1} \\ Z^*_{it-1} \end{bmatrix}\right) \begin{bmatrix} X'_{it-s-1} & Z'_{it-s-1} \end{bmatrix}\right] = 0 \tag{A.1.13}$$

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

$$\mathbb{E}\begin{bmatrix} X^*_{it-1}X'_{it-1} & X^*_{it-1}Z'_{it-1} \\ Z^*_{it-1}X'_{it-1} & Z^*_{it-1}Z'_{it-1} \end{bmatrix} \tag{A.1.14}$$

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

$$
\begin{aligned}
&\mathbb{E}[X^*_{it-1}X'_{it-1}] \\
=&\mathbb{E}\left[\left(X_{it-1}-\left(\sum_{s=t+1}^{T}X_{is-1}W'_{is-1}\right)\left(\sum_{s=t+1}^{T}W_{is-1}W'_{is-1}\right)^{-1}W_{it-1}\right)X'_{it-1}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(X_{it-1}-\left(\sum_{s=t+1}^{T}X_{is-1}W'_{is-1}\right)\left(\sum_{s=t+1}^{T}W_{is-1}W'_{is-1}\right)^{-1}W_{it-1}\right)X'_{it-1}\,\middle|\,W_i^{t:T-1}\right]\right] \\
=&\mathbb{E}[X_{it-1}X'_{it-1}]-\frac{1}{T-h}\Bigg(\sum_{s=t+1}^{T}\mathbb{E}\Big[\mathbb{E}[X_{is-1}X_{it-1}|W_i^{t:T-1}] \\
&\times W'_{is-1}\left(\frac{1}{T-h}\sum_{s=t+1}^{T}W_{is-1}W'_{is-1}\right)^{-1}W_{it-1}\Big]\Bigg) \\
=&\mathbb{E}[X_{it-1}X'_{it-1}]-\frac{1}{T-h}\sum_{s=t+1}^{T}\kappa_s\mathbb{E}[X_{is-1}X'_{it-1}]=I+II,\text{ say.}
\end{aligned}
$$

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.1.14).

A.1.2 Proofs for Section 2.5

Sufficient Conditions for Assumption 2.5.3(iii)

The high-level condition in Assumption 2.5.3(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 $|a| \le \kappa_N \to 0$,

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

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 $|y| \le \mathcal{Y}_\lambda^\pi \cap [-C'_N, C_N]$ and $\lambda \in \Lambda^\pi$, by the change-of-variable with $y^* = \frac{\tilde{y}-y}{B_N}$, we have

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

Split the integration into two, one over $|y^*| \le D_N$ and other one over $|y^*| > D_N$. By Assumption 2.5.3(i) and (iii)-(a), uniformly in $|y^*| \le D_N$ and other one over $|y^*| > D_N$,

$$\begin{aligned}\left|\int_{|y^*|>D_N}\phi(y^*)\left(\frac{\pi(y+B_Ny^*|\lambda)-\pi(y|\lambda)}{\pi(y|\lambda)}\right)dy^*\right| &\le \frac{M\int_{|y^*|>D_N}\phi(y^*)dy^*}{\inf_{y\in\mathcal{Y}_\lambda^\pi\cap[-C'_N,C_N],\lambda\in\Lambda^\pi}\pi(y|\lambda)}\\ &\le \frac{M\exp\left(-\frac{D_N^2}{2}\right)}{\inf_{y\in\mathcal{Y}_\lambda^\pi\cap[-C'_N,C_N],\lambda\in\Lambda^\pi}\pi(y|\lambda)}\\ &= o(1)\end{aligned}$$

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

$$\begin{aligned}\left|\int_{|y^*|\le D_N}\phi(y^*)\left(\frac{\pi(y+B_Ny^*|\lambda)-\pi(y|\lambda)}{\pi(y|\lambda)}\right)dy^*\right| &\le \int\phi(y^*)y^*dy^*\left|\frac{\delta(y,\lambda)}{\pi(y|\lambda)}B_N\right|\\ &= Mo(1) = o(1)\end{aligned}$$

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

An Example of a $\pi(y|\lambda)$ That Satisfies Assumption 2.5.3

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 2.5.3(i) is satisfied. To verify Assumption 2.5.3(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 2.5.3(ii), regardless of the specific rate of C_N. To verify Assumption 2.5.3(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

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

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 2.5.2.

Main Theorem

Proof of Theorem 2.5.5. The goal is to prove that for a given $\epsilon_0 > 0$

$$\limsup_{N\to\infty} \frac{R_N(\widehat{Y}^N_{T+1}) - R^{\text{opt}}_N}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_\theta\left[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}^i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 0, \tag{A.1.15}$$

where

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

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

$$\limsup_{N\to\infty} \frac{N\mathbb{E}^{\mathcal{Y}^N,\lambda_i}_\theta\left[\left(\lambda_i + \rho Y_{iT} - \widehat{Y}_{iT+1}\right)^2\right]}{N\mathbb{E}^{\mathcal{Y}^i,\lambda_i}_\theta\left[(\lambda_i - \mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}^i}[\lambda_i])^2\right] + N^{\epsilon_0}} \leq 1. \tag{A.1.16}$$

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:

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

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

$$p_*(\hat{\lambda}_i, y_{i0}) = \mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})]. \tag{A.1.17}$$

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

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

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

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

and change the order of integration. This leads to:

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

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

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

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 2.3.2:

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

The required bounds are provided in Lemmas A.1.2 (term A_{1i}), A.1.3 (term A_{2i}), A.1.4 (term A_{3i}). ■

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 2.5.4, 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 2.5.4, 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

$$\begin{aligned} N\mathbb{P}(\mathcal{T}_3^c) &= N\mathbb{P}\{\max_{1\leq i\leq N} |\bar{U}_{i,-1}(\rho)| \geq L_N\} \\ &\leq N\sum_{i=1}^{N} \mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} \\ &= N^2\mathbb{P}\{|\bar{U}_{i,-1}(\rho)| \geq L_N\} \\ &\leq 2\exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N\right). \end{aligned} \tag{A.1.18}$$

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

$$\begin{aligned} N\mathbb{P}\mathcal{T}_4^c &= N\mathbb{P}\{\max_{1\leq i\leq N} |Y_{i0}| \geq L_N\} \\ &\leq N\sum_{i=1}^{N} \mathbb{P}\{|Y_{i0}| \geq L_N\} \\ &= N^2 \int \left[\int_{L_N}^{\infty} \pi(y_0|\lambda)dy_0 + \int_{-\infty}^{-L_N} \pi(y_0|\lambda)dy_0\right] \pi_\lambda(\lambda)d\lambda \\ &\leq 2N^2 \int \exp\left[-m\left(L_N, \lambda\right)\right] \pi(\lambda)d\lambda \\ &\leq 2C_N \left(\sup_{|\lambda|\leq C_N} \exp\left[-m\left(L_N, \lambda\right) + 2\ln N\right]\right), \end{aligned} \tag{A.1.19}$$

where the last three lines hold by Assumptions 2.5.1 and 2.5.3.

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 2.5.1 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 $\mathcal{T}_3 \cap \mathcal{T}_4$:

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

which leads to

$$\max_{1\leq i,j\leq N}|\bar{Y}_{j,-1} - \bar{Y}_{i,-1}| \leq 2\max_{1\leq i\leq N}|\bar{Y}_{i,-1}| \leq 2\left(C_N + TL_N + \exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N\right)\right). \tag{A.1.20}$$

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

$$\max_{1\leq i,j\leq N}|(\hat{\rho}-\rho)(\bar{Y}_{j,-1} - \bar{Y}_{i,-1})| \leq \frac{2\left(C_N + TL_N + \exp\left(-\frac{L_N^2}{2\sigma_{\bar{U}}^2} + 2\ln N\right)\right)}{L_N^2} \tag{A.1.21}$$

Recall that $C_N = o(N^\epsilon)$ is the truncation for the support of the prior of λ (Assumption 2.5.1). We will choose

$$L_N = o(N^\epsilon) \text{ such that } L_N = \max\left\{\sigma_{\bar{U}}\sqrt{2(2+\epsilon)\ln N}, K(\sqrt{\ln N} + C_N), \frac{1}{B_N}, C_N\right\}, \tag{A.1.22}$$

so that we can deduce

$$\begin{aligned}&N\mathbb{P}\mathcal{T}_1^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_2^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_3^c = o(N^\epsilon), \quad N\mathbb{P}\mathcal{T}_4^c = o(N^\epsilon) \\ &(A.1.20) = o(N^\epsilon), \quad (A.1.21) = o(N^\epsilon).\end{aligned} \tag{A.1.23}$$

for any ϵ.

Term A_{1i}

Lemma A.1.2. *Suppose the assumptions in Theorem 2.5.5 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.1.2. We begin with the following bound:

$$\begin{aligned}
&|A_{1i}|\\
=\;&\left|\left[\mu\big(\hat{\lambda}_i(\hat{\rho}),\hat{\sigma}^2/T+B_N^2,\hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}),Y_{i0})\big)\right]^{C_N}-\mu\big(\hat{\lambda}_i(\rho),\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)\right|\\
\leq\;&\left|\left[\mu\big(\hat{\lambda}_i(\hat{\rho}),\hat{\sigma}^2/T+B_N^2,\hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}),Y_{i0})\big)\right]^{C_N}\right|+\left|\mu\big(\hat{\lambda}_i(\rho),\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)\right|\\
\leq\;&2C_N. \qquad \text{(A.1.24)}
\end{aligned}$$

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 2.5.1, 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,

$$\begin{aligned}
N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2]\;&\leq\;N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2\mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)]+C_N^2N\left(\mathbb{P}\mathcal{T}_1^c+\mathbb{P}\mathcal{T}_2^c+\mathbb{P}\mathcal{T}_3^c+\mathbb{P}\mathcal{T}_4^c\right)\\
&\leq\;N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2\mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)]+o(N^{\epsilon_0}). \qquad \text{(A.1.25)}
\end{aligned}$$

The bound for the second term follows from the fact that (A.1.23) and (A.1.24) 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.1.25). 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.1.28) 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

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i}^c)] \leq o(N^{\epsilon_0}) \tag{A.1.26}$$

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}[A_{1i}^2 \mathbb{I}(\mathcal{T}_{5i}^c)] \leq o(N^{\epsilon_0}). \tag{A.1.27}$$

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.1.26). Using the bound for $|A_{1i}|$ in (A.1.24) we obtain:

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

The last equality holds by the definition of C'_N found in (A.1.28) below. This establishes (A.1.26).

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:"

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

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

$$C'_N = (1+k)\left(\sqrt{\ln N}+C_N\right), \quad k > \max\{0, \sqrt{2\sigma^2/T}-1\}. \tag{A.1.28}$$

This is the rate that appears in Assumption 2.5.2.

For B_2, notice that under Assumption 2.5.3(ii) we obtain

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

for any ϵ. This leads to the desired bound in (A.1.27).

Step 2. It remains to be shown that

$$N\mathbb{E}_{\theta}^{\mathcal{Y}^N}\left[A_{1i}^2 \mathbb{I}(\mathcal{T}_1)\mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)\mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i})\right] \leq o(N^{\epsilon_0}). \tag{A.1.29}$$

We introduce the following notation:

$$\begin{aligned} \widetilde{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \qquad (A.1.30) \\ d\widetilde{p}_i^{(-i)} &= \frac{1}{\partial\hat{\lambda}_i(\hat{\rho})}\partial\hat{p}^{(-i)}(\hat{\lambda}_i(\hat{\rho}), Y_{i0}) \\ \hat{p}_i^{(-i)} &= \hat{p}^{(-i)}(\hat{\lambda}_i(\rho), Y_{i0}) \\ d\hat{p}_i^{(-i)} &= \frac{1}{\partial\hat{\lambda}_i(\rho)}\partial\hat{p}^{-i}(\hat{\lambda}_i(\rho), Y_{i0}) \\ p_i &= p(\hat{\lambda}_i(\rho), Y_{i0}) \\ p_{*i} &= p_*(\hat{\lambda}_i(\rho), Y_{i0}) \\ dp_{*i} &= \frac{1}{\partial\hat{\lambda}_i(\rho)}\partial p_*(\hat{\lambda}_i(\rho), Y_{i0}). \end{aligned}$$

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 inequal-

ity, we obtain

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

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

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

First, using a slightly more general argument than the one used in the proof of Lemma A.1.4, 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

$$\left(\frac{\sigma^2}{T} + B_N^2\right)\left|\frac{dp_{*i}}{p_{*i}}\right| = \left|\hat{\lambda}_i(\rho) - \mathbb{E}_\theta\big[\lambda_i\big|\hat{\lambda}_i(\rho), Y_{i0}; p_*(\hat{\lambda}_i(\rho), Y_{i0})\big]\right| \leq C_N' + C_N, \qquad \text{(A.1.31)}$$

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 2.5.4 we obtain the bound

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

Finally, note that

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

Thus, the desired result follows if we show

$$N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}} - \frac{dp_{*i}}{p_{*i}}\right)^2 \mathbb{I}(\mathcal{T}_2)\mathbb{I}(\mathcal{T}_3)\mathbb{I}(\mathcal{T}_4)\mathbb{I}(\mathcal{T}_{5i})\mathbb{I}(\mathcal{T}_{6i})\right] = o(N^{\epsilon_0}) \tag{A.1.32}$$

To show (A.1.32), we have to control the denominator and consider the following truncation region:

$$\mathcal{T}_{7i} = \left\{(\hat{\lambda}_i, Y_{i0}) \,\middle|\, \widetilde{p}_i^{(-i)} > \frac{p_{*i}}{2}\right\}. \tag{A.1.33}$$

We first analyze (A.1.32) 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

$$\begin{aligned}
N\mathbb{E}_\theta^{\mathcal{Y}^N}&\left[\left(\frac{d\widetilde{p}_i^{(-i)}}{\widetilde{p}_i^{(-i)}}-\frac{dp_{*i}}{p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\\
&\leq 2N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{d\widetilde{p}_i^{(-i)}-dp_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\\
&\quad+2o(N^{\epsilon_0})N\mathbb{E}_\theta^{\mathcal{Y}^N}\left[\left(\frac{\widetilde{p}_i^{(-i)}-p_{*i}}{\widetilde{p}_i^{(-i)}-p_{*i}+p_{*i}}\right)^2\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4\mathcal{T}_{5i}\mathcal{T}_{6i}\mathcal{T}_{7i})\right]\\
&= 2B_{1i}+2o(N^{\epsilon_0})B_{2i},
\end{aligned}$$

say. The $o(N^{\epsilon_0})$ bound follows from (A.1.31). 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

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

say.

According to Lemma A.1.7(c) (see Section A.1.2)

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

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.1.7(e) (see Section A.1.2)

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

Because $1/B_N^2=o(N^{\epsilon})$ according to Assumption 2.5.2, 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 2.5.4. According to Lemma A.1.7(a) (see

Section A.1.2)

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

where $L_N = o(N^{\epsilon_0})$ was defined in (A.1.22). 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.1.7(e) (see Section A.1.2).

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

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

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,

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

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.1.7(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.1.7(f) (see Section A.1.2) 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.1.7(f) (see Section A.1.2)

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

which gives us the required result for Step 2.2. Combining the results from Steps 2.1 and 2.2 yields (A.1.29).

The bound in (A.1.25) now follows from (A.1.26), (A.1.27), and (A.1.29), which completes the proof of the lemma. ■

Term A_{2i}

Lemma A.1.3. *Suppose the assumptions in Theorem 2.5.5 hold. Then,*

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

Proof of Lemma A.1.3. 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

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

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

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

$$
\begin{aligned}
& N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right] \\
&\quad\leq\ N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right] \\
&\quad=\ N\int\int p_*(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2 d\hat{\lambda}_i dy_{i0} \\
&\quad\leq\ N\int\int p_*(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat{\lambda}_i dy_{i0} \\
&\qquad+N4C_N^2\mathbb{P}(\mathcal{T}_{5i}^c\cup\mathcal{T}_{6i}^c) \\
&\quad=\ N\int\int p_*(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2 \mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat{\lambda}_i dy_{i0}+o(N^{\epsilon_0}).
\end{aligned}
$$

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.1.2.

According to Lemma A.1.6, 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,

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

In turn, we obtain the following bound:

$$
\begin{aligned}
& N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]+\tau_N \\
&\quad\leq\ (1+o(1))N\int\int p(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T,p(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat{\lambda}_i dy_{i0} \\
&\qquad +o(N^{\epsilon_0}) \\
&\quad\leq\ (1+o(1))N\mathbb{E}_{\theta}^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i-\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right]+o(N^{\epsilon_0}) \\
&\quad\leq\ (1+o(1))N\mathbb{E}_{\theta}^{\mathcal{Y}^i,\lambda_i}\left[(\lambda_i-\mathbb{E}_{\theta,\mathcal{Y}^i}^{\lambda_i}[\lambda_i])^2\right]+N^{\epsilon_0},
\end{aligned}
$$

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

$$
\begin{aligned}
& N\mathbb{E}_{\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right] \\
=& N\iint p(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2 d\hat{\lambda}_i dy_{i0} \\
\leq& \iint p_*(\hat{\lambda}_i,y_{i0})\frac{p(\hat{\lambda}_i,y_{i0})}{p_*(\hat{\lambda}_i,y_{i0})}\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat{\lambda}_i dy_{i0} \\
&+N4C_N^2\mathbb{P}(\mathcal{T}_{5i}^c\cup\mathcal{T}_{6i}^c) \\
=&(1+o(1))N\iint p_*(\hat{\lambda}_i,y_{i0})\left(\mu\big(\hat{\lambda}_i(\rho),y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),y_{i0})\big)-\lambda_i\right)^2 \\
&\times\mathbb{I}(\mathcal{T}_{5i}\mathcal{T}_{6i})d\hat{\lambda}_i dy_{i0}+o(N^{\epsilon}),\quad \text{any } \epsilon>0 \\
\leq&(1+o(1))N\mathbb{E}_{*,\theta}^{\mathcal{Y}_i,\lambda_i}\left[\left(\mu\big(\hat{\lambda}_i(\rho),Y_{i0},\sigma^2/T+B_N^2,p_*(\hat{\lambda}_i(\rho),Y_{i0})\big)-\lambda_i\right)^2\right]+\tau_N.
\end{aligned}
$$

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

Term A_{3i}

Lemma A.1.4. *Suppose the assumptions in Theorem 2.5.5 hold. Then, for any $\epsilon > 0$:*

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

Proof of Lemma A.1.4. Using the Cauchy-Schwarz inequality, we can bound

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

By Assumption 2.5.4, we have

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

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 2.5.1 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):

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

for some finite constant M, where $\bar{C}$ is the constant in Assumption 2.5.3(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 2.5.3(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

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

Further Details

We now provide more detailed derivations for some of the bounds used in Section A.1.2. Recall that

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

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,

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

Consider the region $\mathcal{T}_2\cap\mathcal{T}_3\cap\mathcal{T}_4$. First, using (A.1.21) we can bound

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

Thus,

$$\begin{aligned}
&\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-(\widetilde{\rho}-\rho)\left(\frac{\bar{Y}_{j,-1}-\bar{Y}_{i,-1}}{B_N}\right)\right)\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4)\\
&\leq\ \phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}+\left(\frac{M}{L_NB_N}\right)\right)\mathbb{I}\left\{\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\leq-\frac{M}{L_NB_N}\right\}\\
&\quad+\phi(0)\mathbb{I}\left\{\left|\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right|\leq\frac{M}{L_NB_N}\right\}\\
&\quad+\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}-\left(\frac{M}{L_NB_N}\right)\right)\mathbb{I}\left\{\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\geq\frac{M}{L_NB_N}\right\}\\
&=\ \bar{\phi}\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right),
\end{aligned}$$

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

Second, we can use the bound

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

Third, for the region $\mathcal{T}_3\cap\mathcal{T}_4$ we can deduce from (A.1.20) that

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

Therefore,

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

Now, define the function

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

Because for random variables with bounded densities and Gaussian tails all moments exist and because $L_NB_N>1$ by definition of L_N in (A.1.22), 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 $R_{2i}(\widetilde{\rho})$:

$$\left|R_{2i}(\widetilde{\rho})\mathbb{I}(\mathcal{T}_2\mathcal{T}_3\mathcal{T}_4)\right| \leq \frac{ML_N}{N-1}\sum_{j\neq i}^{N}\frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right). \tag{A.1.34}$$

For the subsequent analysis it is convenient define the function

$$f(\hat{\lambda}_j-\hat{\lambda}_i, Y_{j0}-Y_{i0}) = \frac{1}{B_N^2}\bar{\phi}_*\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\phi\left(\frac{Y_{j0}-Y_{i0}}{B_N}\right). \tag{A.1.35}$$

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.1.2. 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.1.5. *Suppose the assumptions required for Theorem 2.5.5 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.1.5. We have

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

The inner integral is

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

say.

Notice that

$$\begin{aligned}
I_1 &= \int \frac{1}{B_N}\bar{\phi}_*\left(\frac{\hat{\lambda}-\hat{\lambda}_i}{B_N}\right)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)d\hat{\lambda} \\
&= \int \bar{\phi}_*(\lambda^*)^m \frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i+B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^* \\
&= \int \bar{\phi}_*(\lambda^*)^m \exp\left(-\left((\hat{\lambda}_i-\lambda_i)B_N\lambda^*\right)\frac{1}{\sigma^2/T}\right)\exp\left(-\frac{1}{2}\left(\frac{B_N\lambda^*}{\sigma/\sqrt{T}}\right)^2\right)d\lambda^* \\
&\quad \times \left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&\leq M\left(\int \bar{\phi}_*(\lambda^*)^m \exp\left(v_N\lambda^*\right)d\lambda^*\right)\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] \\
&\leq M\left[\frac{1}{\sigma/\sqrt{T}}\exp\left(-\frac{1}{2}\left(\frac{\hat{\lambda}_i-\lambda_i}{\sigma/\sqrt{T}}\right)^2\right)\right] = Mp(\hat{\lambda}_i|\lambda_i, Y_{i0}).
\end{aligned}$$

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 2.5.1). Thus, $v_N = B_N(C'_N + 2C_N)$. According to Assumption 2.5.2 $v_N = B_N(C'_N + 2C_N) = o(1)$. Thus, the last inequality holds because

$\int \bar{\phi}_*(\lambda^*)^m \exp(v_N\lambda^*)\, d\lambda^*$ 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 2.5.3(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.1.6. *Suppose the assumptions required for Theorem 2.5.5 are satisfied. Then,*

$$\sup_{(\hat{\lambda}_i, Y_{i0})\in\mathcal{T}_{5i}\cap\mathcal{T}_{6i}} \frac{p_i}{p_{*i}} = 1 + o(1) \tag{A.1.36}$$

$$\sup_{(\hat{\lambda}_i, Y_{i0})\in\mathcal{T}_{5i}\cap\mathcal{T}_{6i}} \frac{p_{*i}}{p_i} = 1 + o(1). \tag{A.1.37}$$

Proof of Lemma A.1.6. We begin by verifying (A.1.36). 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 2.5.1 and $|\hat{\lambda}_i| \leq C'_N$ in the region $\mathcal{T}_{5i}$, for some finite constant M we have

$$\begin{aligned}
\frac{1}{\sqrt{\sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{\sigma^2/T}}\right) &= \frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right) \\
&\quad \times \frac{\sqrt{B_N^2 + \sigma^2/T}}{\sqrt{\sigma^2/T}} \exp\left\{-\frac{1}{2}\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right)^2 \frac{B_N^2}{\sigma^2/T}\right\} \\
&\leq \frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right) \\
&\quad \times \sqrt{1 + MB_N^2} \exp(-M(C'_N + C_N)^2 B_N^2) \\
&= (1 + o(1))\frac{1}{\sqrt{B_N^2 + \sigma^2/T}}\phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{B_N^2 + \sigma^2/T}}\right), \qquad \text{(A.1.38)}
\end{aligned}$$

where $o(1)$ is uniform in $(\hat{\lambda}_i, Y_{i0}) \in \mathcal{T}_{5i} \cap \mathcal{T}_{6i}$. Here we used Assumption 2.5.2 which implies that $v_N = (C'_N + C_N)B_N = o(1)$.

According to Assumption 2.5.3(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

$$\pi(y_{i0}|\lambda_i) \leq (1 + o(1)) \int \frac{1}{B_N}\phi\left(\frac{y_{i0} - \tilde{y}_{i0}}{B_N}\right)\pi(\tilde{y}_{i0}|\lambda_i)d\tilde{y}_{i0}. \qquad \text{(A.1.39)}$$

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

Then, by combining the bounds in (A.1.38) and (A.1.39) we deduce

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

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.1.36). A similar argument can be used to establish (A.1.37). ■

Lemma A.1.7. *Under the assumptions required for Theorem 2.5.5, we obtain the following bounds:*

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

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

Proof of Lemma A.1.7. **Part (a).** Recall the following definitions

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

First, recall that according to (A.1.34), in the region $\mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$

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

Then,

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

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

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

where

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

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

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

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

Using the C_R inequality, Jensen's inequality, the conditional independence of a_j^2 and a_i^2 and Lemma A.1.5, 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.1.6 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

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

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

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

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 [1] (e.g., Lemma 19.32 in van der Vaart (1998)),

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

Using an argument similar to the proof of Lemma A.1.5 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.1.6 we can find a constant c such that $p_i \leq (1+c)p_{*i}$ and $p_{*i} \leq (1+c)p_i$.

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

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

This leads to

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

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

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

as desired. ■

A.1.3 Derivations for Section 2.6

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

$$(\theta_*, \xi_*) = \text{argmax}_{\theta,\xi}\ \mathbb{E}_{\theta_0}^{\mathcal{Y}}\left[\ln p(Y, X_2|H, \theta, \xi)\right], \tag{A.1.40}$$

and verifying that $\theta_* = \theta_0$. Here, $p(y, x_2|h, \theta, \xi)$ is given in (2.4.10). Because the observations are conditionally independent across i and the likelihood function is symmetric with respect to i, we can drop the i subscripts.

We make some adjustment to the notation. The covariance matrix Σ only depends on γ, but not on (ρ, α). Moreover, we will split ξ into the parameters that characterize the

conditional mean of λ, denoted by Φ, and ω, which are the non-redundant elements of the prior covariance matrix $\underline{\Omega}$. Finally, we define

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

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

$$\begin{aligned} &\ln p(y, x_2|h, \theta_1, \gamma, \phi, \omega) \qquad\qquad (A.1.41)\\ &= C - \frac{1}{2}\ln|\Sigma(\gamma)| - \frac{1}{2}\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big)'\Sigma^{-1}(\gamma)\big(\tilde{y}(\theta_1) - w\hat{\lambda}(\theta)\big)\\ &\quad -\frac{1}{2}\ln\big|\underline{\Omega}\big| + \frac{1}{2}\ln\big|\bar{\Omega}(\gamma,\omega)\big|\\ &\quad -\frac{1}{2}\Big(\hat{\lambda}(\theta)'w'\Sigma^{-1}(\gamma)w\hat{\lambda}(\theta) + \phi'\tilde{h}\underline{\Omega}^{-1}\tilde{h}'\phi - \bar{\lambda}'(\theta,\xi)\bar{\Omega}^{-1}(\gamma,\omega)\bar{\lambda}(\theta,\xi)\Big), \end{aligned}$$

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

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

where

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

Note that

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

In turn, we can write

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

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

$$\begin{aligned}&\mathbb{E}\big[\ln p(Y|H,\rho,\gamma,\phi,\omega)\big] \qquad (A.1.42)\\ &= C-\frac{T-1}{2}\ln|\gamma^2|-\frac{1}{2\gamma^2}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big]-\frac{1}{2}\ln\left|\omega^2+\gamma^2/T\right|\\ &\quad -\frac{1}{2(\omega^2+\gamma^2/T)}\Big(\big(\phi-\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big)'\mathbb{E}[\tilde{H}\tilde{H}']\big(\phi-\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big)\\ &\quad -\mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]+\mathbb{E}[\hat{\lambda}^2(\rho)]\Big).\end{aligned}$$

We deduce that

$$\phi_*(\rho)=\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]. \qquad (A.1.43)$$

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,

$$\phi_*(\rho_0) = \big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.1.44}$$

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

Plugging the pseudo-true value for ϕ into (A.1.42) yields the concentrated objective function

$$\begin{aligned} &\mathbb{E}\big[\ln p(Y|H,\rho,\gamma,\phi_*(\rho),\omega)\big] \qquad (A.1.45)\\ &= C - \frac{T-1}{2}\ln|\gamma^2| - \frac{1}{2\gamma^2}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big] \\ &\quad -\frac{1}{2}\ln\big|\omega^2+\gamma^2/T\big| - \frac{1}{2(\omega^2+\gamma^2/T)}\big(\mathbb{E}[\hat{\lambda}^2(\rho)] - \mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big). \end{aligned}$$

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

$$\begin{aligned} \gamma_*^2(\rho) &= \frac{1}{T-1}\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big] \qquad (A.1.46)\\ \omega_*^2(\rho)+\gamma_*^2(\rho)/T &= \big(\mathbb{E}[\hat{\lambda}^2(\rho)] - \mathbb{E}[\hat{\lambda}(\rho)\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\hat{\lambda}(\rho)]\big). \end{aligned}$$

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

$$\gamma_*^2(\rho_0) = \gamma_0^2, \quad \omega_*^2(\rho_0) = \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]. \tag{A.1.47}$$

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.1.46), we can now concentrate out γ^2 and ω^2 from the objective function (A.1.45):

$$\begin{aligned}
&\mathbb{E}\big[\ln p(Y|H,\rho,\gamma_*(\rho),\phi_*(\rho),\omega_*(\rho)\big] \qquad (A.1.48)\\
&= C - \frac{T-1}{2}\ln\big|\mathbb{E}\big[\tilde{Y}(\rho)'(I-\iota\iota'/T)\tilde{Y}(\rho)\big]\big| \\
&\quad -\frac{1}{2}\ln\big|\mathbb{E}[\tilde{Y}'(\rho)\iota\iota'\tilde{Y}(\rho)] - \mathbb{E}[\tilde{Y}'(\rho)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho)]\big|.
\end{aligned}$$

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.1.48) with respect to ρ yields

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

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

$$\begin{aligned}
\text{F.O.C.}(\rho_0) &= (T-1)\mathbb{E}\big[X'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \qquad (A.1.49)\\
&\quad \times\Big(\mathbb{E}[\tilde{Y}'(\rho_0)\iota\iota'\tilde{Y}(\rho_0)] - \mathbb{E}[\tilde{Y}'(\rho_0)\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\Big) \\
&\quad +\mathbb{E}\big[\tilde{Y}(\rho_0)'(I-\iota\iota'/T)\tilde{Y}(\rho_0)\big] \\
&\quad \times\Big(\mathbb{E}[X'\iota\iota'\tilde{Y}(\rho_0)] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\iota'\tilde{Y}(\rho_0)]\Big).
\end{aligned}$$

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.1.49) as follows:

$$
\begin{aligned}
&\text{F.O.C.}(\rho_0) \qquad\qquad (\text{A.1.50}) \\
&= -(T-1)\big(\mathbb{E}[X'\iota\iota'u]\big)\Big(\gamma_0^2 + T\big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big) \\
&\quad + \Big(\mathbb{E}[X'\iota\iota'u] + T\big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\big)\Big)(T-1)\gamma_0^2 \\
&= T(T-1)\Big[\gamma_0^2\Big(\mathbb{E}[X'\iota\lambda] - \mathbb{E}[X'\iota\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big) \\
&\quad - \mathbb{E}[X'\iota\iota'u]\Big(\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}']\big(\mathbb{E}[\tilde{H}\tilde{H}']\big)^{-1}\mathbb{E}[\tilde{H}\lambda]\Big)\Big].
\end{aligned}
$$

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,

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

Thus, we obtain

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

Using these expressions, most terms that appear in (A.1.50) cancel out and the condition simplifies to

$$\text{F.O.C.}(\rho_0) = T(T-1)\gamma_0 a_T \Big(\mathbb{E}[Y_0\lambda] - \mathbb{E}[Y_0\tilde{H}'] \big(\mathbb{E}[\tilde{H}\tilde{H}'] \big)^{-1} \mathbb{E}[\tilde{H}\lambda] \Big). \tag{A.1.51}$$

Now consider

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

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

$$\begin{aligned}
\rho_* &= \rho_0, \quad \gamma_*^2 = \gamma_0, \quad \phi_* = \big(\mathbb{E}[\tilde{H}\tilde{H}'] \big)^{-1} \mathbb{E}[\tilde{H}\lambda], \\
\omega_*^2 &= \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda\tilde{H}'] \big(\mathbb{E}[\tilde{H}\tilde{H}'] \big)^{-1} \mathbb{E}[\tilde{H}\lambda]. \quad \blacksquare
\end{aligned} \tag{A.1.52}$$

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

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

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

$$\lambda|(a_{1:T}, Y_{0:T}) \sim N\big(\bar{\lambda}(a_{1:T}), \bar{\Omega}(a_{1:T})\big), \tag{A.1.53}$$

where

$$\begin{aligned}
\bar{\Omega}(a_{1:T}) &= \big(\underline{\Omega}^{-1} + \sum_{t=1}^{T}(\gamma^2(a_t))^{-1}\big)^{-1} \\
\bar{\lambda}(a_{1:T}) &= \bar{\Omega}(a_{1:T})\big(\underline{\Omega}^{-1}(\phi_0 + \phi_1 Y_0) + \sum_{t=1}^{T}(\gamma^2(a_t))^{-1}\tilde{Y}_t\big).
\end{aligned}$$

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

$$\mathbb{P}\big(a_t = 1|\lambda, Y_{0:T}\big) \tag{A.1.54}$$
$$= \frac{p_u(\gamma_+)^{-1}\exp\left\{-\frac{1}{2\gamma_+^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\}}{p_u(\gamma_+)^{-1}\exp\left\{-\frac{1}{2\gamma_+^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\} + (1-p_u)(\gamma_-)^{-1}\exp\left\{-\frac{1}{2\gamma_-^2}(Y_t - \rho Y_{t-1} - \lambda)^2\right\}}.$$

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.1.53) and (A.1.54). Then,

$$\begin{aligned}
\widehat{\mathbb{E}}[\lambda|Y_{0:T}] &= \frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}}\bar{\lambda}(a_{1:T}^s), \qquad (A.1.55)\\
\widehat{\mathbb{V}}[\lambda|Y_{0:T}] &= \left(\frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}}\bar{\Omega}(a_{1:T}^s) + \bar{\lambda}^2(a_{1:T}^s)\right) - \left(\frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}}\bar{\lambda}(a_{1:T}^s)\right)^2.
\end{aligned}$$

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

$$\lambda|(a_{1:T}, Y_{0:T}) \sim N\big(\bar{\lambda}(a_{1:T}), \bar{\Omega}\big), \tag{A.1.56}$$

where

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

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

$$\begin{aligned}
&\mathbb{P}\big(a_t = 1|\lambda, Y_{0:T}\big) \qquad\qquad (A.1.57)\\
&= \frac{p_u \exp\Big\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda - \mu_+)^2\Big\}}{p_u \exp\Big\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda - \mu_+)\Big\} + (1-p_u)\exp\Big\{-\frac{1}{2\gamma^2}(Y_t - \rho Y_{t-1} - \lambda + \mu_-)^2\Big\}}.
\end{aligned}$$

The posterior mean $\mathbb{E}[\lambda|Y_{0:T}]$ can be approximated with the following Gibbs sampler. Generate a sequence of draws $\{\lambda^s, a^s_{1:T}\}_{s=1}^{N_{sim}}$ by iterating over the conditional distributions given in (A.1.56) and (A.1.57). Then,

$$\begin{aligned}
\widehat{\mathbb{E}}[\lambda|Y_{0:T}] &= \frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}} \bar{\lambda}(a^s_{1:T}), \qquad\qquad (A.1.58)\\
\widehat{\mathbb{V}}[\lambda|Y_{0:T}] &= \left(\bar{\Omega} + \frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}} \bar{\lambda}^2(a^s_{1:T})\right) - \left(\frac{1}{N_{sim}}\sum_{s=1}^{N_{sim}} \bar{\lambda}(a^s_{1:T})\right)^2.
\end{aligned}$$

A.2 Data Set

The construction of our data is based on Covas *et al.* (2014). We downloaded 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

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

where

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

Here net interest income is the difference between total interest income and expenses. It excludes provisions for loan and lease losses. Non-interest income includes various types of fees, trading revenue, as well as net gains on asset sales. Non-interest expenses include, for instance, salaries and employee benefits and expenses of premises and fixed assets. As in Covas *et al.* (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 21. 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

Table 21: Size of Adjusted Rolling Samples

Sample		Adjustment Step					
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

Table 22: 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.

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.

Descriptive statistics for the $T = 3$ and $T = 11$ rolling samples are reported in Table 21. 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.

A.3 Additional Empirical Results

Table 23: Parameter Estimates: $\hat{\theta}_{QMLE}$, Parametric Tweedie Correction

			Intercept			Unemployment			
τ	$\hat{\rho}$	$\hat{\sigma}^2$	$\hat{\phi}_{10}$	$\hat{\phi}_{11}$	$\hat{\underline{\omega}}_1^2$	$\hat{\phi}_{20}$	$\hat{\phi}_{21}$	$\hat{\underline{\omega}}_2^2$	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 $T = 5$.

APPENDIX B

Density Forecasts and Young Firm Dynamics

B.1 Notations

$U(a,b)$ represents a **uniform distribution** with minimum value a and maximum value b. If $a=0$ and $b=1$, we obtain the standard uniform distribution, $U(0,1)$.

$N\left(\mu,\sigma^2\right)$ or $N\left(x;\mu,\sigma^2\right)$ stands for a **Gaussian distribution** with mean μ and variance σ^2. Its probability distribution function (pdf) is given by $\phi\left(x;\mu,\sigma^2\right)$. When $\mu=0$ and $\sigma^2=1$ (i.e. standard normal), we reduce the notation to $\phi(x)$. The corresponding cumulative distribution functions (cdf) are denoted as $\Phi\left(x;\mu,\sigma^2\right)$ and $\Phi(x)$, respectively. The same convention holds for multivariate normal, where $N(\mu,\Sigma)$, $N(x;\mu,\Sigma)$, $\phi(x;\mu,\Sigma)$, and $\Phi(x;\mu,\Sigma)$ are for the distribution with the mean vector μ and the covariance matrix Σ.

$TN\left(\mu,\sigma^2,a,b\right)$ denotes a **truncated normal distribution** with μ and σ^2 being the mean and variance before truncation, and a and b being the lower and upper end of the truncated interval.

The **gamma distribution** is denoted as $\mathrm{Ga}(x;a,b)$ with probability density function being $f_{\mathrm{Ga}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$. The according **inverse-gamma distribution** is given by $\mathrm{IG}(x;a,b)$ with probability density function being $f_{\mathrm{IG}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{-a-1}e^{-b/x}$. The $\Gamma(\cdot)$ in the denominators is the gamma function.

The **inverse Wishart distribution** is a generalization of the inverse gamma distribution to multi-dimensional setups. Let Ω be a $d\times d$ matrix, then the inverse Wishart distribution is denoted as $\mathrm{IW}(\Omega;\Psi,\nu)$, and its pdf is $f_{\mathrm{IW}}(\Omega;\Psi,\nu)=\frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu d}{2}}\Gamma_d(\frac{\nu}{2})}|\Omega|^{-\frac{\nu+d+1}{2}}e^{-\frac{1}{2}tr(\Psi\Omega^{-1})}$. When Ω is a scalar, the inverse Wishart distribution is reduced to a inverse-gamma distri-

bution with $a = \nu/2,\ b = \Psi/2$.

$\mathbf{1}(\cdot)$ is an **indicator function** that equals 1 if the condition in the parenthesis is satisfied and equals 0 otherwise.

I_N is an $N \times N$ **identity matrix**.

In the **panel data** setup, for a generic variable z, which can be $v,\ w\ , x,$ or y, z_{it} is a $d_z \times 1$ vector, and $z_{i,t_1:t_2} = (z_{it_1}, \cdots, z_{it_2})$ is a $d_z \times (t_2 - t_1 + 1)$ matrix.

$\|\cdot\|$ represents the **Euclidean norm**, i.e. for a n-dimensional vector $z = [z_1, z_2, \cdots, z_n]'$, $\|z\| = \sqrt{z_1^2 + \cdots + z_n^2}$.

$\text{supp}(\cdot)$ denotes the **support** of a probability measure.

B.2 Algorithms

B.2.1 Hyperparameters

Recall the prior for the common parameters:

$$(\beta, \sigma^2) \sim N\left(m_0^\beta, \psi_0^\beta \sigma^2\right) \text{IG}\left(\sigma^2;\ a_0^{\sigma^2}, b_0^{\sigma^2}\right).$$

The hyperparameters are chosen in a relatively ignorant sense without inferring too much from the data except aligning the scale according to the variance of the data.

$$a_0^{\sigma^2} = 2, \tag{B.2.1}$$

$$b_0^{\sigma^2} = \hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right) \cdot \left(a_0^{\sigma^2} - 1\right) = \hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right), \tag{B.2.2}$$

$$m_0^\beta = 0.5, \tag{B.2.3}$$

$$\psi_0^\beta = \frac{1}{b_0^{\sigma^2} / \left(a_0^{\sigma^2} - 1\right)} = \frac{1}{\hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right)}. \tag{B.2.4}$$

In equation (B.2.2) here and equation (B.2.5) below, $\hat{E}_i^t$ and $\widehat{Var}_i^t$ stand for the sample mean and variance for firm i over $t = 1, \cdots, T$, and $\hat{E}^i$ and $\widehat{Var}^i$ are the sample mean and variance over the whole cross-section $i = 1, \cdots, N$. Equation (B.2.2) ensures that on average the prior and the data have a similar scale. Equation (B.2.3) conjectures that the young firm dynamics are highly likely persistent and stationary. Since we don't have strong prior information in the common parameters, their priors are chosen to be not very restrictive. Equation (B.2.1) characterizes a rather less informative prior on σ^2 with infinite variance, and Equation (B.2.4) assumes that the prior variance of β is equal to 1 on average.

The hyperpriors for the DPM prior are specified as:

$$G_0\left(\mu_k, \omega_k^2\right) = N\left(\mu_k;\ m_0^\lambda, \psi_0^\lambda \omega_k^2\right) \mathrm{IG}\left(\omega_k^2;\ a_0^\lambda, b_0^\lambda\right),$$
$$\alpha \sim \mathrm{Ga}\left(\alpha;\ a_0^\alpha, b_0^\alpha\right).$$

Similarly, the hyperparameters are chosen to be:

$$a_0^\lambda = 2,\ b_0^\lambda = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right) \cdot \left(a_0^\lambda - 1\right) = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right), \tag{B.2.5}$$
$$m_0^\lambda = 0,\ \psi_0^\lambda = 1,$$
$$a_0^\alpha = 2,\ b_0^\alpha = 2. \tag{B.2.6}$$

where b_0^λ is selected to match the scale, while a_0^λ, m_0^λ, and ψ_0^λ yields a relatively ignorant and diffuse prior. Following Ishwaran and James (2001, 2002), the hyperparameters for the DP scale parameter α in equation (B.2.6) allows for a flexible component structure with a wide range of component numbers. The truncated number of components is set to be $K = 50$, so that the approximation error is uniformly bounded by Ishwaran and James (2001) Theorem 2:

$$\left\| f^{\lambda,K} - f^\lambda \right\| \sim 4N \exp\left(-\frac{K-1}{\alpha}\right) \leq 2.10 \times 10^{-18},$$

at the prior mean of α ($\bar{\alpha} = 1$) and cross-sectional sample size $N = 1000$.

I have also examined other choices of hyperparameters, and results are not very sensitive to hyperparameters as long as the implied priors are flexible enough to cover the range of observables.

B.2.2 Random-Walk Metropolis-Hastings

When there is no closed-form conditional posterior distribution in some MCMC steps, it is helpful to employ the Metropolis-within-Gibbs sampler and use the random-walk Metropolis-Hastings (RWMH) algorithm for those steps. The adaptive RWMH algorithm below is based on Atchadé and Rosenthal (2005) and Griffin (2016), which adaptively adjust the random walk step size in order to keep acceptance rates around certain desirable percentage.

Algorithm B.2.1. *(Adaptive RWMH)*

Let us consider a generic variable θ. For each iteration $s = 1, \cdots, n_{sim}$,

1. *Draw candidate $\tilde{\theta}$ from the random-walk proposal density $\tilde{\theta} \sim N\left(\theta^{(s-1)}, \zeta^{(s)}\Sigma\right)$.*
2. *Calculate the acceptance rate*

$$a.r.(\tilde{\theta}|\theta^{(s-1)}) = \min\left(1, \frac{p(\tilde{\theta}|\cdot)}{p(\theta^{(s-1)}|\cdot)}\right),$$

   where $p(\theta|\cdot)$ is the conditional posterior distribution of interest.
3. *Accept the proposal and set $\theta^{(s)} = \tilde{\theta}$ with probability $a.r.(\tilde{\theta}|\theta^{(s-1)})$. Otherwise, reject the proposal and set $\theta^{(s)} = \theta^{(s-1)}$.*
4. *Update the random-walk step size for the next iteration,*

$$\log \zeta^{(s+1)} = \rho\left(\log \zeta^{(s)} + s^{-c}\left(a.r.(\tilde{\theta}|\theta^{(s-1)}) - a.r.^{\star}\right)\right),$$

   where $0.5 < c \leq 1$, $a.r.^{\star}$ is the target acceptance rate, and

$$\rho(x) = \min(|x|, \bar{x}) \cdot sgn(x),$$

where $\bar{x} > 0$ is a very large number.

Remark B.2.2. (i) In step 1, since the algorithms in this paper only consider RWMH on conditionally independent scalar variables, Σ is simply taken to be 1.

(ii) In step 4, I choose $c = 0.55$, a.r.$^\star = 30\%$ in the numerical exercises, following Griffin (2016).

B.2.3 Details on Posterior Samplers

The formulas below focus on the (correlated) random coefficients model in Algorithms 3.5.1 and 3.5.2 where the (correlated) random effects model in Algorithms 3.3.1 and 3.3.2 are special cases with solely univariate λ_i.

Step 2: Component Parameters

Random Coefficients Model For $z = \lambda,\ l$ and $k^z = 1, \cdots, K^z$, draw $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a scalar) $p\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$
\begin{aligned}
\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right) &\sim N\left(\mu_{k^z}^{z(s)};\ m_{k^z}^z, \psi_{k^z}^z \Omega_{k^z}^{z(s)}\right) \mathrm{IW}\left(\Omega_{k^z}^{z(s)};\ \Psi_{k^z}^z, \nu_{k^z}^z\right), \\
\hat{m}_{k^z}^z &= \frac{1}{n_{k^z}^{z(s-1)}} \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)}, \\
\psi_{k^z}^z &= \left(\left(\psi_0^z\right)^{-1} + n_{k^z}^{z(s-1)}\right)^{-1}, \\
m_{k^z}^z &= \psi_{k^z}^z \left(\left(\psi_0^z\right)^{-1} m_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)}\right), \\
\nu_{k^z}^z &= \nu_0^z + n_{k^z}^{z(s-1)}, \\
\Psi_{k^z}^z &= \Psi_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} \left(z_i^{(s-1)}\right)^2 + m_0^{z\prime} \left(\psi_0^z\right)^{-1} m_0^z - m_{k^z}^{z\prime} \left(\psi_{k^z}^z\right)^{-1} m_{k^z}^z.
\end{aligned}
$$

Correlated Random Coefficients Model Due to the complexity arising from the conditional structure, I break the updating procedure for $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ into two steps. For $z = \lambda,\ l$ and $k^z = 1, \cdots, K^z$,

(a) Draw $\mu_{k^z}^{z(s)}$ from a matricvariate-normal distribution (or a multivariate-normal distribution if z is a scalar) $p\left(\mu_{k^z}^{z(s)} \middle| \Omega_{k^z}^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$
\begin{aligned}
\operatorname{vec}\left(\mu_{k^z}^{z(s)}\right) &\sim N\left(\operatorname{vec}\left(\mu_{k^z}^{z(s)}\right);\ \operatorname{vec}\left(m_{k^z}^{z}\right), \psi_{k^z}^{z}\right), \\
\hat{m}_{k^z}^{z,zc} &= \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)} \left[1, c_{i0}'\right], \\
\hat{m}_{k^z}^{z,cc} &= \sum_{i \in J_{k^z}^{z(s-1)}} \left[1, c_{i0}'\right]' \left[1, c_{i0}'\right], \\
\hat{m}_{k^z}^{z} &= \hat{m}_{k^z}^{z,zc} \left(\hat{m}_{k^z}^{z,cc}\right)^{-1}, \\
\psi_{k^z}^{z} &= \left[(\psi_0^z)^{-1} + \hat{m}_{k^z}^{z,cc} \otimes \left(\Omega_{k^z}^{z(s-1)}\right)^{-1}\right]^{-1}, \\
\operatorname{vec}\left(m_{k^z}^{z}\right) &= \psi_{k^z}^{z} \left[(\psi_0^z)^{-1} \operatorname{vec}\left(m_0^z\right) + \left(\hat{m}_{k^z}^{z,cc} \otimes \left(\Omega_{k^z}^{z(s-1)}\right)^{-1}\right) \operatorname{vec}\left(\hat{m}_{k^z}^{z}\right)\right],
\end{aligned}
$$

where $\operatorname{vec}(\cdot)$ denotes matrix vectorization, and $\otimes$ is the Kronecker product.

(b) Draw $\Omega_{k^z}^{z(s)}$ from an inverse-Wishart distribution (or an inverse-gamma distribution if z is a scalar) $p\left(\Omega_{k^z}^{z(s)} \middle| \mu_{k^z}^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$
\begin{aligned}
\Omega_{k^z}^{z(s)} &\sim \mathrm{IW}\left(\Omega_{k^z}^{z(s)};\ \Psi_{k^z}^{z}, \nu_{k^z}^{z}\right), \\
\nu_{k^z}^{z} &= \nu_0^z + n_{k^z}^{z(s-1)}, \\
\Psi_{k^z}^{z} &= \Psi_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} \left(z_i^{(s-1)} - \mu_{k^z}^{z(s)} \left[1, c_{i0}'\right]'\right) \left(z_i^{(s-1)} - \mu_{k^z}^{z(s)} \left[1, c_{i0}'\right]'\right)'.
\end{aligned}
$$

Step 4: Individual-specific Parameters

For $i = 1, \cdots, N$, draw $\lambda_i^{(s)}$ from a multivariate-normal distribution (or a normal distribution if λ is a scalar) $p\left(\lambda_i^{(s)} \left| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A \right.\right)$:

$$\begin{aligned}
\lambda_i^{(s)} &\sim N\left(m_i^\lambda, \Sigma_i^\lambda\right), \\
\Sigma_i^\lambda &= \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=1}^{T} w_{i,t-1} w_{i,t-1}'\right)^{-1}, \\
m_i^\lambda &= \Sigma_i^\lambda \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} \tilde{\mu}_i^\lambda + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=1}^{T} w_{i,t-1} \left(y_{it} - \beta^{(s-1)\prime} x_{i,t-1}\right)\right),
\end{aligned}$$

where the conditional "prior" mean is characterized by

$$\tilde{\mu}_i^\lambda = \begin{cases} \mu_{\gamma_i^\lambda}^{\lambda(s)}, & \text{for the random coefficients model,} \\ \mu_{\gamma_i^\lambda}^{\lambda(s)} \left[1, c_{i0}'\right]', & \text{for the correlated random coefficients model.} \end{cases}$$

Step 5: Common parameters

Cross-sectional Homoskedasticity Draw $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ from a linear regression model with "unknown" variance, $p\left(\beta^{(s)}, \sigma^{2(s)} \left| \left\{\lambda_i^{(s)}\right\}, D \right.\right)$:

$$\begin{aligned}
\left(\beta^{(s)}, \sigma^{2(s)}\right) &\sim N\left(\beta^{(s)};\ m^\beta, \psi^\beta \sigma^{2(s)}\right) \mathrm{IG}\left(\sigma^{2(s)};\ a^{\sigma^2}, b^{\sigma^2}\right), \\
\psi^\beta &= \left(\left(\psi_0^\beta\right)^{-1} + \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t-1} x_{i,t-1}'\right)^{-1}, \\
m^\beta &= \psi^\beta \left(\left(\psi_0^\beta\right)^{-1} m_0^\beta + \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t-1} \left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right), \\
a^{\sigma^2} &= a_0^{\sigma^2} + \frac{NT}{2} \\
b^{\sigma^2} &= b_0^{\sigma^2} + \frac{1}{2}\left(\sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)^2 + m_0^{\beta\prime} \left(\psi_0^\beta\right)^{-1} m_0^\beta - m^{\beta\prime} \left(\psi^\beta\right)^{-1} m^\beta\right).
\end{aligned}$$

Cross-sectional Heteroskedasticity Draw $\beta^{(s)}$ from a linear regression model with "known" variance, $p\left(\beta^{(s)} \left| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right.\right)$:

$$\begin{aligned}
\beta^{(s)} &\sim N\left(m^{\beta}, \Sigma^{\beta}\right), \\
\Sigma^{\beta} &= \left(\left(\Sigma_0^{\beta}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} x_{i,t-1} x_{i,t-1}'\right)^{-1}, \\
m^{\beta} &= \Sigma^{\beta}\left(\left(\Sigma_0^{\beta}\right)^{-1} m_0^{\beta} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right).
\end{aligned}$$

Remark B.2.3. For unbalanced panels, the summations and products in steps 4 and 5 (Subsections B.2.3 and B.2.3) are instead over $t = t_{0i}, \cdots, t_{1i}$, the observed periods for individual i.

B.2.4 Slice-Retrospective Samplers

The next algorithm borrows the idea from some recent development in DPM sampling strategies (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015), which integrates the slice sampler (Walker, 2007; Kalli *et al.*, 2011) and the retrospective sampler (Papaspiliopoulos and Roberts, 2008). By adding extra auxiliary variables, the sampler is able to avoid hard truncation in Ishwaran and James (2001, 2002). I experiment with it to check whether the approximation error due to truncation would significantly affect the density forecasts or not, and the results do not change much. The following algorithm is designed for the random coefficient case. A corresponding version for the correlated random coefficient case can be constructed in a similar manner.

The auxiliary variables u_i^z, $i = 1, \cdots, N$, are i.i.d. standard uniform random variables, i.e. $u_i^z \sim U(0,1)$. Then, the mixture of components in equation (3.2.6) can be rewritten as

$$z \sim \sum_{k^z=1}^{\infty} \mathbf{1}\left(u_i^z < p_{ik^z}^z\right) f^z\left(z;\, \theta_{k^z}^z\right),$$

where $z = \lambda,\ l$. By marginalizing over u_i^z, we can recover equation (3.2.6). Accordingly, we

can define the number of active components as

$$K^{z,A} = \max_{1 \le i \le N} \gamma_i^z,$$

and the number of potential components (including active components) as

$$K^{z,P} = \min \left\{ k^z : \left(1 - \sum_{j=1}^{k^z} p_j^z \right) < \min_{1 \le i \le N} u_i^z \right\}.$$

Although the number of components is infinite literally, we only need to care about the components that can potentially be occupied. Therefore, $K^{z,P}$ serves as an upper limit on the number of components that need to be updated at certain iteration. Here I suppress the iteration indicator s for exposition simplicity, but note that both $K^{z,A}$ and $K^{z,P}$ can change over iterations; this is indeed the highlight of this sampler.

Algorithm B.2.4. *(General Model: Random Coefficients III (Slice-Retrospective))*

For each iteration $s = 1, \cdots, n_{sim}$, steps 1-3 in Algorithm 3.5.1 are modified as follows:

For $z = \lambda,\ l$,

1. *Active components:*
   (a) *Number of active components:*

$$K^{z,A} = \max_{1 \le i \le N} \gamma_i^{z(s-1)}.$$

   (b) *Component probabilities: for $k^z = 1, \cdots, K^{z,A}$, draw $p_{k^z}^{z*}$ from the stick breaking process $p\left(\{p_{k^z}^{z*}\} \left| \alpha^{z(s-1)}, \left\{ n_{k^z}^{z(s-1)} \right\} \right. \right)$:*

$$p_{k^z}^{z*} \sim SB\left(n_{k^z}^{z(s-1)},\ \alpha^{z(s-1)} + \sum_{j=k^z+1}^{K^{z,A}} n_j^{z(s-1)} \right), \quad k^z = 1, \cdots, K^{z,A}.$$

   (c) *Component parameters: for $k^z = 1, \cdots, K^{z,A}$, draw $\theta_{k^z}^{z*}$ from*

$p\left(\theta_{k^z}^{z*} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$ *as in Algorithm 3.3.1 step 2.*

(d) *Label switching:*

jointly update $\left\{p_{k^z}^{z(s)}, \theta_{k^z}^{z(s)}, \gamma_i^{z*}\right\}_{k^z=1}^{K^{z,A}}$ *based on* $\left\{p_{k^z}^{z*}, \theta_{k^z}^{z*}, \gamma_i^{z(s-1)}\right\}_{k^z=1}^{K^{z,A}}$ *by three Metropolis-Hastings label-switching moves:*

i. *randomly select two non-empty components, switch their component labels* (γ_i^z)*, while leaving component parameters* ($\theta_{k^z}^z$) *and component probabilities* ($p_{k^z}^z$) *unchanged;*

ii. *randomly select two adjacent components, switch their component labels* (γ_i^z) *and component "stick lengths"* ($\zeta_{k^z}^z$)*, while leaving component parameters* ($\theta_{k^z}^z$) *unchanged;*

iii. *randomly select two non-empty components, switch their component labels* (γ_i^z) *and component parameters* ($\theta_{k^z}^z$)*, as well as update their component probabilities* ($p_{k^z}^z$).

Then, adjust $K^{z,A}$ *accordingly.*

2. *Auxiliary variables: for* $i = 1, \cdots, N$*, draw* $u_i^{z(s)}$ *from a uniform distribution* $p\left(u_i^{z(s)} \middle| \left\{p_{k^z}^{z(s)}\right\}, \gamma_i^{z*}\right)$*:*

$$u_i^{z(s)} \sim U\left(0,\ p_{\gamma_i^{z*}}^{z(s)}\right).$$

3. *DP scale parameter:*

(a) *Draw the latent variable* $\xi^{z(s)}$ *from a beta distribution* $p\left(\xi^{z(s)} \middle| \alpha^{z(s-1)}, N\right)$*:*

$$\xi^{z(s)} \sim Beta\left(\alpha^{z(s-1)} + 1, N\right).$$

(b) *Draw* $\alpha^{z(s)}$ *from a mixture of two gamma distributions* $p\left(\alpha^{z(s)} \middle| \xi^{z(s)}, K^{z,A}, N\right)$ *:*

$$\begin{aligned}
\alpha^{z(s)} \sim\ & p^{\alpha^z}\, Ga\left(\alpha^{z(s)};\ a^{\alpha^z} + K^{z,A},\ b^{\alpha^z} - \log \xi^{z(s)}\right) \\
& + \left(1 - p^{\alpha^z}\right) Ga\left(\alpha^{z(s)};\ a^{\alpha^z} + K^{z,A} - 1,\ b^{\alpha^z} - \log \xi^{z(s)}\right), \\
p^{\alpha^z} =\ & \frac{a^{\alpha^z} + K^{z,A} - 1}{N\left(b^{\alpha^z} - \log \xi^{z(s)}\right)}.
\end{aligned}$$

4. *Potential components:*

   (a) *Component probabilities: start with* $K^{z*} = K^{z,A}$,

      i. *if* $\left(1 - \sum_{j=1}^{K^{z*}} p_j^{z(s)}\right) < \min_{1 \le i \le N} u_i^{z(s)}$, *set* $K^{z,P} = K^{z*}$ *and stop;*

      ii. *otherwise, let* $K^{z*} = K^{z*} + 1$, *draw* $\zeta_{K^{z*}}^z \sim Beta\left(1, \alpha^{z(s)}\right)$, *update* $p_{K^{z*}}^{z(s)} = \zeta_{K^{z*}}^z \prod_{j<K^{z*}} \left(1 - \zeta_j^z\right)$, *and go to step (a-i).*

   (b) *Component parameters: for* $k^z = K^{z,A} + 1, \cdots, K^{z,P}$, *draw* $\theta_{k^z}^{z(s)}$ *from the DP base distribution* G_0^z.

5. *Component memberships: For* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \middle| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\}, \ u_i^{z(s)}, z_i^{(s-1)}\right)$:

$$\gamma_i^{z(s)} = k^z, \text{ with probability } p_{ik^z}^z, \ k^z = 1, \cdots, K^{z,P},$$

$$p_{ik^z}^z \propto p_{k^z}^{z(s)} \phi\left(z_i^{(s-1)}; \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right) \mathbf{1}\left(u_i^{z(s)} < p_{k^z}^{z(s)}\right), \quad \sum_{k^z=1}^{K^{z,P}} p_{ik^z}^z = 1.$$

The remaining part of the algorithm resembles steps 4 and 5 in Algorithm 3.5.1.

Remark B.2.5. Note that:

(i) Steps 1-b,c,d are sampling from "marginal" posterior of $(p_{k^z}^z, \theta_{k^z}^z, \gamma_i^z)$ for the active components with the auxiliary variables u_i^z's being integrated out. Thus, extra caution is needed in dealing with the order of the steps.

(ii) The label switching moves 1-d-i and 1-d-ii are based on Papaspiliopoulos and Roberts (2008), and 1-d-iii is suggested by Hastie *et al.* (2015). All these label switching moves aim to improve numerical convergence.

(iii) Step 3 for DP scale parameter α^z follows Escobar and West (1995). It is different from step 1-a in Algorithm 3.5.1 due to the unrestricted number of components in the current sampler.

(iv) Steps 4-a-ii and 4-b that update potential components are very similar to steps 1-b and 1-c that update active components—just take $J_{k^z}^z$ as an empty set and draw directly from

the prior.

(v) The auxiliary variable u_i^z also appears in step 5 that updates component memberships. The inclusion of auxiliary variables helps determine a finite set of relevant components for each individual i without mechanically truncating the infinite mixture.

B.3 Proofs for Baseline Model

B.3.1 Posterior Consistency: Random Effects Model

Skills vs Shocks

Proof. **(Proposition 3.4.7)**

Based on the Schwartz (1965) theorem stated in Lemma 3.4.6, two sufficient conditions guarantee the posterior consistency: KL requirement and uniformly exponentially consistent tests.

(i) KL requirement

The proposition assumes that the KL property holds for the distribution of λ, i.e. for all $\epsilon > 0$,

$$\Pi^f \left(f \in \mathcal{F} : \int f_0(\lambda) \log \frac{f_0(\lambda)}{f(\lambda)} d\lambda < \epsilon \right) > 0,$$

whose sufficient conditions are stated in Lemmas 3.4.8 and B.5.1. On the other hand, the KL requirement is specified on the observed y in order to guarantee that the denominator in equation (3.4.2) is large enough. In this sense, we need to establish that for all $\epsilon > 0$,

$$\Pi \left(f \in \mathcal{F} : \int f_0(y-u)\phi(u) \log \frac{\int f_0(y-u')\phi(u')du'}{\int f(y-u')\phi(u')du'} dudy < \epsilon \right) > 0.$$

Let $g(x) = x \log x$, $a(u) = f_0(y-u)\phi(u)$, $A = \int a(u)du$, $b(u) = f(y-u)\phi(u)$, $B =$

$\int b(u)\,du$. We can rewrite the integral over u as

$$\begin{aligned}
&\int f_0(y-u)\,\phi(u)\log\frac{\int f_0(y-u')\,\phi(u')\,du'}{\int f(y-u')\,\phi(u')\,du'}du = A\cdot\log\frac{A}{B} = B\cdot g\left(\frac{A}{B}\right)\\
=&B\cdot g\left(\int\frac{b(u)}{B}\cdot\frac{f_0(y-u)}{f(y-u)}du\right)\leq\int b(u)\,g\left(\frac{f_0(y-u)}{f(y-u)}\right)du\\
=&\int\phi(u)\,f_0(y-u)\log\frac{f_0(y-u)}{f(y-u)}du, \qquad \text{(B.3.1)}
\end{aligned}$$

where the inequality is given by Jensen's inequality. Then, further integrating the above expression over y, we have

$$\begin{aligned}
&\int f_0(y-u)\,\phi(u)\log\frac{\int f_0(y-u')\,\phi(u')\,du'}{\int f(y-u')\,\phi(u')\,du'}dudy\leq\int\phi(u)\,f_0(y-u)\log\frac{f_0(y-u)}{f(y-u)}dudy\\
=&\int\phi(u)\,du\cdot\int f_0(\lambda)\log\frac{f_0(\lambda)}{f(\lambda)}d\lambda=\epsilon
\end{aligned}$$

The inequality follows the above expression (B.3.1), the next equality is given by change of variables, and the last equality is given by the KL property of the distribution of λ.

(ii) Uniformly exponentially consistent tests

(ii-a) When λ is observed

Note that by the Hoeffding's inequality, the uniformly exponentially consistent tests are equivalent to strictly unbiased tests, so we only need to construct a test function $\varphi^\star$ such that

$$\mathbb{E}_{f_0}(\varphi^\star)<\inf_{f\in U^c}\mathbb{E}_f(\varphi^\star).$$

Without loss of generality, let us consider a weak neighborhood defined on $\epsilon>0$ and a bounded continuous function φ ranging from 0 to 1. Then, the corresponding neighborhood is given by

$$U_{\epsilon,\varphi}(f_0)=\left\{f:\ \left|\int\varphi f-\int\varphi f_0\right|<\epsilon\right\}.$$

We can divide the alternative region into two parts[37]

$$U^c_{\epsilon,\varphi}(f_0) = A_1 \cup A_2$$

where

$$A_1 = \left\{ f : \int \varphi f > \int \varphi f_0 + \epsilon \right\},$$
$$A_2 = \left\{ f : \int \varphi f < \int \varphi f_0 - \epsilon \right\}.$$

For A_1, we can choose the test function $\varphi^\star$ to be φ. For A_2, we can choose $\varphi^\star$ to be $1-\varphi$. Then, in either case $A = A_1,\ A_2$, type I error $\mathbb{E}_{f_0}(\varphi^\star) = \int \varphi^\star f_0$, and power $\inf_{f\in A} \mathbb{E}_f(\varphi^\star) \geq \int \varphi^\star f_0 + \epsilon$, hence the tests exist when λ is observed.

(ii-b) When y is observed instead of λ

Define $g(\lambda) = f(\lambda) - f_0(\lambda)$. Then, by definition, $\int g(\lambda)\, d\lambda = 0$ for all g. There are always tests if we observe λ, then for any g, there exists a $\epsilon > 0$ such that

$$\int |g(\lambda)|\, d\lambda > \epsilon. \tag{B.3.2}$$

The next step is to prove that there are tests when y is observed instead of λ, which is done via proof by contradiction. Suppose there is no test when we only observe y, then there exists a $\tilde{g}$ such that

$$\tilde{h}(y) = \int \tilde{g}(y-u)\,\phi(u)\, du = 0 \text{ for all } y,$$

due to the continuity of $\tilde{h}$. Employing the Fourier transform, we have

$$F_y(\xi) = F_\lambda(\xi) \cdot c_1 \exp\left(-c_2\xi^2\right) = 0 \text{ for all } \xi.$$

[37] It is legitimate to divide the alternatives into sub-regions. Intuitively, with different alternative sub-regions, the numerator in equation (3.4.2) is composed of integrals over different domains, and all of them converge to 0.

Since $c_1 \exp\left(-c_2 \xi^2\right) \neq 0$, then

$$F_\lambda(\xi) = 0 \text{ for all } \xi.$$

Finally, the inverse Fourier transform leads to

$$\tilde{g}(\lambda) = 0 \text{ for all } \lambda,$$

which contradicts equation (B.3.2). Therefore, there are also tests when y is observed instead of λ.

Combining (i) and (ii-b), f achieves posterior consistency even when we only observe y. $\square$

Unknown Shocks Sizes

Proof. **(Proposition 3.4.9)**

(i) KL requirement

Based on the observed sufficient statistics $\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$, the KL requirement can be written as follows: for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ \sigma^2 \in \mathbb{R}^+ : \\ \int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda} < \epsilon \end{array}\right) > 0.$$

Under the prior specification together with hyperparameters specified in Appendix B.2.1, the integral is bounded with probability one. Following the dominated convergence theorem,

$$\begin{aligned} &\lim_{\sigma^2 \to \sigma_0^2} \int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda} \\ =&\int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda}, \end{aligned}$$

where the upper bound of the right hand side can be characterized by the KL property of the distribution of λ as in the proof of Proposition 3.4.7 part (i). The sufficient conditions of the KL property of the distribution of λ are stated in Lemmas 3.4.8 and B.5.1.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $\left|\sigma^2 - \sigma_0^2\right| > \Delta$

Orthogonal forward differencing yields $\tilde{y}_{it} \sim N\left(0, \sigma_0^2\right)$. Then, as $N \to \infty$,

$$\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\tilde{y}_{it}\right)^2}{\sigma_0^2} \sim \chi^2_{N(T-1)} \xrightarrow{d} N\left(1, \frac{2}{N\left(T-1\right)}\right).$$

Note that for a generic variable $x \sim N\left(0,1\right)$, for $x^* > 0$,

$$\mathbb{P}\left(x > x^*\right) \leq \frac{\phi\left(x^*\right)}{x^*}. \tag{B.3.3}$$

Then, we can directly construct the following test function

$$\boldsymbol{\varphi_N}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} (\tilde{y}_{it})^2}{\sigma_0^2} < 1 - \frac{\Delta}{2\sigma_0^2}\right), & for\ \sigma^2 < \sigma_0^2 - \Delta, \\ \mathbf{1}\left(\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} (\tilde{y}_{it})^2}{\sigma_0^2} > 1 + \frac{\Delta}{2\sigma_0^2}\right), & for\ \sigma^2 > \sigma_0^2 + \Delta, \end{cases}$$

which satisfies the requirements (3.4.1) for the uniformly exponentially consistent tests.

(ii-b) $\left|\sigma^2 - \sigma_0^2\right| < \Delta,\ f \in U^c_{\epsilon,\Phi}\left(f_0\right)$

Without loss of generality, let $\Phi = \{\varphi\}$ be a singleton and $\varphi^\star$ be the test function that distinguishes $f = f_0$ versus $f \in U^c_{\epsilon,\varphi}\left(f_0\right)$ when σ_0^2 is known. Then, we can express the

difference between $\mathbb{E}_f(\varphi^\star)$ and $\mathbb{E}_{f_0}(\varphi^\star)$ as

$$\begin{aligned}
&\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)d\hat{u}d\hat{\lambda}-\int \varphi^\star\left(\hat{\lambda}\right) f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda}\\
>&\int \varphi^\star\left(\hat{\lambda}\right)\left(f\left(\hat{\lambda}-\hat{u}\right)-f_0\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda}\\
&-\left|\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left(\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right)d\hat{u}d\hat{\lambda}\right|. \qquad \text{(B.3.4)}
\end{aligned}$$

Since $\varphi^\star$ is the test function when σ_0^2 is known, the first term

$$\int \varphi^\star\left(\hat{\lambda}\right)\left(f\left(\hat{\lambda}-\hat{u}\right)-f_0\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda}>\epsilon. \qquad \text{(B.3.5)}$$

For the second term,

$$\begin{aligned}
&\left|\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left(\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right)d\hat{u}d\hat{\lambda}\right|\\
\leq&\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left|\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right|d\hat{u}d\hat{\lambda}\\
\leq&\int \left|\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right|d\hat{u}\\
\leq&\sqrt{\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}}. \qquad \text{(B.3.6)}
\end{aligned}$$

The second inequality is given by the fact that $\varphi^\star\left(\hat{\lambda}\right)\in[0,1]$. The last inequality follows Pinsker's inequality that bounds the total variation distance by the KL divergence, which has an explicit form for normal distributions

$$d_{KL}\left(\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right),\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)\right)=\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}\right).$$

We can choose $\Delta>0$ such that for any $\left|\sigma^2-\sigma_0^2\right|<\Delta$,

$$\sqrt{\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}}<\frac{\epsilon}{2}.$$

Plugging expressions (B.3.5) and (B.3.6) into (B.3.4), we obtain

$$\int \varphi^{\star}\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right) \phi\left(\hat{u};\, 0, \frac{\sigma^2}{T}\right) d\hat{u}d\hat{\lambda} - \int \varphi^{\star}\left(\hat{\lambda}\right) f_0\left(\hat{\lambda}-\hat{u}\right) \phi\left(\hat{u};\, 0, \frac{\sigma_0^2}{T}\right) d\hat{u}d\hat{\lambda}$$
$$> \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2},$$

so $\varphi^{\star}$ is the test function with respect to the alternative sub-region $\{\left|\sigma^2-\sigma_0^2\right| < \Delta,\ f \in U^c_{\epsilon,\Phi}(f_0)\}$. □

Lagged Dependent Variables

Proof. **(Proposition 3.4.11)**

(i) KL requirement

Define the sufficient statistics $\hat{\lambda}(\beta) = \frac{1}{T}\sum_{t=1}^{T} y_{it} - \beta y_{i,t-1}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. The KL requirement is satisfied as long as for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ \left(\beta, \sigma^2\right) \in \mathbb{R}\times\mathbb{R}^{+}: \\ \int f_0\left(\hat{\lambda}(\beta_0)-\hat{u}\right)\phi\left(\hat{u};\, 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}(\beta_0)-\hat{u}'\right)\phi\left(\hat{u}';\, 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}(\beta)-\hat{u}'\right)\phi\left(\hat{u}';\, 0, \frac{\sigma^2}{T}\right)d\hat{u}'} d\hat{u}d\hat{\lambda} < \epsilon \end{array}\right) > 0.$$

Similar to the previous case, the dominated convergence theorem and the KL property of the distribution of λ complete the proof.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $|\beta-\beta_0| > \Delta$ or $\left|\sigma^2-\sigma_0^2\right| > \Delta'$

Orthogonal forward differencing yields $\tilde{y}_{it} = \beta \tilde{y}_{i,t-1} + \tilde{u}_{it},\ \tilde{u}_{it} \sim N\left(0, \sigma_0^2\right)$. Then, as $N \to \infty$,

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\tilde{y}_{i,t-1}\right)^2 \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \tilde{y}_{i,t-1} \tilde{y}_{it} \right) \xrightarrow{d} N\left(\beta_0, \frac{\sigma_0^2}{N \sum_{t=1}^{T-1} \mathbb{E}\left(\tilde{y}_{i,t-1}\right)^2} \right)$$

$$\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\tilde{y}_{it} - \hat{\beta}_{OLS} \tilde{y}_{i,t-1} \right)^2}{\sigma_0^2} \sim \chi^2_{N(T-1)-1} \xrightarrow{d} N\left(1, \frac{2}{N(T-1)-1}\right).$$

Since the upper tail of a normal distribution is bounded as in expression (B.3.3), we can directly construct the following test function

$$\boldsymbol{\varphi_N} = 1 - \left(1 - \boldsymbol{\varphi_{N,\beta}}\right)\left(1 - \boldsymbol{\varphi_{N,\sigma^2}}\right),$$

where

$$\boldsymbol{\varphi_{N,\beta}}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\hat{\beta}_{OLS} < \beta_0 - \frac{\Delta}{2}\right), & for\ \beta < \beta_0 - \Delta, \\ \mathbf{1}\left(\hat{\beta}_{OLS} > \beta_0 + \frac{\Delta}{2}\right), & for\ \beta > \beta_0 + \Delta, \end{cases}$$

$$\boldsymbol{\varphi_{N,\sigma^2}}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\tilde{y}_{it} - \hat{\beta}_{OLS} \tilde{y}_{i,t-1} \right)^2}{\sigma_0^2} < 1 - \frac{\Delta'}{2\sigma_0^2} \right), & for\ \sigma^2 < \sigma_0^2 - \Delta', \\ \mathbf{1}\left(\frac{\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\tilde{y}_{it} - \hat{\beta}_{OLS} \tilde{y}_{i,t-1} \right)^2}{\sigma_0^2} > 1 + \frac{\Delta'}{2\sigma_0^2} \right), & for\ \sigma^2 > \sigma_0^2 + \Delta', \end{cases}$$

which satisfies the requirements (3.4.1) for the uniformly exponentially consistent tests.

(ii-b) $|\beta - \beta_0| < \Delta,\ \left|\sigma^2 - \sigma_0^2\right| < \Delta',\ f \in U^c_{\epsilon,\Phi}\left(f_0\right)$

The following proof is analogous to the proofs of Proposition 3.3 in Amewou-Atisso *et al.* (2003) except the inclusion of shocks u_{it}'s in the current setup, which prohibits direct inference of λ_i. Without loss of generality, let $\Phi = \{\varphi\}$ and $\varphi^\star\left(\mathring{y}\right)$ be the corresponding test function on $\mathring{y} = y_{i1} - \beta_0 y_{i0} = \lambda_i + u_{i1}$ when β_0 and σ_0^2 are known. Then, we can construct

a uniformly continuous test function

$$\varphi^{\star\star}\left(\mathring{y}\right)=\begin{cases}\varphi^{\star}\left(\mathring{y}\right), & if\ \ |\mathring{y}|<M_1,\\ 1, & if\ \ |\mathring{y}|>M_2,\\ \max\left\{\varphi^{\star}\left(\mathring{y}\right),\ \varphi^{\star}\left(M_1\right)+\frac{1-\varphi^{\star}(M_1)}{M_2-M_1}\left(\mathring{y}-M_1\right)\right\}, & if\ \mathring{y}\in[M_1,M_2],\\ \max\left\{\varphi^{\star}\left(\mathring{y}\right),\ 1+\frac{\varphi^{\star}(-M_1)-1}{M_2-M_1}\left(\mathring{y}+M_2\right)\right\} & if\ \mathring{y}\in[-M_2,-M_1],\end{cases}$$

where M_1 is chosen such that

$$\int_{|\mathring{y}|>M_1} f_0\left(\mathring{y}-u\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1<\frac{\epsilon}{4}.$$

Then,

$$\int\varphi^{\star\star}\left(\mathring{y}\right)f\left(\mathring{y}-u\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1-\int\varphi^{\star\star}\left(\mathring{y}\right)f_0\left(\mathring{y}-u\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1>\frac{3}{4}\epsilon. \tag{B.3.7}$$

Due to uniform continuity, given $\epsilon>0$, there exists $\delta>0$ such that $|\varphi^{\star\star}\left(\mathring{y}'\right)-\varphi^{\star\star}\left(\mathring{y}\right)|<\epsilon/4$ for any $|\mathring{y}'-\mathring{y}|<\delta$. As y_{i0} is compacted supported, we can choose Δ such that $|\left(\beta-\beta_0\right)y_{i0}|<\delta$.

Let y_1 be a generic variable representing y_{i1}. Define the test function for the non-i.i.d. case to be $\varphi_i\left(y_1\right)=\varphi^{\star\star}\left(y_1-\beta_0 y_{i0}\right)$. Then, the difference between $\mathbb{E}_f\left(\varphi_i\right)$ and $\mathbb{E}_{f_0}\left(\varphi_i\right)$ is

$$\begin{aligned}&\int\varphi_i\left(y_1\right)f\left(y_1-\beta y_{i0}-u\right)\phi\left(u;\ 0,\sigma^2\right)dudy_1\\&-\int\varphi_i\left(y_1\right)f_0\left(y_1-\beta_0 y_{i0}-u\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1\\>&\int\varphi_i\left(y_1\right)\left(f\left(y_1-\beta_0 y_{i0}-u\right)-f_0\left(y_1-\beta_0 y_{i0}-u\right)\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1\\&+\int\varphi_i\left(y_1\right)\left(f\left(y_1-\beta y_{i0}-u\right)-f\left(y_1-\beta_0 y_{i0}-u\right)\right)\phi\left(u;\ 0,\sigma_0^2\right)dudy_1\\&-\left|\int\varphi_i\left(y_1\right)f\left(y_1-\beta y_{i0}-u\right)\left(\phi\left(u;\ 0,\sigma^2\right)-\phi\left(u;\ 0,\sigma_0^2\right)\right)dudy_1\right|.\end{aligned}$$

From expression (B.3.7), the first term is bounded below by $3\epsilon/4$. Similar to the proof of Proposition 3.4.9 part (ii-b), the third term is bounded above by $\epsilon/4$. For the second term, note that for any δ,

$$\int \varphi^{\star\star}(y_1 - \delta) f(y_1 - \delta - u)\, dy_1 = \int \varphi^{\star\star}(y_1) f(y_1 - u)\, dy_1$$

Then,

$$\begin{aligned}
&\int \varphi_i(y_1) \left(f(y_1 - \beta y_{i0} - u) - f(y_1 - \beta_0 y_{i0} - u)\right) dy_1 \\
=&\int \varphi^{\star\star}(y_1 + (\beta - \beta_0) y_{i0}) f(y_1 - u)\, dy_1 - \int \varphi^{\star\star}(y_1) f(y_1 - u)\, dy_1 \\
\geq& -\int \left|\varphi^{\star\star}(y_1 + (\beta - \beta_0) y_{i0}) - \varphi^{\star\star}(y_1)\right| f(y_1 - u)\, dy_1 \\
\geq& -\frac{\epsilon}{4}
\end{aligned}$$

where the last inequality is given by the uniform continuity of $\varphi^{\star\star}$. Hence, $\mathbb{E}_f(\varphi_i) - \mathbb{E}_{f_0}(\varphi_i) > \epsilon/4$, and $\{\varphi_i\}$ constitutes the tests with respect to the alternative sub-region $\left\{|\beta - \beta_0| < \Delta,\ \left|\sigma^2 - \sigma_0^2\right| < \Delta',\ f \in U^c_{\epsilon,\Phi}(f_0)\right\}$. □

B.3.2 Posterior Consistency: Correlated Random Effects Model

Recall that h, f, and q are the joint, conditional, and marginal densities, respectively. In addition,

$$h_0(\lambda, c) = f_0(\lambda|c) \cdot q_0(c), \quad h(\lambda, c) = f(\lambda|c) \cdot q_0(c).$$

Proof. **(Proposition 3.4.15)**

(i) KL requirement

Define the sufficient statistics $\hat{\lambda}(\beta) = \frac{1}{T}\sum_{t=1}^{T} y_{it} - \beta y_{i,t-1}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Considering joint density characterization, the observations are i.i.d. across i in the correlated random effects setup. The KL requirement can be specified as follows: for all

$\epsilon > 0$,

$$\Pi \left(\begin{array}{l} f \in \mathcal{F},\ (\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : \\ \int h_0 \left(\hat{\lambda} (\beta_0) - \hat{u},\ y_0 \right) \phi \left(\hat{u};\ 0, \frac{\sigma_0^2}{T} \right) \\ \quad \cdot \log \frac{\int h_0 \left(\hat{\lambda} (\beta_0) - \hat{u}',\ y_0 \right) \phi \left(\hat{u}';\ 0, \frac{\sigma_0^2}{T} \right) d\hat{u}'}{\int h \left(\hat{\lambda} (\beta) - \hat{u}',\ y_0 \right) \phi \left(\hat{u}';\ 0, \frac{\sigma^2}{T} \right) d\hat{u}'} d\hat{u} d\hat{\lambda} dy_0 < \epsilon \end{array} \right) > 0.$$

The rest of the proof is similar to the previous cases employing the dominated convergence theorem and the KL property of the joint distribution of (λ, y_0) with sufficient conditions stated in Assumption 3.4.14.

(ii) Uniformly exponentially consistent tests

It follows the proof of Proposition 3.4.11 part (ii) except that in case $|\beta - \beta_0| < \Delta,\ \left|\sigma^2 - \sigma_0^2\right| < \Delta',\ f \in U^c_{\epsilon,\Phi} (f_0)$, the test function φ is defined on (y_1, y_0) that distinguishes the true h_0 from alternative h. □

B.3.3 Density Forecasts

Proof. **(Proposition 3.4.16)**

(i) Random Effects: Result 1

In this part, I am going to prove that for any i and any $U_{\epsilon,\Phi} \left(f^{oracle}_{i,T+1} \right)$, as $N \to \infty$,

$$\mathbb{P} \left(f^{cond}_{i,T+1} \in U_{\epsilon,\Phi} \left(f^{oracle}_{i,T+1} \right) \middle| y_{1:N,0:T} \right) \to 1,\ a.s.$$

This is equivalent to proving that for any bounded continuous function φ,

$$\mathbb{P} \left(f \in \mathcal{F} : \left| \begin{array}{l} \int \varphi (y)\, f^{cond}_{i,T+1} \left(y | \beta, \sigma^2, f, y_{i,0:T} \right) dy \\ - \int \varphi (y)\, f^{oracle}_{i,T+1} (y)\, dy \end{array} \right| < \epsilon \middle| y_{1:N,0:T} \right) \to 1,\ a.s.$$

where

$$
\begin{aligned}
&\left|\int \varphi(y)\, f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)dy - \int \varphi(y)\, f_{i,T+1}^{oracle}(y)\,dy\right| \\
=&\left|\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right) p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right) d\lambda_i dy\right. \\
&\left.-\int \varphi(y)\,\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right) p\left(\lambda_i\left|\beta_0,\sigma_0^2,f_0,y_{i,0:T}\right.\right) d\lambda_i dy\right| \\
=&\left|\frac{\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)\,d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)\,d\lambda_i}\right. \\
&\left.-\frac{\int \varphi(y)\,\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0(\lambda_i)\,d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0(\lambda_i)\,d\lambda_i}\right|.
\end{aligned}
$$

The last equality is given by plugging in

$$
p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right)=\frac{\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)}{\int \prod_t p\left(y_{it}\left|\lambda_i',\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i')\,d\lambda_i'}.
$$

Set

$$
\begin{aligned}
A&=\int \prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) d\lambda_i, \\
B&=\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) d\lambda_i dy.
\end{aligned}
$$

with A_0 and B_0 being the counterparts for the oracle predictor. Then, we want to make sure the following expression is arbitrarily small,

$$
\left|\frac{B}{A}-\frac{B_0}{A_0}\right|\leq\frac{|B_0|\,|A-A_0|}{|A_0|\,|A|}+\frac{|B-B_0|}{|A|},
$$

and it is sufficient to establish the following four statements.

(a) $|A - A_0| < \epsilon'$

$$
\begin{aligned}
&|A - A_0| \\
\leq &\left| \int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \left(f\left(\lambda_i\right) - f_0\left(\lambda_i\right)\right) d\lambda_i \right| \\
&+ \left| \int \left(\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right) f_0\left(\lambda_i\right) d\lambda_i \right|
\end{aligned}
$$

The first term is less than $\epsilon'/2$ with probability one due to the posterior consistency of f and that

$$
\prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) = C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) \quad \text{(B.3.8)}
$$

is a bounded continuous function in λ_i, with $C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)$ being

$$
C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) = \frac{1}{\sqrt{T}\left(2\pi\sigma_0^2\right)^{\frac{T-1}{2}}} \exp\left(-\frac{\sum_t \left(y_{it} - \beta_0 y_{i,t-1}\right)^2 - \frac{1}{T}\left(\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right)\right)^2}{2\sigma_0^2} \right).
$$

For the second term,

$$
\begin{aligned}
&\left| \int \left(\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right) f_0\left(\lambda_i\right) d\lambda_i \right| \\
\leq &M \int \left| \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right| d\lambda_i \\
\leq &MC\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \int \left| \begin{array}{l} \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) \\ - \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) \end{array} \right| d\lambda_i \\
&+ M\left|C\left(\beta, \sigma^2, y_{i,0:T}\right) - C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)\right| \int \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) d\lambda_i.
\end{aligned} \quad \text{(B.3.9)}
$$

where the last inequality is given by rewriting $\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right)$ as a distribution of

λ_i (equation B.3.8). Following Pinsker's inequality that bounds the total variation distance by the KL divergence,

$$\begin{aligned}
&\int \left| \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) - \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right)\right| d\lambda_i \\
\leq &\sqrt{2 d_{KL}\left(\phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right),\ \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right)\right)} \\
\leq &\sqrt{\frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2} + \frac{(\beta-\beta_0)^2\left(\sum_t y_{i,t-1}\right)^2}{T\sigma^2}}. \qquad \text{(B.3.10)}
\end{aligned}$$

As $\left(\beta, \sigma^2\right)$ enjoy posterior consistency, both $\left|C\left(\beta, \sigma^2, y_{i,0:T}\right) - C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)\right|$ in expression (B.3.9) and $\sqrt{\frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2} + \frac{(\beta-\beta_0)^2\left(\sum_t y_{i,t-1}\right)^2}{T\sigma^2}}$ in expression (B.3.10) can be arbitrarily small. Therefore, the second term is less than $\epsilon'/2$ with probability one.

(b) $|B - B_0| < \epsilon'$

$$\begin{aligned}
&|B - B_0| \\
\leq &\left| \int \varphi(y)\, \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \left(f\left(\lambda_i\right) - f_0\left(\lambda_i\right)\right) d\lambda_i dy \right| \\
&+ \left| \int \varphi(y) \begin{pmatrix} \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) \\ \\ -\phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \end{pmatrix} f_0\left(\lambda_i\right) d\lambda_i dy \right|
\end{aligned}$$

Similar to (a), the first term is small due to the posterior consistency of f, while Pinsker's inequality together with the posterior consistency of $\left(\beta, \sigma^2\right)$ ensure a small second term.

(c) There exists $\underline{A} > 0$ such that $|A_0| > \underline{A}$.

$$\begin{aligned}
A_0 &= \int \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i \\
&= C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \int \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) f_0\left(\lambda_i\right) d\lambda_i
\end{aligned}$$

Since $\phi\left(\lambda_i;\ \frac{1}{T}\sum_T\left(y_{it}-\beta_0 y_{i,t-1}\right),\frac{\sigma_0^2}{T}\right)$ and $f_0\left(\lambda_i\right)$ share the same support on $\mathbb{R}$, the integral is bounded below by some positive $\underline{A}$. Moreover, we have $|A-A_0|<\epsilon'$ from (a), then $|A|>|A_0|-\epsilon'>\underline{A}-\epsilon'$. Therefore, both $|A_0|$ and $|A|$ are bounded below.

(d) $|B_0|<\infty$

$$\begin{aligned}
|B_0| &= \left|\int\varphi\left(y\right)\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right)f_0\left(\lambda_i\right)d\lambda_i dy\right| \\
&\leq M_\varphi\cdot\frac{1}{\left(2\pi\sigma_0^2\right)^{\frac{T}{2}}}\cdot\left|\int\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i dy\right| \\
&= M_\varphi\cdot\frac{1}{\left(2\pi\sigma_0^2\right)^{\frac{T}{2}}}
\end{aligned}$$

(ii) Random Effects: Result 2

Now the goal is to prove that for any i, any y, and any $\epsilon>0$, as $N\to\infty$,

$$\left|f_{i,T+1}^{sp}\left(y\right)-f_{i,T+1}^{oracle}\left(y\right)\right|<\epsilon,\ a.s.$$

where

$$
\begin{aligned}
&\left| f_{i,T+1}^{sp}(y) - f_{i,T+1}^{oracle}(y) \right| \\
=&\left| \int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) p\left(\lambda_i \left| \beta, \sigma^2, f, y_{i,0:T}\right.\right) d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\lambda_i d\beta d\sigma^2 df \right. \\
&\left. - \int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) p\left(\lambda_i \left| \beta_0, \sigma_0^2, f_0, y_{i,0:T}\right.\right) d\lambda_i \right| \\
=&\left| \int \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i} \right. \\
&\cdot d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\beta d\sigma^2 df \\
&\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i} \right| \\
\leq& \int \left| \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i} \right. \\
&\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i} \right| \\
&\cdot d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\beta d\sigma^2 df.
\end{aligned}
$$

Note that along the same lines as part (i) "Random Effects: Result 1", the integrand

$$
\left| \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f\left(\lambda_i\right) d\lambda_i} \right.
$$
$$
\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0\left(\lambda_i\right) d\lambda_i} \right| < \epsilon.
$$

(iii) Correlated Random Effects: Result 1

As the posterior consistency for conditional density estimation is characterized by the joint distribution over (λ_i, y_{i0}), the convergence of "joint" predictive distribution $(y_{i,T+1}, y_{i0})$ follows the same logic as part (i) "Random Effects: Result 1". Hence for any bounded contin-

uous function $\tilde{\varphi}(y, y_{i0})$, and any $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\begin{array}{l|l} f \in \mathcal{F},\ (\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : & \\ \left| \int \tilde{\varphi}(y, y_{i0}) f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) q_0(y_{i0}) dy_{i0} dy \right. & y_{1:N,0:T} \\ \left. - \int \tilde{\varphi}(y, y_{i0}) f_{i,T+1}^{oracle}(y|y_{i0}) q_0(y_{i0}) dy_{i0} dy \right| < \epsilon & \end{array}\right) \to 1,\ a.s.$$

where

$$\begin{aligned} & \left| \int \tilde{\varphi}(y, y_{i0}) f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) q_0(y_{i0}) dy_{i0} dy \right. \\ & \left. - \int \tilde{\varphi}(y, y_{i0}) f_{i,T+1}^{oracle}(y|y_{i0}) q_0(y_{i0}) dy_{i0} dy \right| \\ = & \left| \frac{\int \tilde{\varphi}(y, y_{i0}) \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i|y_{i0}) q_0(y_{i0}) d\lambda_i dy_{i0} dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i|y_{i0}) q_0(y_{i0}) d\lambda_i dy_{i0}} \right. \\ & \left. - \frac{\int \tilde{\varphi}(y, y_{i0}) \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0}) q_0(y_{i0}) d\lambda_i dy_{i0} dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0}) q_0(y_{i0}) d\lambda_i dy_{i0}} \right|. \end{aligned} \tag{B.3.11}$$

However, it is more desirable to establish the convergence of "conditional" predictive distribution $y_{i,T+1}|y_{i0}$, i.e. for any bounded continuous function on y, $\varphi(y)$ and any $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\begin{array}{l|l} f \in \mathcal{F},\ (\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : & \\ \left| \int \varphi(y) f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) dy \right. & y_{1:N,0:T} \\ \left. - \int \varphi(y) f_{i,T+1}^{oracle}(y|y_{i0}) dy \right| < \epsilon & \end{array}\right) \to 1,\ a.s.$$

where

$$\begin{aligned}
&\left| \int \varphi(y) f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right) dy - \int \varphi(y) f_{i,T+1}^{oracle}\left(y|y_{i0}\right) dy \right| \\
=&\left| \frac{\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f\left(\lambda_i|y_{i0}\right) d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f\left(\lambda_i|y_{i0}\right) d\lambda_i} \right. \\
&\left. - \frac{\int \varphi(y)\,\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0\left(\lambda_i|y_{i0}\right) d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0\left(\lambda_i|y_{i0}\right) d\lambda_i} \right|. \qquad \text{(B.3.12)}
\end{aligned}$$

Set $\tilde{\varphi}(y,y_{i0}) = \frac{\varphi(y)}{q_0(y_{i0})}$. Note that $q_0(y_{i0})$ is continuous and bounded below due to condition 2-b in Proposition 3.4.16, so $\tilde{\varphi}(y,y_{i0})$ is a bounded continuous continuous function. Then, the right hand side of equation (B.3.11) coincides with the right hand side of equation (B.3.12), so we achieve the convergence of "conditional" predictive distribution $y_{i,T+1}|y_{i0}$.

(iv) Correlated Random Effects: Result 2

Combining (ii) and (iii) completes the proof. □

B.4 Proofs for General Model

B.4.1 Identification

Proof. **(Proposition 3.5.6)**

Part (iii) follows Liu *et al.* (2016), which is based on the early work by Arellano and Bonhomme (2012b). Part (ii) for cross-sectional heteroskedasticity is new.

(i) The identification of common parameters β is given by Assumption 3.5.5 (1).

(ii) Identify the distribution of shock sizes f^{σ^2}

First, let us perform orthogonal forward differencing, i.e. for $t = 1, \cdots, T - d_w$,

$$\begin{aligned}
\tilde{y}_{it} &= y_{it} - w'_{i,t-1} \left(\sum_{s=t+1}^{T} w_{i,s-1} w'_{i,s-1} \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} y_{is}, \\
\tilde{x}_{i,t-1} &= x_{i,t-1} - w'_{i,t-1} \left(\sum_{s=t+1}^{T} w_{i,s-1} w'_{i,s-1} \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} x_{i,s-1}.
\end{aligned}$$

Then, define

$$\begin{aligned}
\tilde{u}_{it} &= \tilde{y}_{it} - \beta' \tilde{x}_{i,t-1}, \\
\hat{\sigma}_i^2 &= \sum_{t=1}^{T-d_w} \tilde{u}_{it}^2 = \sigma_i^2 \chi_i^2.
\end{aligned}$$

where $\chi_i^2 \sim \chi^2 (T - d_w)$ follows an i.i.d. chi-squared distribution with $(T - d_w)$ degrees of freedom.

Note that Fourier transformation (i.e. characteristic functions) is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004). For a generic variable x, the Mellin transform of $f(x)$ is specified as

$$M_x (\xi) = \int x^{i\xi} f(x) \, dx,$$

which exists for all ξ.

Considering that $\sigma_i^2 | c$ and χ_i^2 are independent, we have

$$M_{\hat{\sigma}^2} (\xi | c) = M_{\sigma^2} (\xi | c) \, M_{\chi^2} (\xi).$$

Note that the non-vanishing characteristic function of σ^2 implies non-vanishing Mellin transform $M_{\sigma^2} (\xi | c)$ (almost everywhere), so it is legitimate to take the logarithm of both sides,

$$\log M_{\hat{\sigma}^2} (\xi | c) = \log M_{\sigma^2} (\xi | c) + \log M_{\chi^2} (\xi).$$

Taking the second derivative with respect to ξ, we get

$$\frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\sigma^2}(\xi|c) = \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\hat{\sigma}^2}(\xi|c) - \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\chi^2}(\xi).$$

The Mellin transform of chi-squared distribution $M_{\chi^2}(\xi)$ is a known functional form. In addition, we have

$$\begin{aligned}
\log M_{\sigma^2}(0|c) &= \log M_{\hat{\sigma}^2}(0|c) - \log M_{\chi^2}(0) = 0, \\
\frac{\partial}{\partial\xi}\log M_{\sigma^2}(0|c) &= \frac{\partial}{\partial\xi}\log M_{\hat{\sigma}^2}(0|c) - \frac{\partial}{\partial\xi}\log M_{\chi^2}(0) \\
&= i\left(\mathbb{E}\left(\log\hat{\sigma}^2\middle|c\right) - \mathbb{E}\left(\chi^2\middle|c\right)\right).
\end{aligned}$$

Based on Pav (2015),

$$\mathbb{E}\left(\chi^2\middle|c\right) = \log 2 + \psi\left(\frac{T-d_w}{2}\right),$$

where $\psi(\cdot)$ is the derivative of the log of the Gamma function.

Given $\log M_{\sigma^2}(0|c)$, $\frac{\partial}{\partial\xi}\log M_{\sigma^2}(0|c)$, and $\frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\sigma^2}(\xi|c)$, we can fully recover $\log M_{\sigma^2}(\xi|c)$ and hence uniquely determine f^{σ^2}. Please refer to Theorem 1.19 in Galambos and Simonelli (2004) for the uniqueness.

(iii) Identify the distribution of individual effects f^{λ}

Define

$$\mathring{y}_{i,1:T} = y_{i,1:T} - \beta' x_{i,0:T-1} = \lambda_i' w_{i,0:T-1} + u_{i,1:T}.$$

Let $\mathring{Y} = \mathring{y}_{i,1:T}$, $W = w_{i,0:T-1}'$, $\Lambda = \lambda_i$ and $U = u_{i,1:T}$. The above expression can be simplified as

$$\mathring{Y} = W\Lambda + U.$$

Denote $F_{\mathring{Y}}$, F_{Λ} and F_U as the conditional characteristic functions for $\mathring{Y}$, Λ and U, respectively. Based on Assumption (3.5.5) (4), F_{Λ} and F_U are non-vanishing almost everywhere.

Then, we obtain

$$\log F_{\Lambda}\left(W'\xi|c\right) = \log F_{\mathring{Y}}\left(\xi|c\right) - \log F_{U}\left(\xi|c\right).$$

Let $\zeta = W'\xi$ and $A_W = \left(W'W\right)^{-1}W'$, then the second derivative of $\log F_{\Lambda}\left(\zeta|c\right)$ is characterized by

$$\frac{\partial^2}{\partial\zeta\partial\zeta'}\log F_{\Lambda}\left(\zeta|c\right) = A_W\left(\frac{\partial^2}{\partial\xi\partial\xi'}\left(\log F_{\mathring{Y}}\left(\xi|c\right) - \log F_{U}\left(\xi|c\right)\right)\right)A_W'.$$

Moreover,

$$\begin{aligned}\log F_{\Lambda}\left(0|c\right) &= 0,\\ \frac{\partial}{\partial\zeta}\log F_{\Lambda}\left(0|c\right) &= i\mathbb{E}\left(A_W\mathring{Y}\middle|c\right),\end{aligned}$$

so we can pin down $\log\Lambda\left(\zeta|c\right)$ and f^{λ}. □

The proof of Proposition (3.5.8) for unbalanced panels follows in a similar manner.

B.4.2 Cross-sectional Heteroskedasticity

Proof. **(Proposition 3.5.9)**

(i) KL requirement

As λ and σ^2 are independent, we have

$$d_{KL}\left(f_0^{\lambda}f_0^{\sigma^2},\ f^{\lambda}f^{\sigma^2}\right) = d_{KL}\left(f_0^{\lambda},\ f^{\lambda}\right) + d_{KL}\left(f_0^{\sigma^2},\ f^{\sigma^2}\right).$$

Based on the observed sufficient statistics $\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}y_{it}$ with corresponding errors $\hat{u} =$

$\frac{1}{T}\sum_{t=1}^{T} u_{it}$, the KL requirement is: for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ f^{\sigma^2} \in \mathcal{F}^{\sigma^2} :: \\ \int f_0^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) \\ \quad \cdot \log \frac{\int f_0^{\lambda}\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u};\ 0, \frac{\sigma^{2\prime}}{T}\right) f_0^{\sigma^2}\left(\sigma^{2\prime}\right) d\hat{u}' d\sigma^{2\prime}}{\int f^{\lambda}\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u};\ 0, \frac{\sigma^{2\prime}}{T}\right) f^{\sigma^2}\left(\sigma^{2\prime}\right) d\hat{u}' d\sigma^{2\prime}} d\hat{u} d\sigma^2 d\hat{\lambda} < \epsilon \end{array}\right) > 0.$$

As in the proof of Proposition 3.4.7 part (i), similar convexity reasoning can be applied to bound the KL divergence on y by $d_{KL}\left(f_0^{\lambda} f_0^{\sigma^2},\ f^{\lambda} f^{\sigma^2}\right)$. The sufficient conditions for KL properties on λ and l are listed in Lemmas 3.4.8 and B.5.1. Note that since the KL divergence is invariant under variable transformations, the KL property of the distribution of l is equivalent to the KL property of the distribution of σ^2.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $f^{\sigma^2} \in U^c_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right)$

Orthogonal forward differencing yields $\tilde{y}_{it} \sim N\left(0, \sigma_i^2\right)$. Define $\hat{\sigma}_i^2 = \sum_{t=1}^{T-d_w} \tilde{y}_{it}^2 = \sigma_i^2 \chi_i^2$, where $\chi_i^2 \sim \chi^2\left(T - d_w\right)$ follows an i.i.d. chi-squared distribution with $(T - d_w)$ degrees of freedom. Here and below, I ignore the subscripts to simplify the notation.

Let $g^{\sigma^2}\left(\sigma^2\right) = f^{\sigma^2}\left(\sigma^2\right) - f_0^{\sigma^2}\left(\sigma^2\right)$. There are always tests if we observe σ^2, then for any g^{σ^2}, there exists a $\epsilon > 0$ such that

$$\int \left| g^{\sigma^2}\left(\sigma^2\right) \right| d\sigma^2 > \epsilon. \tag{B.4.1}$$

Similar to part (ii-b) in the proof of Proposition 3.4.7, here again I utilize the proof-by-contradiction technique. Suppose there is no test when $\hat{\sigma}^2$ is observed instead of σ^2, then

there exist a $\tilde{g}^{\sigma}$ such that

$$\tilde{h}\left(\hat{\sigma}^2\right) = \int \tilde{g}^{\sigma^2}\left(\frac{\hat{\sigma}^2}{\chi^2}\right) f_{\chi^2}\left(\chi^2\right) d\chi^2 = 0 \text{ for all } \hat{\sigma}^2,$$

due to the continuity of $\tilde{h}$. Here I utilize the Mellin transform for products of random variables. As σ^2 and χ^2 are independent, we have

$$M_{\hat{\sigma}^2}(\xi) = M_{\sigma^2}(\xi) \cdot M_{\chi^2}(\xi) = 0 \text{ for all } \xi.$$

The Mellin transform of chi-squared distribution $M_{\chi^2}(\xi) \neq 0$, then

$$M_{\sigma^2}(\xi) = 0 \text{ for all } \xi.$$

Note that $M_{\sigma^2}(\xi)$ uniquely determines $\tilde{g}^{\sigma^2}\left(\sigma^2\right)$. Then, the inverse Mellin transform leads to

$$\tilde{g}^{\sigma^2}\left(\sigma^2\right) = 0 \text{ for all } \sigma^2,$$

which contradicts equation (B.4.1). Therefore, there are also tests distinguishing the true $f_0^{\sigma^2}$ from alternative f^{σ^2} even when we only observe $\hat{\sigma}^2$.

(ii-b') $f^{\sigma^2} = f_0^{\sigma^2},\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)$

This is an intermediate step for part (ii-c). Once again I resort to proof by contradiction. Define $g^{\lambda}(\lambda) = f^{\lambda}(\lambda) - f_0^{\lambda}(\lambda)$. There are always tests if we observe λ, then for any g^{λ}, there exists a $\epsilon > 0$ such that

$$\int \left|g^{\lambda}(\lambda)\right| d\lambda > \epsilon. \tag{B.4.2}$$

Suppose there is no test when y is observed instead of λ, then there exist a $\tilde{g}^{\lambda}$ such that

$$\begin{aligned}
0 = \tilde{h}(y) &= \int \tilde{g}^{\lambda}(y-u)\,\phi\left(u;\ 0,\sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 \text{ for all } y \\
\Longrightarrow 0 = F_y(\xi) &= \int e^{-i\xi y}\tilde{g}^{\lambda}(y-u)\,\phi\left(u;\ 0,\sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 dy \\
&= \int e^{-i\xi(\lambda+\sigma v)}\tilde{g}^{\lambda}(\lambda)\,\phi\left(u;\ 0,\sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 d\lambda \\
&= F_{\lambda}(\xi)\cdot\int c_1 \exp\left(-c_2\xi^2\sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) d\sigma^2 = 0 \text{ for all } \xi \\
\Longrightarrow F_{\lambda}(\xi) &= 0 \text{ for all } \xi \\
\Longrightarrow \tilde{g}^{\lambda}(\lambda) &= 0 \text{ for all } \lambda,
\end{aligned}$$

which contradicts equation (B.4.2). Therefore, there are also tests if we know $f_0^{\sigma^2}$ but only observe y.

(ii-b) $f^{\sigma^2} \in U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right),\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)$

Without loss of generality, let $\Phi = \{\varphi\}$ and $\varphi^{\star}$ be the corresponding test function when $f_0^{\sigma^2}$ is known as in case (ii-b'). Then, the difference between $\mathbb{E}_f(\varphi^{\star})$ and $\mathbb{E}_{f_0}(\varphi^{\star})$ is

$$\begin{aligned}
&\int \varphi^{\star}\left(\hat{\lambda}\right) f^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0,\frac{\sigma^2}{T}\right) f^{\sigma^2}\left(\sigma^2\right) d\hat{u} d\sigma^2 d\hat{\lambda} \\
&- \int \varphi^{\star}\left(\hat{\lambda}\right) f_0^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0,\frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) d\hat{u} d\sigma^2 d\hat{\lambda} \\
> &\int \varphi^{\star}\left(\hat{\lambda}\right)\left(f^{\lambda}\left(\hat{\lambda}-\hat{u}\right) - f_0^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\ 0,\frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) d\hat{u} d\sigma^2 d\hat{\lambda} \\
&- \left|\int \varphi^{\star}\left(\hat{\lambda}\right) f^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0,\frac{\sigma^2}{T}\right)\left(f^{\sigma^2}\left(\sigma^2\right) - f_0^{\sigma^2}\left(\sigma^2\right)\right) d\hat{u} d\sigma^2 d\hat{\lambda}\right|.
\end{aligned}$$

Case (ii-b') implies that the first term is greater than some $\epsilon > 0$. Meanwhile, we can choose $\epsilon' = \epsilon/2$ and $\Phi' = \left\{\varphi'\left(\sigma^2\right) = 1\right\}$ for $U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right)$ so that the second term is bounded by $\epsilon/2$. Hence, $\mathbb{E}_f(\varphi^{\star}) - \mathbb{E}_{f_0}(\varphi^{\star}) > \epsilon/2$, and $\varphi^{\star}$ is the test function with respect to the alternative sub-region $\left\{f^{\sigma^2} \in U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right),\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)\right\}$. □

B.5 Extension: Heavy Tails

Lemma B.5.1 gives one set of conditions accommodating f_0^z with heavy tails using the Gaussian-mixture DPM prior. It follows Tokdar (2006) Theorem 3.3. The notation is slightly different from Tokdar (2006). Here G_0^z is defined on $\left(\mu_i^z, (\omega_i^z)^2\right)$, the mean and the variance, while Tokdar (2006) has the mean and the standard deviation as the arguments for G_0^z.

Lemma B.5.1. *(Tokdar, 2006)*

If f_0^z and the DP base distribution G_0^z satisfy the following conditions:

1. $\left|\int f_0^z(z)\log f_0^z(z)\,dz\right| < \infty$.
2. *For some* $\eta \in (0,1)$, $\int |z|^\eta f_0^z(z)\,dz < \infty$.
3. *There exist* $\omega_0 > 0$, $0 < b_1 < \eta$, $b_2 > b_1$, *and* $c_1, c_2 > 0$ *such that for large* $\mu > 0$,

$$\max\left\{\begin{array}{c} G_0^z\left(\left[\mu-\omega_0\mu^{\frac{\eta}{2}},\infty\right)\times\left[\omega_0^2,\infty\right)\right),\ G_0^z\left([0,\infty)\times\left(\mu^{2-\eta},\infty\right)\right), \\ G_0^z\left(\left(-\infty,-\mu+\omega_0\mu^{\frac{\eta}{2}}\right]\times\left[\omega_0^2,\infty\right)\right),\ G_0^z\left((-\infty,0]\times\left(\mu^{2-\eta},\infty\right)\right) \end{array}\right\} \geq c_1\mu^{-b_1},$$

$$\max\left\{\begin{array}{c} G_0^z\left((-\infty,\mu)\times(0,\exp\left(2\mu^\eta-1\right))\right), \\ G_0^z\left((-\mu,\infty)\times(0,\exp\left(2\mu^\eta-1\right))\right) \end{array}\right\} > 1-c_2\mu^{-b_2}.$$

Then, $f_0^z \in KL\left(\Pi^z\right)$.

The next lemma extends Lemma B.5.1 to the multivariate case. Then, Proposition B.5.3 largely parallels Proposition (3.5.10) with different condition sets for the KL property, which accounts for heavy tails in the true unknown distributions..

Lemma B.5.2. *(Heavy Tails: Multivariate)*

If f_0^z and the DP base distribution G_0^z satisfy the following conditions:

1. $\left|\int f_0^z(z)\log f_0^z(z)\,dz\right| < \infty$.
2. *For some* $\eta \in (0,1)$, $\int \|z\|^\eta f_0^z(z)\,dz < \infty$.

3. *There exist* $\omega_0 > 0$, $0 < b_1 < \eta$, $b_2 > b_1$, *and* $c_1, c_2 > 0$ *such that for large* $\mu > 0$, *for all directional vectors* $\|z^*\| = 1$,

$$\max \left\{ \begin{array}{c} G_0^z\left(\left[\mu - \omega_0 \mu^{\frac{\eta}{2}}, \infty\right) \times \left[\omega_0^2, \infty\right) | z^*\right), \; G_0^z\left([0, \infty) \times \left(\mu^{2-\eta}, \infty\right) | z^*\right), \\ G_0^z\left(\left(-\infty, -\mu + \omega_0 \mu^{\frac{\eta}{2}}\right] \times \left[\omega_0^2, \infty\right) | z^*\right), \; G_0^z\left((-\infty, 0] \times \left(\mu^{2-\eta}, \infty\right) | z^*\right) \end{array} \right\} \geq c_1 \mu^{-b_1},$$

$$\max \left\{ \begin{array}{c} G_0^z\left((-\infty, \mu) \times \left(0, \exp\left(2\mu^\eta - 1\right)\right) | z^*\right), \\ G_0^z\left((-\mu, \infty) \times \left(0, \exp\left(2\mu^\eta - 1\right)\right) | z^*\right) \end{array} \right\} > 1 - c_2 \mu^{-b_2},$$

where $G_0^z\left(\cdot | z^*\right)$ *represents the conditional distribution that is induced from* $G_0^z\left(\cdot\right)$ *conditional on the direction* z^*.

Then, $f_0^z \in KL\left(\Pi^z\right)$

Proposition B.5.3. *(General Model: Random Coefficients II)*

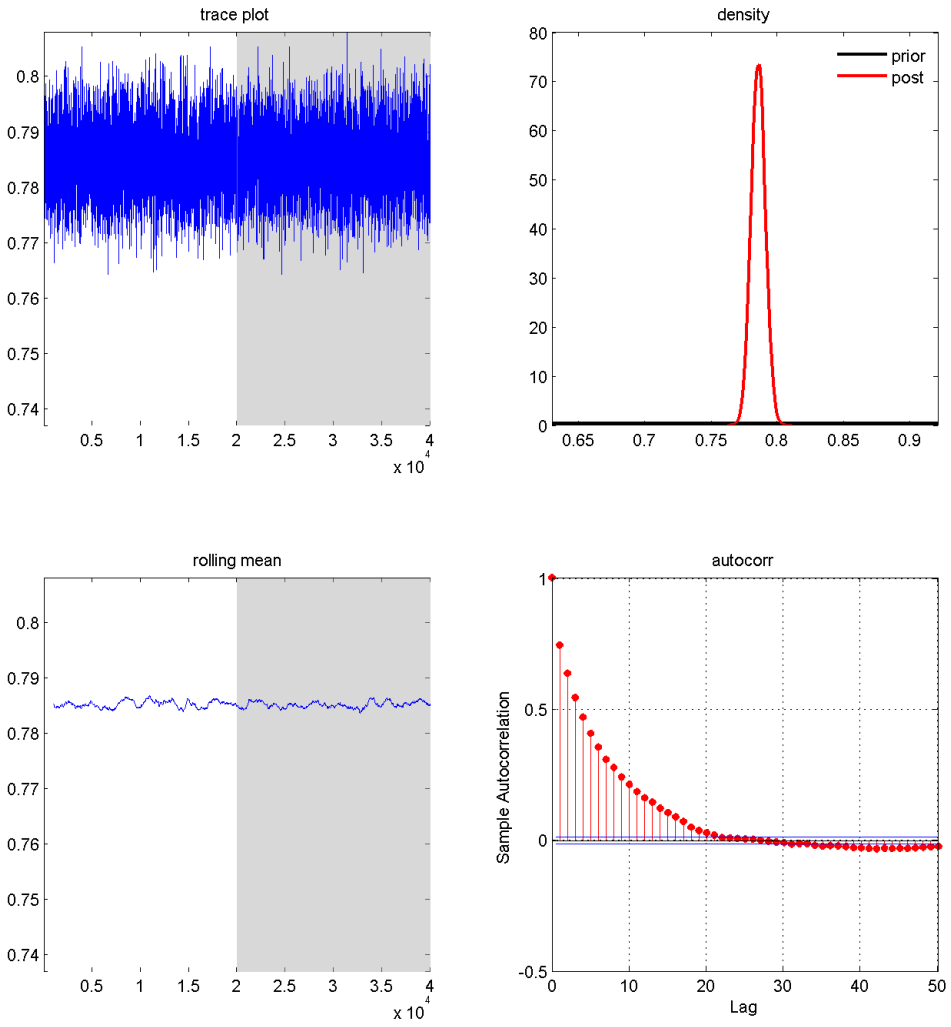
Suppose we have:

1. *Assumptions 3.5.3, 3.5.5 (3-4), 3.5.7, and 3.4.10.*
2. *Lemma B.5.2 on* λ *and Lemma B.5.1 on* l.
3. $\beta_0 \in \operatorname{supp}\left(\Pi^\beta\right)$.

Then, the posterior is weakly consistent at $\left(\beta_0, f_0^\lambda, f_0^{\sigma^2}\right)$.

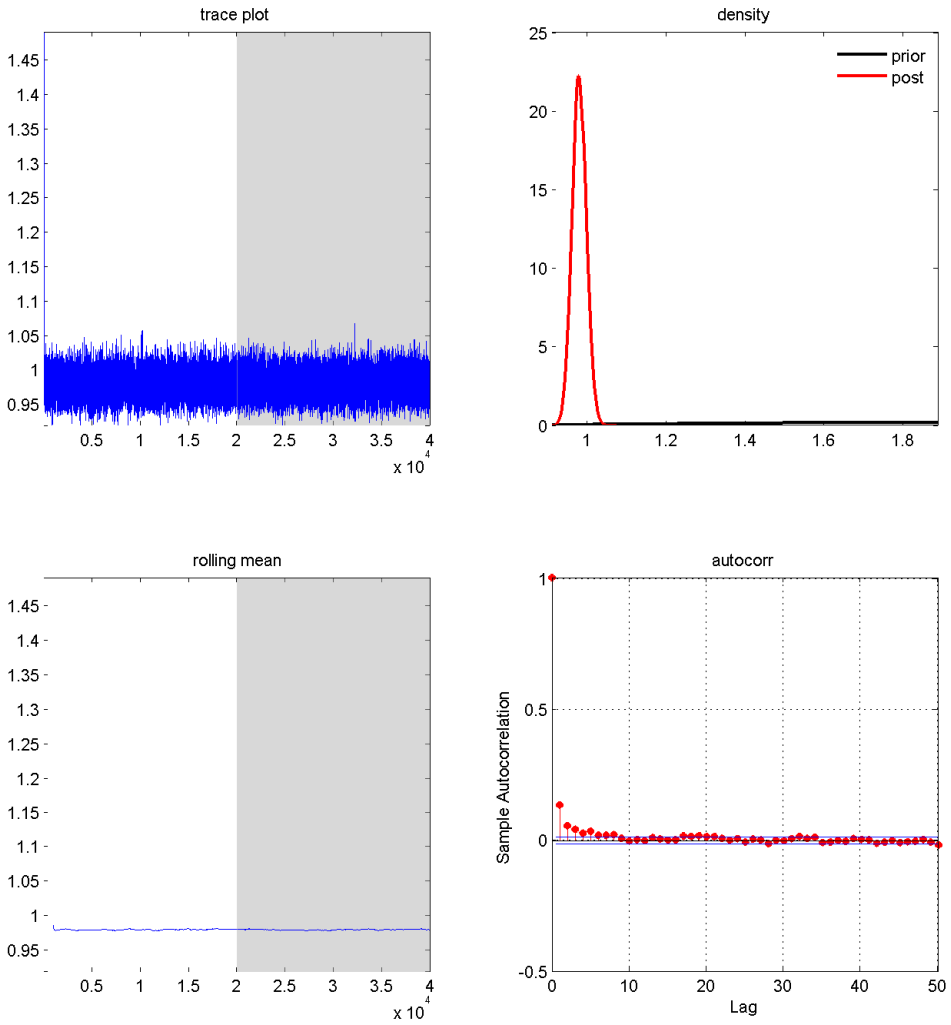
B.6 Simulations

Figure 15: Convergence Diagnostics: β



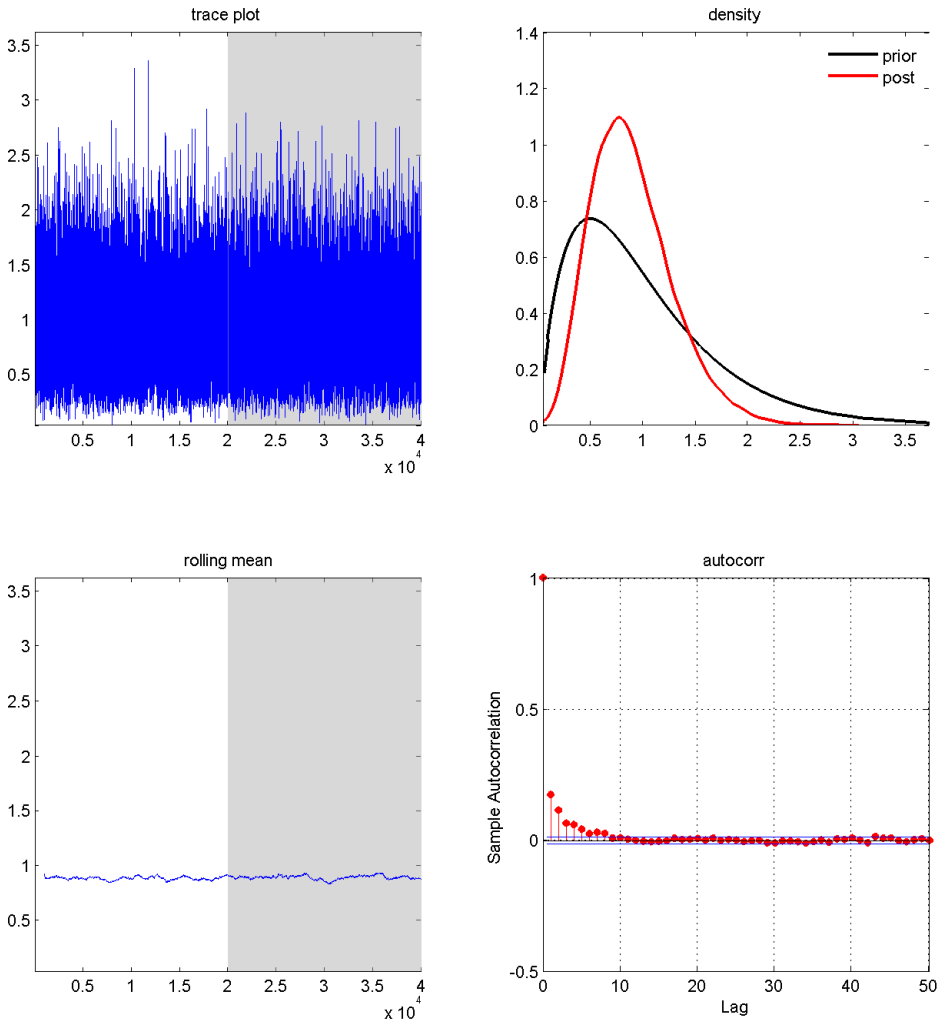
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 16: Convergence Diagnostics: σ^2



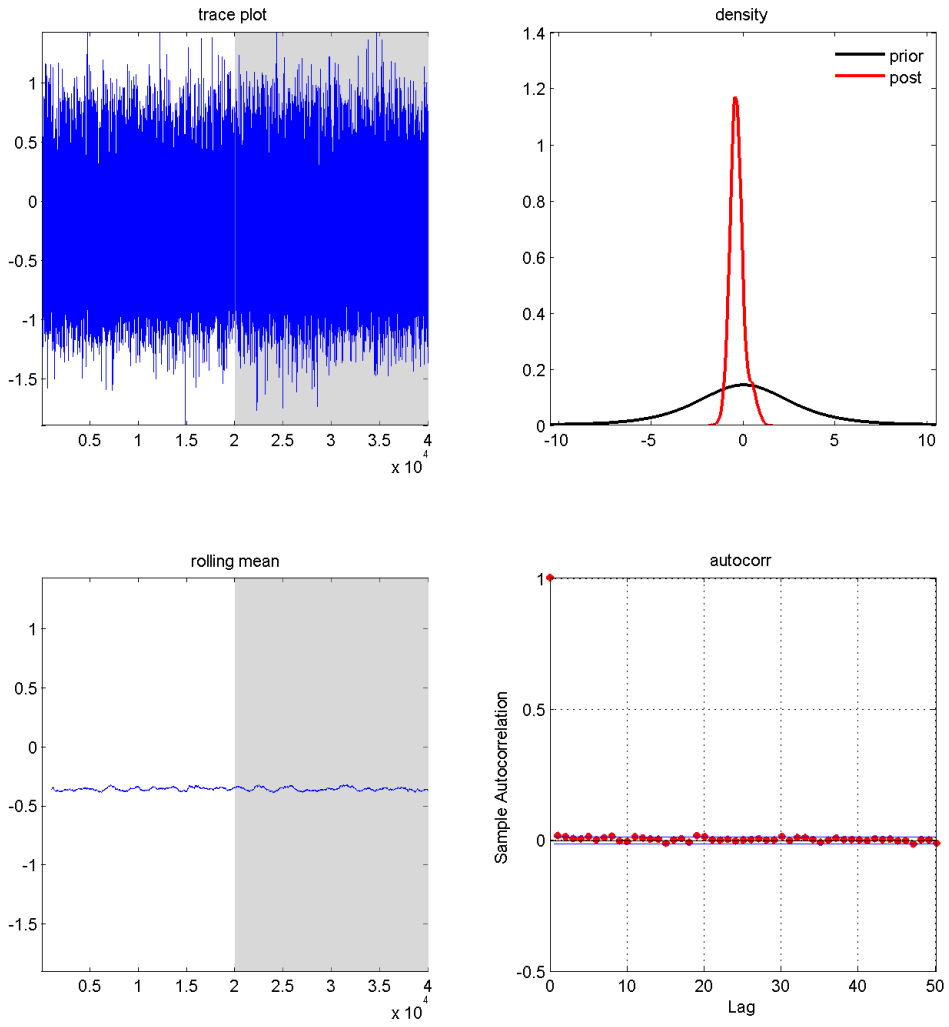
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 17: Convergence Diagnostics: α



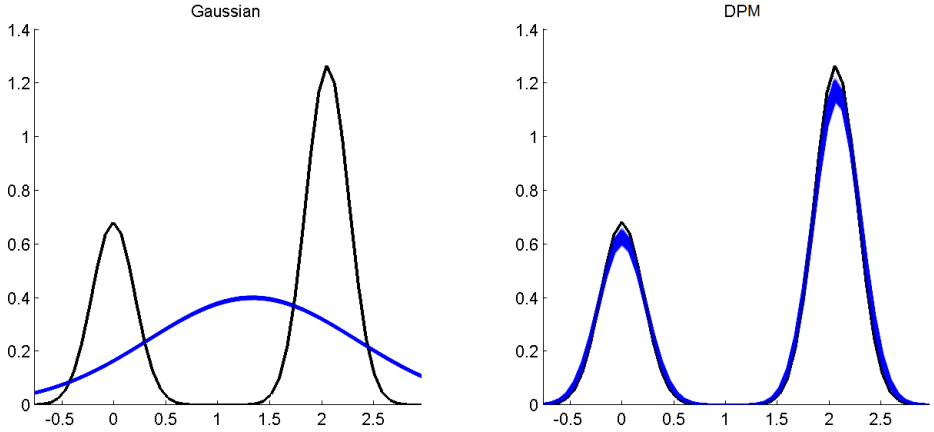
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 18: Convergence Diagnostics: λ_1



For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 19: f_0 vs $\Pi(f \mid y_{1:N,0:T})$: Baseline Model, $N = 10^5$



The black solid line represents the true λ_i distribution, f_0. The blue bands show the posterior distribution of f, $\Pi(f \mid y_{1:N,0:T})$.

BIBLIOGRAPHY

Akcigit, U. and Kerr, W. R. (2010). Growth through heterogeneous innovations.

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, **71** (4), 1121–1159.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, **9** (2), 291–312.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, **25** (2), 177–190.

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American statistical Association*, **76** (375), 598–606.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

— and Bond, S. (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.

— and Bonhomme, S. (2012a). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, **79** (3), 987–1020.

— and — (2012b). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, **79** (3), 987–1020.

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

— and Honoré, B. (2001). Panel data models: some recent developments. *Handbook of econometrics*, **5**, 3229–3296.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11** (5), 815–828.

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.

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, **98** (461), 224–235.

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, **33** (3), 882–886.

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

Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, **38** (5), 2916–2957.

Brown, L. D. and Greenshtein, E. (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.

Burda, M. and Harding, M. (2013). Panel probit with flexible correlated effects: quantifying technology spillovers in the presence of latent heterogeneity. *Journal of Applied Econometrics*, **28** (6), 956–981.

—, — and Hausman, J. (2012). A Poisson mixture model of discrete choice. *Journal of Econometrics*, **166** (2), 184–203.

Canova, F. and Ciccarelli, M. (2013). *Panel Vector Autoregressive Models: A Survey*. Working Paper Series, European Central Bank 1507, European Central Bank.

Chamberlain, G. and Hirano, K. (1999). Predictive distributions based on longitudinal earnings data. *Annales d'Economie et de Statistique*, pp. 211–242.

Chung, Y. and Dunson, D. B. (2012). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*.

Covas, F. B., Rump, B. and Zakrajsek, E. (2014). Stress-testing US bank holding companies: a dynamic panel quantile regression approach. *International Journal of Forecasting*, **30** (3), 691–713.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pp. 665–685.

Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *The Annals of Statistics*, pp. 68–87.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39** (4), 863–883.

— and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13** (3).

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pp. 23–27.

Dunson, D. B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika*, **96** (2), 249–262.

— and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, **95** (2), 307–323.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, **106** (496), 1602–1614.

— (2012). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1. Cambridge University Press.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90** (430), 577–588.

Fernandez, C. and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, **16** (01), 80–101.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pp. 1386–1403.

— (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, **36** (2), 454–456.

Galambos, J. and Simonelli, I. (2004). *Products of Random Variables: Applications to Problems of Physics and to Arithmetical Functions*. Marcel Dekker.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, **26** (2), 216–230.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R. *et al.* (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, **27** (1), 143–158.

Ghosh, J. K. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer-Verlag.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, **57** (298), 369–375.

Griffin, J. E. (2016). An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing*, **26** (1), 423–441.

Gu, J. and Koenker, R. (2016a). Empirical bayesball remixed: empirical bayes methods for longitudinal data. *Journal of Applied Economics (Forthcoming)*.

— and — (2016b). Unobserved heterogeneity in income dynamics: an empirical bayes perspective. *Journal of Business & Economic Statistics (Forthcoming)*.

Hall, B. H. and Rosenberg, N. (2010). *Handbook of the Economics of Innovation*, vol. 1. Elsevier.

Hastie, D. I., Liverani, S. and Richardson, S. (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, **25** (5), 1023–1037.

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

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hsiao, C. (2014). *Analysis of panel data*. 54, Cambridge university press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96** (453), 161–173.

— and — (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11** (3), 508–532.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif.: University of California Press, pp. 361–379.

Jensen, M. J., Fisher, M. and Tkac, P. (2015). Mutual fund performance when learning the distribution of stock-picking skill.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, **37** (4), 1647–1684.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21** (1), 93–105.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 419–430.

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

Lee, Y., Amaral, L. A. N., Canning, D., Meyer, M. and Stanley, H. E. (1998). Universal features in the growth dynamics of complex organizations. *Physical Review Letters*, **81** (15), 3275.

Liu, L. (2016). Density forecasts in panel data models: a semiparametric bayesian perspective. *Manuscript, University of Pennsylvania*.

—, Moon, H. R. and Schorfheide, F. (2016). Forecasting with dynamic panel data models.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2015). PReMiuM: an R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, **64** (7).

Marcellino, M., Stock, J. H. and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, **135** (1), 499–526.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9** (2), 249–265.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, **38** (3), 1733–1766.

— and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, **168** (2), 332–346.

— and — (2014). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, **30**, 606–646.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95** (1), 169–186.

Pati, D., Dunson, D. B. and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, **116**, 456–472.

Pav, S. E. (2015). Moments of the log non-central chi-square distribution. *arXiv preprint arXiv:1503.06266*.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, **178**, 624–638.

Robb, A., Ballou, J., DesRoches, D., Potter, F., Zhao, Z. and Reedy, E. (2009). An overview of the Kauffman Firm Survey: results from the 2004-2007 data. *Available at SSRN 1392292*.

— and Seamans, R. (2014). The role of R&D in entrepreneurial finance and performance. *Available at SSRN 2341631*.

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.

Rossi, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press.

Santarelli, E., Klomp, L. and Thurik, A. R. (2006). Gibrat's law: an overview of the empirical literature. In *Entrepreneurship, Growth, and Innovation*, Springer, pp. 41–73.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **4** (1), 10–26.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pp. 639–650.

Shin, M. (2014). Bayesian GMM.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pp. 90–110.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, **36** (1), 45–54.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, **2**, 298–331.

Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73** (1), 37–57.

Zarutskie, R. and Yang, T. (2015). How did young firms fare during the great recession? Evidence from the Kauffman Firm Survey. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, University of Chicago Press.