Density Forecasts in Panel Data Models:

A Semiparametric Bayesian Perspective*

Laura Liu†

March 18, 2018

Abstract

This paper constructs individual-specific density forecasts for a panel of firms or households using a dynamic linear model with common and heterogeneous coefficients and cross-sectional heteroskedasticity. The panel considered in this paper features a large cross-sectional dimension N but short time series T. Due to the 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, I assume that there is an underlying distribution of heterogeneous parameters, model this distribution nonparametrically allowing for correlation between heterogeneous parameters and initial conditions as well as individual-specific regressors, and then estimate this distribution by pooling the information from the whole cross-section together. Theoretically, 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. Methodologically, I develop a simulation-based posterior sampling algorithm specifically addressing the nonparametric density estimation of unobserved heterogeneous parameters. Monte Carlo simulations and an application to young firm dynamics demonstrate improvements in density forecasts relative to alternative approaches.

JEL Codes: C11, C14, C23, C53, L25

Keywords: Bayesian, Semiparametric Methods, Panel Data, Density Forecasts, Posterior Consistency, Young Firms Dynamics

---

*First version: November 15, 2016. Latest version: https://goo.gl/8zZZwn. I am indebted to my advisors, Francis X. Diebold and Frank Schorfheide, for much help and guidance at all stages of this research project. I also thank the other members of my committee, Xu Cheng and Francis J. DiTraglia, for their advice and support. I further benefited from many helpful discussions with Stéphane Bonhomme, Benjamin Connault, Hyungsik R. Moon, and seminar participants at the University of Pennsylvania, the Federal Reserve Bank of Philadelphia, the Federal Reserve Board, the University of Virginia, Microsoft, the University of California, Berkeley, the University of California, San Diego (Rady), Boston University, the University of Illinois at Urbana–Champaign, Princeton University, and Libera Università di Bolzano, as well as conference participants at the 26th Annual Meeting of the Midwest Econometrics Group, NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics, the 11th Conference on Bayesian Nonparametrics, Microeconometrics Class of 2017 Conference, Interactions Workshop 2017, and First Italian Workshop of Econometrics and Empirical Economics: Panel Data Models and Applications. I would also like to acknowledge the Kauffman Foundation and the NORC Data Enclave for providing researcher support and access to the confidential microdata. All remaining errors are my own. This paper does not necessarily reflect the views of the Board of Governors or the Federal Reserve System.

†Federal Reserve Board, laura.liu@frb.gov.

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 applications. For example, PSID can be used to analyze income dynamics (Hirano, 2002; Gu and Koenker, 2017b), and bank balance sheet data can help conduct bank stress tests (Liu *et al.*, 2017). 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:

$$\underbrace{y_{it}}_{\text{performance}} = \beta y_{i,t-1} + \underbrace{\lambda_i}_{\text{skill}} + \underbrace{u_{it}}_{\text{shock}}, \quad u_{it} \sim N\left(0, \sigma^2\right), \tag{1.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. y_{it} is the observed firm performance such as the log of employment,[1] 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. Based on the observed panel from period 0 to period T, I am interested in forecasting the future performance of any specific firm in period $T+1$, $y_{i,T+1}$.[2]

The panel considered in this paper features a large cross-sectional dimension N but short time series T.[3] This framework is appealing to the young firm dynamics example because the number of observations for each young firm is restricted by its age.[4] Good estimates of the unobserved skill λ_is facilitate good forecasts of $y_{i,T+1}$s. Because of the short T, traditional methods have difficulty in disentangling the unobserved skill λ_i from the shock u_{it}, which contaminates the estimates of λ_i. The naive estimators that only utilize the firm-specific observations are inconsistent, even if N goes to infinity.

[1] Employment is one of the standard measures in the firm dynamics literature (Akcigit and Kerr, 2016; Zarutskie and Yang, 2015).

[2] In the main body of the paper, I consider a more general specification that accommodates many important features of real-world empirical studies, such as strictly exogenous and predetermined covariates, correlated random coefficients, and cross-sectional heteroskedasticity.

[3] Which T can be considered small depends on the dimension of individual heterogeneity (which can be multi-dimensional in the general model), the cross-sectional dimension, and the size of the shocks. There can still be a significant gain in density forecasts even when T exceeds 100 in simulations with fairly standard data generating processes. Roughly speaking, the proposed predictor would provide a sizable improvement as long as the time series for individual i is not informative enough to fully reveal its individual effects.

[4] Although I describe the econometric intuition using the young firm dynamics application as an example, the method can be applied to many other 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 and holding company data).

To tackle this problem, I assume that λ_i is drawn from an underlying skill distribution f and estimate this distribution by pooling the information from the whole cross-section. In terms of modeling f, the parametric Gaussian density misses many features in real-world data, such as asymmetricity, heavy tails, and multiple peaks. For example, as good ideas are scarce, the skill distribution of young firms may be highly skewed. In this sense, the challenge now is how we can model f more carefully and flexibly. Here I estimate f via a nonparametric Bayesian approach where the prior is constructed from a mixture model and allows for correlation between λ_i and the initial condition y_{i0} (i.e. a correlated random effects model).

Conditional on f, we can, intuitively speaking, treat it as a prior distribution and combine it with firm-specific data to obtain the firm-specific posterior. In a special case where the common parameters are set to $\beta = 0$ and $\sigma^2 = 1$, the firm-specific posterior is characterized by Bayes' theorem,

$$p\left(\lambda_i \left| f, y_{i,0:T}\right.\right) = \frac{p\left(y_{i,1:T}\right| \lambda_i\right) f\left(\lambda_i \left| y_{i0}\right.\right)}{\int p\left(y_{i,1:T}\right| \lambda_i\right) f\left(\lambda_i \left| y_{i0}\right.\right) d\lambda_i}. \tag{1.2}$$

This firm-specific posterior helps provide a better inference about the unobserved skill λ_i of each individual firm and a better forecast of the firm-specific future performance, thanks to the underlying distribution f that integrates the information from the whole panel in an efficient and flexible way.[5]

It is natural to construct density forecasts based on the firm-specific posterior. In general, forecasting can be done in point, interval, or density fashion, with density forecasts giving the richest insight regarding future outcomes. By definition, a density forecast provides a predictive distribution of firm i's future performance and summarizes all sources of uncertainties; hence, it is preferable in the context of young firm dynamics and other applications with large uncertainties and nonstandard distributions. In particular, for the dynamic panel data model as specified in equation (1.1), the density forecasts reflect uncertainties arising from the future shock $u_{i,T+1}$, individual heterogeneity λ_i, and estimation uncertainty of common parameters $\left(\beta, \sigma^2\right)$ and skill distribution f. Moreover, once the density forecasts are obtained, one can easily recover the point and interval forecasts.

The contributions of this paper are threefold. First, I establish the theoretical properties of the proposed Bayesian predictor when the cross-sectional dimension N tends to infinity. To begin, I provide conditions for identifying both the parametric component and the nonparametric component.[6] Then, I prove that both the estimated common parameters and the estimated distribution of the individual heterogeneity achieve posterior consistency in strong topology. Compared with previous literature on posterior consistency, there are several challenges in the panel data framework: (1) a deconvolution problem disentangling unobserved individual effects and independent shocks, (2) an unknown common shock size in cross-sectional homoskedastic cases, (3) unknown individual-

[5] Note that this is only an intuitive explanation why the skill distribution f is crucial. In the actual implementation, the estimation of the correlated random effect distribution f, the estimation of common parameters $\left(\beta, \sigma^2\right)$, and the inference of firm-specific skill λ_i are all done simultaneously.

[6] In the baseline model, the parametric component is $\left(\beta, \sigma^2\right)$, and the nonparametric component is f.

specific shock sizes in cross-sectional heteroskedastic cases, (4) strictly exogenous and predetermined variables (including lagged dependent variables) as covariates, and (5) correlated random effects[7] addressed by flexible conditional density estimation.

Building on the posterior consistency of the estimates, we can bound the discrepancy between the proposed density predictor and the oracle to be arbitrarily small, where the oracle predictor is 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.

Second, I develop a posterior sampling algorithm specifically addressing nonparametric density estimation of the unobserved individual effects. For a random effects model, which is a special case where the individual effects are independent of the conditioning variables,[8] the f part becomes a relatively simple unconditional density estimation problem. I adopt a Dirichlet Process Mixture (DPM) prior for 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).

Third, Monte Carlo simulations demonstrate improvements in density forecasts relative to alternative predictors with various parametric priors on f, evaluated by the log predictive score. An application to young firm dynamics also shows that the proposed predictor provides more 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 of firm-specific coefficients and cross-sectional heteroskedasticity, as well as whether and how these unobserved heterogeneous features depend on the initial condition of the firms.

Moreover, the method proposed in this paper is applicable beyond forecasting. Here estimating heterogeneous parameters is important because we want to generate good forecasts, but in other cases, the heterogeneous parameters themselves could be the objects of interest. For example, the technique developed here can be adapted to infer individual-specific treatment effects.

Related Literature First, this paper contributes to the literature on individual forecasts in a panel data setup, and is closely related to Liu *et al.* (2017) and Gu and Koenker (2017a,b). Liu *et al.* (2017) focus on point forecasts. They utilize the idea of Tweedie's formula to steer away from the complicated deconvolution problem in estimating λ_i and establish the ratio optimality of point forecasts. Unfortunately, the Tweedie shortcut is not applicable to the inference of the underlying λ_i distribution and therefore not suitable for density forecasts. In addition, this paper addresses

[7] In heteroskedastic cases, the terminologies "random effects" and "correlated random effects" also apply to individual-specific σ_i^2, which is slightly different from the traditional definitions focusing on λ_i.

[8] The conditioning set can include initial values of predetermined covariates and entire sequences of strictly exogenous covariates.

cross-sectional heteroskedasticity where σ_i^2 is an unobserved random quantity, while Liu *et al.* (2017) incorporate cross-sectional and time-varying heteroskedasticity via a deterministic function of observed conditioning variables.[9]

Gu and Koenker (2017b) address the density estimation problem, but with a different method. 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 cross-sectional heterogeneity (including cross-sectional heteroskedasticity) interacting with the initial conditions, whereas the Gu and Koenker (2017b) approach focuses on random effects models without this interaction.

In their recent paper, Gu and Koenker (2017a) 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 the case in many empirical setups. Second, unlike their set of fixed scale parameters, 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 an "automatic" model selection.[10]

There have also been empirical works on the DPM model with panel data,[11] 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 with setups being slightly different from this paper. Hirano (2002) considers flexibility in the u_{it} distribution instead of the 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, Li and Vuong (1998), Delaigle *et al.* (2008), Evdokimov (2010), and Hu (2017) have studied the similar deconvolution problem and estimated the λ_i distribution in a frequentist way. Also see Compiani and Kitamura (2016) for a review of frequentist applications of mixture

[9] Note that, in this paper, the identification restriction to ensure unobserved random σ_i^2 can only permit time-varying distribution for v_{it} while keeping zero mean and unit variance (see Remark 3.2 (iv)). However, considering that this paper focuses on the scenarios with a short time dimension, lack of time-varying heteroskedasticity would not be a major concern.

[10] Section 5 shows the simulation results comparing the DP prior vs the DPM prior, where both incorporate a hyperprior for the scale parameter.

[11] For earlier works regarding full Bayesian analyses with parametric priors on λ_i, see Lancaster (2002) (orthogonal reparametrization and a flat prior), Chamberlain and Hirano (1999), Chib and Carlin (1999), Sims (2000) (Gaussian prior), and Chib (2008) (student-t and finite mixture priors).

models. However, 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. See the handbooks, Ghosh and Ramamoorthi (2003), Hjort *et al.* (2010), and Ghosal and van der Vaart (2017), for a more thorough review and discussion on posterior consistency in Bayesian nonparametric problems. In particular, Canale and De Blasi (2017) relax the tail conditions to accommodate multivariate location-scale mixtures for unconditional density estimation. To handle conditional density estimation, the mixing probabilities can be characterized by a multinomial choice model (Norets, 2010; Norets and Pelenis, 2012), a kernel stick-breaking process (Norets and Pelenis, 2014; Pelenis, 2014; Norets and Pati, 2017), 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 more flexible in the conditional measure. This paper builds on these previous works and establishes the strong consistency for a multivariate conditional density estimator featuring infinite location-scale mixtures with a probit stick-breaking process. Then, this paper further takes into account the deconvolution and dynamic panel data structure, as well as obtains the convergence of the proposed predictor to the oracle predictor in strong topology.

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 (Escobar and West, 1995; Neal, 2000). The other resorts to the stick-breaking process (Sethuraman, 1994) and directly incorporates the unknown distribution into the sampling procedure. This paper utilizes a sampler from the second category, the blocked Gibbs sampler by Ishwaran and James (2001, 2002), as a building block for the proposed algorithm. It incorporates truncation approximation and augments the data with auxiliary component probabilities, which breaks down the complex posterior structure and thus enhances mixing properties as well as reduces computation time.[12] 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 (2016) 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 slow and jobless recovery. See the handbook by Hall and Rosenberg (2010) for a thorough review on young firm innovation. The empirical analysis of this paper builds

[12]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 to the simpler truncation sampler.

on these previous findings. Besides more accurate density forecasts, we can also obtain the latent heterogeneity structure of firm-specific coefficients and cross-sectional heteroskedasticity.

The rest of the paper is organized as follows. Section 2 introduces the general panel data model, the predictors for density forecasts, and the nonparametric Bayesian priors. Section 3 characterizes identification conditions and large sample properties. Section 4 proposes the posterior sampling algorithms. Section 5 examines the performance of the semiparametric Bayesian predictor using simulated data, and Section 6 applies the proposed predictor to the confidential microdata from the Kauffman Firm Survey and analyzes the empirical findings on young firm dynamics. Finally, Section 7 concludes and sketches future research directions. Notations, proofs, as well as additional algorithms and results are in the Appendix.

2 Model

2.1 General Panel Data Model

The general panel data model with (correlated) random coefficients and potential cross-sectional heteroskedasticity can be specified as

$$y_{it} = \beta' x_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right) \tag{2.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. Similar to the baseline setup in equation (1.1), 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 $y_{i,T+h}$ for any individual i. 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).

$w_{i,t-1}$ is a vector of observed covariates that have heterogeneous effects on the outcomes, with λ_i being the unobserved heterogeneous coefficients. $w_{i,t-1}$ is strictly exogenous and captures the key sources of individual heterogeneity. 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 (1.1). Moreover, it is also helpful to include other key covariates of interest whose effects are more diverse cross-sectionally; for example, R&D activities may benefit the young firms in different magnitudes. Furthermore, the current setup can also take into account deterministic or stochastic aggregate effects; for example, different young firms may respond differently to the financial crisis. For notation clarity, I partition $w_{i,t-1} = \left[1, w_{t-1}^{A\prime}, w_{i,t-1}^{I\prime}\right]'$, where w_{t-1}^{A} stands for a vector of aggregate variables, and $w_{i,t-1}^{I}$ is composed of individual-specific variables.

$x_{i,t-1}$ 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 and $x_{i,t-1}^{P}$ 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 characteristics as well as local and national economic conditions, which help control for other sources of variation and facilitate forecasts. In addition, we let $x_{i,t-1}^{P*}$ denote the subgroup of $x_{i,t-1}^{P}$ excluding lagged outcomes and decompose $\beta = \left[\beta^{O\prime}, \beta^{P*\prime}, \rho\right]'$, where $\left(\beta^{O\prime}, \beta^{P*\prime}, \rho\right)$ are the coefficients corresponding to $\left(x_{i,t-1}^{O\prime}, x_{i,t-1}^{P*\prime}, y_{i,t-1}\right)$, respectively. Here, the distinction between homogeneous effects $\beta' x_{i,t-1}$ versus heterogeneous effects $\lambda_i' w_{i,t-1}$ reveals the latent nonstandard structures for the key effects while avoiding the curse-of-dimensionality problem.

u_{it} is an individual-time-specific shock characterized by zero mean and potential cross-sectional heteroskedasticity σ_i^2,[13] with cross-sectional homoskedasticity being a special case where $\sigma_i^2 = \sigma^2$. In a unified framework, denote ϑ as the common parameters, h_i as the individual heterogeneity, and f as the underlying distribution of h_i.

$$\vartheta = \left(\beta, \sigma^2\right),\ h_i = \lambda_i,\ \text{in cross-sectional homoskedastic cases,}$$
$$\vartheta = \beta,\ h_i = \left(\lambda_i, \sigma_i^2\right),\ \text{in cross-sectional heteroskedastic cases.}$$

We can define the conditioning set at period t to be

$$c_{i,t-1} = \left(x_{i,0:t-1}^{P}, x_{i,0:T}^{O}, w_{0:T}^{A}, w_{i,0:T}^{I}\right). \tag{2.2}$$

Note that as $x_{i,t-1}^{P}$ is predetermined, the sequences of $x_{i,t-1}^{P}$ in the conditioning set $c_{i,t-1}$ start from period 0 to period $t-1$; $x_{i,t-1}^{O}$, w_{t-1}^{A}, and $w_{i,t-1}^{I}$ are strictly exogenous, so the conditioning set $c_{i,t-1}$ contains their entire sequences. Moreover, we can define the part of $c_{i,t-1}$ that is composed of individual-specific variables as $c_{i,t-1}^{*} = \left(x_{i,0:t-1}^{P}, x_{i,0:T}^{O}, w_{i,0:T}^{I}\right)$. Then, we can further denote $D = \left(\{D_i\}_{i=1}^{N}, D_A\right)$ as a shorthand for the data sample used for estimation, which constitutes the conditioning set for posterior inference. $D_i = c_{i,T}^{*}$ contains the observed data for individual i, and $D_A = w_{0:T}^{A}$ contains the aggregate regressors with heterogeneous effects.

As stressed in the motivation, the underlying distribution of individual effects is the key to better density forecasts. In the literature, there are usually two kinds of assumptions imposed on this distribution. One is the random coefficients model, where the individual effects h_i are independent of the conditioning variables c_{i0}, which include initial values of predetermined covariates and full sequences of strictly exogenous covariates.[14] The other is the correlated random coefficients model, where h_i and c_{i0} could be correlated with each other. This paper considers both random coefficients

[13] In many empirical applications, such as the young firm analysis in Section 6, risk may largely vary over the cross-section, which also contributes considerably to more precise density forecasts.

[14] In the baseline setup as a special case, the conditioning set is a singleton with the initial outcome y_{i0} being the only element.

and correlated random coefficients models while focusing on the latter. The random coefficients model is more parsimonious and easier to implement, but the correlated random coefficients model is more realistic for young firm dynamics as well as many other empirical setups,[15] and random coefficients can be viewed as a special case of correlated random coefficients with zero dependence.

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.

Extension: Unbalanced Panels The above setup can be extended to unbalanced panels with randomly omitted observations, which incorporates more data into the estimation and elicits more information for the prediction. Conditional on the covariates, the common parameters, and the distributions of individual heterogeneities, $y_{i,t+1}$s are cross-sectionally independent, so the theoretical argument and numerical implementation are still valid in like manner.

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

$$c_{i,t-1} = \left(x^P_{i,\tau_{i,t-1}}, x^O_{i,\tau_{iT}}, w^A_{\tau_{iT}}, w^I_{i,\tau_{iT}} \right), \tag{2.3}$$

where the set for time periods $\tau_{i,t-1} = \{0, t_{0i}-1, \cdots, t_{1i}-1, T\} \cap \{0, \cdots, t-1\}$. Note that t_{i0} can be 1, and t_{i1} can be T, so this structure is also able to accommodate balanced panels. Accordingly, the individual-specific component of $c_{i,t-1}$ is $c^*_{i,t-1} = \left(x^P_{i,\tau_{i,t-1}}, x^O_{i,\tau_{iT}}, w^I_{i,\tau_{iT}} \right)$.

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,

$$f^{cond}_{i,T+1}(y|\vartheta, f) = \int \underbrace{p(y|\, h_i, \vartheta, w_{iT}, x_{iT})}_{\text{future shock}} \cdot \underbrace{p(h_i\,|\vartheta, f, D_i, D_A)}_{\text{individual heterogeneity}} dh_i, \tag{2.4}$$

which provides the density forecasts of $y_{i,T+1}$ conditional on the common parameters ϑ, underlying distribution f, and individual i's and aggregate data (D_i, D_A). The first term $p(y|\, h_i, \vartheta, w_{iT}, x_{iT})$ captures individual i's uncertainty due to the future shock $u_{i,T+1}$. The second term

$$p(h_i\,|\vartheta, f, D_i, D_A) = \frac{\prod_{t=1}^T p(y_{it}|\, h_i, \vartheta, w_{i,t-1}, x_{i,t-1})\, f(h_i\,|c_{i0})}{\int \prod_{t=1}^T p(y_{it}|\, h_i, \vartheta, w_{i,t-1}, x_{i,t-1})\, f(h_i\,|c_{i0})\, dh_i}$$

[15] In the baseline setup, the correlated random coefficients model can be interpreted as saying that a young firm's initial performance may reflect its underlying skill, which is a more sensible assumption.

is the individual-specific posterior. It characterizes individual i's uncertainty due to heterogeneity that arises from insufficient time-series information to infer individual h_i. The common distribution f helps in formulating this source of uncertainty and hence contributes to individual i's 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 ϑ_0 and the underlying distribution f_0, but not the individual effects h_i. Then, the oracle predictor is formulated by plugging the true values (ϑ_0, f_0) into the conditional predictor in equation (2.4),

$$f_{i,T+1}^{oracle}(y) = f_{i,T+1}^{cond}(y|\vartheta_0, f_0). \tag{2.5}$$

In practice, (ϑ, f) are all unknown and need to be estimated, thus introducing another source of uncertainty. I adopt a conjugate prior for the common parameters ϑ (mulitvariate normal inverse gamma for cross-sectional homoskedastic cases and mulitvariate normal for cross-sectional heteroskedastic cases) in order to stay close to the linear regression framework. I resort to the nonparametric Bayesian prior (specified in the next subsection) to flexibly model the underlying distribution, which could better approximate the true distribution f_0, and the resulting feasible predictor would be close to the oracle. 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 (ϑ, f),[16]

$$f_{i,T+1}^{sp}(y) = \int \underbrace{f_{i,T+1}^{cond}(y|\vartheta, f)}_{\text{shock \& heterogentity}} \cdot \underbrace{d\Pi(\vartheta, f\,|D)}_{\text{estimation uncertainty}} d\vartheta df. \tag{2.6}$$

The conditional predictor reflects uncertainties due to future shock and individual heterogeneity, whereas the posterior of (ϑ, f) captures estimation uncertainty. Note that the inference of (ϑ, f) pools information from the whole cross-section; once conditioned on (ϑ, f) and the aggregate observables D_A, individuals' outcomes are independent across i, and only individual i's data are further needed for its density forecasts.

2.3 Nonparametric Bayesian Priors

A prior on the 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 mixture models can effectively approximate a general class of distributions (see Section 3) while being relatively easy to implement (see Section 4). The specific functional form of the nonparametric Bayesian prior depends on whether f is characterized by a random coefficients model or a correlated random coefficients model. The correlated random coefficients setup is more involved but can be crucial in some empirical

[16] The superscript "sp" stands for "semiparametric".

studies, such as the young firm dynamics application in this paper.

In cross-sectional heteroskedastic cases, I incorporate another flexible prior on the distribution of σ_i^2. Define $l_i = \log \frac{\bar{\sigma}^2\left(\sigma_i^2 - \underline{\sigma}^2\right)}{\bar{\sigma}^2 - \sigma_i^2}$, where $\underline{\sigma}^2$ is some small positive number and $\bar{\sigma}^2$ is some large positive number. Then, the support of $f_0^{\sigma^2}$ is bounded by $\left[\underline{\sigma}^2, \bar{\sigma}^2\right]$ and thus satisfies the requirement for the asymptotic convergence of the Bayesian estimates and density forecasts in Propositions 3.10, 3.13, and 3.14. This transformation ensures an unbounded support for l_i so that we can employ similar prior structures to λ_i and l_i. Note that because λ_i and σ_i^2 are independent with respect to each other, their mixture structures are completely separate. For a concise exposition, I define a generic variable z that can represent either λ or l, and then include z as a superscript to indicate whether a specific parameter belongs to the λ part or the l part.

2.3.1 Random Coefficients Model

In the random coefficients model, the individual heterogeneity $z_i\,(=\lambda_i \text{ or } l_i)$ is assumed to be independent of the conditioning variables c_{i0}, so the inference of the underlying distribution f can be considered an unconditional density estimation problem.

The first candidate is the Dirichlet Process (DP), which casts a distribution over a set of discrete distributions. We denote $G \sim DP\left(\alpha, G_0\right)$, where the base distribution G_0 characterizes the center of the DP, and the scale parameter α represents the precision (inverse-variance) of the DP. [17]

However, considering the baseline model, imposing a DP prior on the distribution f means 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 extension is to assume z_i follows a continuous parametric distribution $f\left(z;\theta\right)$ where θ are the parameters, and adopt a DP prior for the distribution of θ.[18] Then, the parameters θ are discrete while the individual heterogeneity z enjoys a continuous distribution. This additional layer of mixture leads to the idea of the Dirichlet Process Mixture (DPM) model. For variables supported on the whole real space, like the individual heterogeneity z here, a typical choice of the kernel of $f\left(z;\theta\right)$ is a (multivariate) normal distribution with $\theta = (\mu, \Omega)$ being the mean and covariance matrix of the normal.

With component label k, component probability p_k, and component parameters (μ_k, Ω_k), one draw from the DPM prior can be written as an infinite location-scale mixture of (multivariate) normals,

$$z_i \sim \sum_{k=1}^{\infty} p_k N\left(\mu_k, \Omega_k\right). \tag{2.7}$$

Different draws from the DPM prior are characterized by different combinations of $\{p_k, \mu_k, \Omega_k\}$, and different combinations of $\{p_k, \mu_k, \Omega_k\}$ lead to different shapes of f. That is why the DPM

[17] See Appendix A for a formal definition of DP.

[18] Here and below, I suppress the superscript z in the parameters when there is no confusion.

prior is flexible enough to approximate many distributions. The component parameters (μ_k, Ω_k) are directly drawn from the DP base distribution G_0, which is chosen to be the conjugate multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z_i is a scalar). The component probability p_k is constructed via the stick-breaking process governed by the DP scale parameter α.

$$\begin{aligned}(\mu_k, \Omega_k) &\sim G_0, \\ p_k &\sim \zeta_k \prod_{j<k} (1-\zeta_j), \text{ where } \zeta_k \sim \text{Beta}(1, \alpha).\end{aligned} \tag{2.8}$$

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 conciseness, I denote the p_k part in equation (2.8) as $p_k \sim \text{SB}(1, \alpha)$, where the function name SB is the acronym for "stick-breaking", and the two arguments are from the parameters of the Beta distribution for "stick length" ζ_k's.

One virtue of the nonparametric Bayesian framework is its ability to flexibly elicit the tuning parameter from the data. Namely, we can set up a relatively flexible hyperprior for the DP scale parameter, $\alpha \sim \text{Ga}(a_0^\alpha, b_0^\alpha)$, 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 $E[K^*|\alpha] \approx \alpha \log\left(\frac{\alpha+N}{\alpha}\right)$, $Var[K^*|\alpha] \approx \alpha\left[\log\left(\frac{\alpha+N}{\alpha}\right) - 1\right]$.

2.3.2 Correlated Random Coefficients Model

To accommodate the correlated random coefficients model where the individual heterogeneity z_i $(= \lambda_i$ or $l_i)$ can be potentially correlated with the conditioning variables c_{i0}, it is necessary to consider a non-parametric 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 density to the corresponding ones of the joint density without explicitly modeling the marginal density of c_{i0}. 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, and extend it to the multivariate setup. Conditioning on c_{i0},

$$z_i|c_{i0} \sim \sum_{k=1}^{\infty} p_k(c_{i0}) N\left(\mu_k\left[1, c_{i0}'\right]', \Omega_k\right). \tag{2.9}$$

Similar to the DPM prior, the component parameters can be directly drawn from the base distribution, $\theta_k = (\mu_k, \Omega_k) \sim G_0$. G_0 is again specified as a conjugate form with a multivariate normal

distribution for vec (μ_k) and an inverse Wishart distribution for Ω_k (or a multivariate-normal-inverse-gamma distribution if z_i is a scalar). Now the mixture probabilities are characterized by the probit stick-breaking process

$$p_k(c_{i0}) = \Phi(\zeta_k(c_{i0})) \prod_{j<k} (1 - \Phi(\zeta_j(c_{i0}))), \tag{2.10}$$

where stochastic function ζ_k is drawn from the Gaussian process $\zeta_k \sim GP(0, V_k)$ for $k = 1, 2, \cdots$.[19]

This setup has three key features: (1) component means are linear in c_{i0}; (2) component covariance matrices are independent of c_{i0}; and (3) mixture probabilities are flexible functions of c_{i0}. This framework is general enough to accommodate a broad class of conditional distributions due to the flexibility in the mixture probabilities. Intuitively, the current setup is similar to approximating the conditional density via Bayes' theorem, but does not explicitly model the distribution of the conditioning variables c_{i0}, and thus allows for more relaxed assumptions on it.[20]

3 Theoretical Properties

3.1 Background

Generally speaking, a 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 ensure 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 the 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 counterexample 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

[19] See Appendix A for the definition of Gaussian process.
[20] See Appendix B.1 for a more detailed explanation.

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 time dimension T is fixed and 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 data generating process, and the density forecasts concentrate in an arbitrarily small region around the oracle.

3.2 Identification

To establish the posterior consistency argument, we first need to ensure identification of both the common parameters and the (conditional) distribution of individual effects. Here, I present the identification result in terms of the correlated random coefficients model with cross-sectional heteroskedasticity, where random coefficients and cross-sectional homoskedasticity can be viewed as special cases.

Assumption 3.1. *(Identification: General Model)*

1. *Model setup:*
   (a) *Conditional on* $w_{0:T}^A$, $\left(c_{i0}^*, \lambda_i, \sigma_i^2\right)$ *are i.i.d. across* i.
   (b) *For all* t, *conditional on* $(y_{it}, c_{i,t-1})$, x_{it}^{P*} *is independent of* $\left(\lambda_i, \sigma_i^2, \beta\right)$.
   (c) $\left(x_{i,0:T}^O, w_{i,0:T}\right)$ *are independent of* $\left(\lambda_i, \sigma_i^2, \beta\right)$.
   (d) *Conditioning on* c_{i0}, λ_i *and* σ_i^2 *are independent of each other.*
   (e) *Let* $u_{it} = \sigma_i v_{it}$. v_{it} *is i.i.d. across* i *and* t *and independent of* $c_{i,t-1}$.
2. *Identification:*
   (a) *The characteristic functions for* $\lambda_i|c_{i0}$ *and* $\sigma_i^2|c_{i0}$ *are non-vanishing almost everywhere.*[21]
   (b) *For all* i, $w_{i,0:T-1}$ *has full rank* d_w.
   (c) *After orthogonal forward differencing,*

$$\tilde{x}_{i,t-1} = x_{i,t-1} - w_{i,t-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}' \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} x_{i,s-1}.$$

   Then, the matrix $\mathbb{E}\left[\sum_t \tilde{x}_{i,t-1} \tilde{x}_{i,t-1}'\right]$ *is assumed to have full rank.*

Remark 3.2. (i) Condition 1-a characterizes the correlated random coefficients model, where there can be a potential correlation between the individual heterogeneity $\left(\lambda_i, \sigma_i^2\right)$ and the conditioning variables c_{i0}. Therefore, despite the conditional independence in condition 1-d, λ_i and σ_i^2 can potentially relate to each other through c_{i0}. For example, a young firm's initial performance may reveal its underlying ability and risk.

[21] This assumption can be relaxed based on Evdokimov and White (2012).

(ii) For the random coefficients case, condition 1-a can be altered to "$\left(\lambda_i, \sigma_i^2\right)$ are independent of c_{i0} and i.i.d. across i". Together with condition 1-d, it implies that $\left(\lambda_i, \sigma_i^2, c_{i0}\right)$ are independent among one another.
(iii) "v_{it} is i.i.d. across i and t" in condition 1-e implies that v_{it} is independent of $\left(\lambda_i, \sigma_i^2\right)$, because the individual effects λ_i and σ_i^2 are time invariant.
(iv) It is possible to allow some additional flexibility in the distribution of the shock u_{it}. For example, the identification argument still holds as long as (1) conditional on $c_{i,t-1}$, v_{it} is i.i.d. across i, (2) the distributions of v_{it}, $f_t^v\left(v_{it}|c_{i,t-1}\right)$, have known functional forms, such that $\mathbb{E}[v_{it}|c_{i,t-1}] = 0$, $\mathbb{V}[v_{it}|c_{i,t-1}] = 1$, and (3) the characteristic function for $v_{it}|c_{i,t-1}$ is non-vanishing almost everywhere. 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.

Proposition 3.3. *(Identification: General Model)*

Under Assumption 3.1, 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.

See Appendix B.2 for the proof. Assumption 3.1 and Proposition 3.3 are similar to Assumption 2.1-2.2 and Theorem 2.3 in Liu *et al.* (2017)[22] except for the treatment of cross-sectional heteroskedasticity with σ_i^2 being an unobserved random quantity (see the literature review for a more detailed comparison). The rank condition supports the posterior consistency of β via orthogonal forward differencing. [23] λ_i is additively separable from the shocks, so I follow the original proof based on the characteristic function (i.e. the Fourier transform). Note that unlike λ_i, σ_i^2 interacts with the shocks in a multiplicative way. The Fourier transform is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004) to deliver the identification of f^{σ^2}.

Example: Baseline Model For the baseline setup in equation (1.1), we can reduce Assumption 3.1 and establish the identification result based on a simpler set of assumptions as follows.

Assumption 3.4. *(Identification: Baseline Model)*

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

Intuitively speaking, taking young firm dynamics as the example, the second condition implies that skill is independent of shock (see Remark 3.2 (iii)) and that shock is independent across firms

[22] It is in turn based on early works such as Arellano and Bover (1995) and Arellano and Bonhomme (2012).

[23] The identification of common parameters in panel data models is standard in the literature. See textbooks and handbook chapters, Baltagi (1995), Arellano and Honoré (2001), Arellano (2003), and Hsiao (2014).

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 transform. The last condition guarantees that the time span is long enough to distinguish persistence $\beta y_{i,t-1}$ and individual effects λ_i.

Extension: Unbalanced Panels For the unbalanced panels with randomly omitted observations as specified in Subsection 2.1, we have:

Assumption 3.5. *(Identification: Unbalanced Panels) For all i,*

1. c_{i0} *is observed.*
2. x_{iT} *and* w_{iT} *are observed.*
3. $w_{i,(t_{0i}-1):(t_{1i}-1)}$ *has full rank* d_w.
4. *After orthogonal forward differencing,*

$$\tilde{x}_{i,t-1} = x_{i,t-1} - w'_{i,t-1}\left(\sum_{s=t+1}^{t_{1i}} w_{i,s-1}w'_{i,s-1}\right)^{-1}\sum_{s=t+1}^{t_{1i}} w_{i,s-1}x_{i,s-1}.$$

Then, the matrix $\mathbb{E}\big[\sum_{t=t_{0i}}^{t_{1i}} \tilde{x}_{i,t-1}\tilde{x}'_{i,t-1}\big]$ *is assumed to have full rank.*

The first condition guarantees the existence of the initial conditioning set for the correlated random coefficients model. The second condition ensures that the covariates in the forecast equation are available in order to make predictions. The third and fourth conditions are the unbalanced panel counterparts of Assumption 3.1 (2-b,c). They guarantee that the observed chain is long and informative enough to distinguish different aspects of common effects and individual effects. Now we can obtain similar identification results for unbalanced panels under Assumptions 3.1 (except 2-b,c) and 3.5.

3.3 Posterior Consistency

In this subsection, I establish the posterior consistency of the estimated common parameters ϑ and the estimated (conditional) distribution of individual effects f^z. Note that the estimated individual effects z_i are not consistent because information is accumulated only along the cross-sectional dimension but not along the time dimension.

3.3.1 Random Coefficients Model

In the random coefficients model, f^z is an unconditional distribution. Let Θ be the space of the parametric component ϑ, and let $\mathcal{F}^z$ be the set of densities on $\mathbb{R}^{d_z}$ (with respect to Lebesgue measure) as the space of the nonparametric component f^z. Hence, $\Theta = \mathbb{R}^{d_x} \times \mathbb{R}^+$, $f = f^\lambda$, and $\mathcal{F} = \mathcal{F}^\lambda$ in cross-sectional homoskedastic cases, and $\Theta = \mathbb{R}^{d_x}$, $f = f^\lambda f^l$, and $\mathcal{F} = \mathcal{F}^\lambda\mathcal{F}^l$ in cross-sectional heteroskedastic cases due to the independence between λ_i and σ_i^2. 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 $\Pi(\cdot,\cdot|D)$ with the marginal posteriors indicated with superscripts.

The posterior consistency results are established with respect to the strong topology on f, which is generated by the L_1-metric on integrable functions and is closely related to convergence of the probability distribution function (pdf). Note that the strong topology is stronger than the weak topology, and convergence of the pdf is stronger than convergence in distribution or weak convergence. It in turn implies the convergence of quantiles.

Definition 3.6. The posterior achieves *strong consistency* at (ϑ_0, f_0) if for any $\epsilon > 0$ and any $\delta > 0$, as $N \to \infty$,

$$\Pi\left(\left(\vartheta, f\right):\ \left\|\vartheta - \vartheta_0\right\|_2 < \delta,\ \left\|f - f_0\right\|_1 < \epsilon \middle| D\right) \to 1$$

in probability with respect to the true data generating process.

To give the intuition behind the posterior consistency argument, let us first consider a simpler scenario where we estimate the distribution of observables without the deconvolution and dynamic panel data structure. The following lemma restates Theorem 1 in Canale and De Blasi (2017). Note that space $\mathcal{F}$ is not compact, so we introduce a compact subset $\mathcal{F}_N$ (i.e. sieve) that asymptotically approximates $\mathcal{F}$ and then regularize the asymptotic behavior of $\mathcal{F}_N$ instead of $\mathcal{F}$.

Lemma 3.7. *(Canale and De Blasi, 2017)*

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

1. *Kullback-Leibler (KL) property: f_0 is in the KL support of Π, i.e. for all $\epsilon > 0$,*

$$\Pi\left(f \in \mathcal{F}:\ d_{KL}\left(f_0, f\right) < \epsilon\right) > 0,$$

   where $d_{KL}\left(f_0, f\right) = \int f_0 \log \frac{f_0}{f}$ is the KL divergence of f from f_0.

2. *Sieve property: There exists $\mathcal{F}_N \subset \mathcal{F}$ that can be partitioned as $\mathcal{F}_N = \cup_j \mathcal{F}_{N,j}$ such that, for any $\epsilon > 0$,*
   (a) *For some $\beta > 0$, $\Pi\left(\mathcal{F}_N^c\right) = O\left(\exp\left(-\beta N\right)\right)$.*
   (b) *For some $\gamma > 0$, $\sum_j \sqrt{\mathcal{N}\left(\epsilon, \mathcal{F}_{N,j}\right)\Pi\left(\mathcal{F}_{N,j}\right)} = o\left(\exp\left(\left(1-\gamma\right)N\epsilon^2\right)\right)$, where $\mathcal{N}\left(\epsilon, \mathcal{F}_{N,j}\right)$ is the covering number of $\mathcal{F}_{N,j}$ by balls with a radius ϵ.*[24]

By Bayes' Theorem, the posterior probability of the alternative region $U^c = \{f \in \mathcal{F}:\ \|f - f_0\|_1 \geq \epsilon\}$ can be expressed as the ratio on the right hand side,

$$\Pi\left(U^c | x_{1:N}\right) = \int_{U^c} \left[\int \prod_{i=1}^{N} \frac{f\left(x_i\right)}{f_0\left(x_i\right)} dx_i\right] d\Pi\left(f\right) \Bigg/ \int_{\mathcal{F}} \left[\int \prod_{i=1}^{N} \frac{f\left(x_i\right)}{f_0\left(x_i\right)} dx_i\right] d\Pi\left(f\right).$$

[24] As the covering number increases exponentially with the dimension of z, a direct adoption of Theorem 2 in Ghosal *et al.* (1999) would impose a strong tail restriction on the prior and exclude the case where the base distribution G_0 contains an inverse Wishart distribution for component variances. Hence, I follow the idea of Ghosal and van der Vaart (2007) and Canale and De Blasi (2017), where they relax the assumption on the coverage behavior by a summability condition of covering numbers weighted by their corresponding prior probabilities.

Intuitively speaking, for the numerator, the sieve property ensures that the sieve expands to the entire space and puts an asymptotic upper bound on the number of balls that cover the sieve. As the likelihood ratio is small in each covering ball, the integration over the alternative region is still sufficiently small. For the denominator, the KL property implies that the prior of distributions puts positive weight on the true distribution, so the likelihood ratio integrated over the whole space is large enough. Therefore, the posterior probability of the alternative region is arbitrarily small.

Lemma 3.7 establishes posterior consistency in a density estimation context. However, as mentioned in the introduction, there are a number of challenges in adapting to the dynamic panel data setting. The first challenge is, because we observe y_{it} rather than λ_i, to disentangle the uncertainty generated from unknown cross-sectional heterogeneity λ_i and from independent shocks u_{it}, i.e. a deconvolution problem.[25] Second is to incorporate an unknown shock size σ^2 in cross-sectional homoskedastic cases.[26] Third is to address unknown individual-specific shock sizes σ_i^2 in cross-sectional heteroskedastic cases. Fourth is to take care of strictly exogenous and predetermined variables (including lagged dependent variables) as covariates.

More specifically, in the dynamic panel data model, the posterior probability of the alternative region can be decomposed as

$$\begin{aligned}&\Pi\left(\left(\vartheta, f\right): \ \left\|\vartheta-\vartheta_0\right\|_2 \geq \delta \text{ or } \left\|f-f_0\right\|_1 \geq \epsilon \middle| \, D\right)\\ =&\Pi^{\vartheta}\left(\left\|\vartheta-\vartheta_0\right\|_2 \geq \delta \middle| \, D\right)+\Pi\left(\left\|f-f_0\right\|_1 \geq \epsilon, \left\|\vartheta-\vartheta_0\right\|_2<\delta \middle| \, D\right),\end{aligned}$$

and we want to show that the whole expression tends to zero as N goes to infinity. The first term is the marginal posterior probability of the finite dimensional common parameters. Its posterior consistency is relatively straightforward to obtain. When a candidate ϑ is far from the true ϑ_0, we can employ orthogonal forward differencing to get rid of λ_i (see Appendix B.2), and then apply the traditional posterior consistency argument to a linear regression of the "residues". The second term accounts for the infinite dimensional underlying distribution when ϑ approaches ϑ_0 but f is separated from f_0. The following two paragraphs outline the intuition behind this part of the proof. Basically, we can re-establish the two conditions in Lemma 3.7 by linking the distribution of the observables y_{it} with the underlying distribution of λ_i (and σ_i^2).

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 coefficients model) offers more regularities on the prior Π and

[25] Some previous 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 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 inferrable from the data. However, here the target λ_i intertwines with u_{it} and cannot be easily inferred from the observed y_{it}.

[26] 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.

thus weaker assumptions on the true distribution f_0 (see Assumptions 3.9 and 3.12). In terms of deconvolution,[27] the KL requirement is achieved through the convexity of the KL divergence. Intuitively, convolution with a common distribution would reduce the difference in f. In terms of the common parameters, the KL requirement is delivered via controlling the tail behavior of covariates, so the continuity at ϑ_0 is preserved after integrating out covariates.

The sieve property guarantees that the data are informative enough to differentiate the true distribution from the alternatives. It relies on both the DPM setup and the strong topology characterized by the L_1-norm, where the latter is closely linked to the identification argument. In terms of deconvolution, (de)convolution preserves the L_1-norm as well as the number of balls that cover the sieve. In terms of the common parameters, when ϑ is close to ϑ_0 but f is far from f_0, we want to make sure that the deviation generated from ϑ is small enough so that it cannot offset the difference in f.

Now let us formally state the assumptions and main theorem for random coefficients models. Appendix B.3 provides the complete proof.

Assumption 3.8. *(Covariates)*

1. $w_{i,0:T}^I$ *is bounded.*
2. *Let* $\tilde{c}_{i0} = \left(x_{i0}^P, x_{i,0:T-1}^O\right)$. *As* $C \to 0$, $\int_{\|\tilde{c}_{i0}\|_2 \geq C} \|\tilde{c}_{i0}\|_2^2 \, p\left(\tilde{c}_{i0} \left| c_{i0} \backslash \tilde{c}_{i0} \right.\right) d\tilde{c}_{i0} \to 0$.
3. *As* $C \to 0$, $\int_{\|x_{i,t-1}^{P*}\|_2 \geq C} \left\|x_{i,t-1}^{P*}\right\|_2^2 p\left(x_{i,t-1}^{P*} \left| y_{i,t-1}, c_{i,0:t-2}\right.\right) dx_{i,t-1}^{P*} \to 0$.

Considering that we have a regression model, the tail condition prevents a slight difference in β from obscuring the difference in f, which is satisfied when $x_{i,t-1}$ exhibits a tail behavior similar to λ_i (see Assumption 3.9 (1-e) below). The boundedness of $w_{i,0:T}^I$ serves the same purpose for the heterogenous term $\lambda_i' w_{i,t-1}$.

Assumption 3.9. *(Nonparametric Bayesian Prior: Random Coefficients)*

1. *Conditions on* f_0^z*:*
   (a) $f_0^z(z)$ *is a continuous density.*
   (b) *For some* $0 < M < \infty$, $0 < f_0^z(z) \leq M$ *for all* z.
   (c) $\left|\int f_0^z(z) \log f_0^z(z) \, dz\right| < \infty$.
   (d) $\int f_0^z(z) \log \frac{f_0^z(z)}{\varphi_\delta(z)} dz < \infty$, *where* $\varphi_\delta(z) = \inf_{\|z'-z\|_2 < \delta} f_0^z(z)$, *for some* $\delta > 0$.
   (e) *For some* $\eta > 0$, $\int \|z\|_2^{2(1+\eta)} f_0^z(z) \, dz < \infty$.
2. *Conditions on* $G_0^z(\mu, \Omega)$*:*
   (a) G_0^z *has full support on* $\mathbb{R}^{d_z} \times \mathcal{S}$, *where* $\mathcal{S}$ *is the space of* $d_z \times d_z$ *positive definite matrices with the spectral norm.*[28]
   (b) *For some* c_1, c_2, $c_3 > 0$, $r > (d_z - 1)/2$, *and* $\kappa > d_z(d_z - 1)$, *for sufficiently large* $x > 0$, $G_0^z(\|\mu\|_2 > x) = O\left(x^{-2(r+1)}\right)$, $G_0^z(\Lambda_1 > x) = O\left(\exp\left(-c_1 x^{c_2}\right)\right)$, $G_0^z\left(\Lambda_{d_z} < \frac{1}{x}\right) =$

[27] Here and below, "deconvolution" also refers to the multiplicative deconvolution for cross-sectional heteroskedasticity.

[28] The spectral norm is induced by the L_2-norm on vectors, $\|\Omega\|_2 = \max_{x \neq 0} \frac{\|\Omega x\|_2}{\|x\|_2}$.

$O\left(x^{-c_3}\right)$, $G_0^z\left(\frac{\Lambda_1}{\Lambda_{d_z}} > x\right) = O\left(x^{-\kappa}\right)$, *where* Λ_1 *and* Λ_{d_z} *are the largest and smallest eigenvalues of* Ω^{-1}*, respectively.*

First, the KL property is obtained based on conditions 1 and 2-a. Condition 1 ensures that the true distribution f_0 is well-behaved, and condition 2-a guarantees that the DPM prior is general enough to contain the true distribution. Then, condition 2-b further accounts for the sieve property. According to Corollary 1 in Canale and De Blasi (2017), condition 2-b holds for a multivariate-normal-inverse-Wishart base distribution G_0^z (or a normal-inverse-gamma base distribution if z is a scalar) as long as the degree of freedom of the inverse Wishart component $\nu_0^z > \max\{2d_z, (2d_z+1)(d_z-1)\}$, where $2d_z$ controls the tail behavior of component mean μ and $(2d_z+1)(d_z-1)$ regulates the eigenvalue structure of component variance Ω.

Proposition 3.10. *(Consistency: Random Coefficients)*

Suppose we have:

1. *Model: The random effects version of Assumption 3.1.*[29]
2. *Covariates:* $(x_{i,0:T}, w_{i,0:T})$ *satisfies Assumption 3.8.*
3. *Common parameters:* ϑ_0 *is in the interior of supp* $\left(\Pi^{\vartheta}\right)$.
4. *Distributions of individual effects:*
   (a) f_0^z *and* Π^z *satisfy Assumption 3.9.*
   (b) *For cross-sectional heteroskedastic models, supp* $\left(f_0^{\sigma^2}\right)$ *is bounded above by some large* $\bar{\sigma}^2 > 0$.

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

3.3.2 Correlated Random Coefficients Model

For the correlated random coefficients model, the definitions and notations are parallel with those in the random coefficients model with a slight adjustment considering that f is now a conditional distribution. As in Pati *et al.* (2013), it is helpful to link the properties of the conditional density to the corresponding ones of the joint density without explicitly modeling the marginal density of c_{i0}, which circumvents the difficulty associated with an uncountable set of conditional densities.

Let $q_0(c_0)$ be the true marginal density of the conditioning variables. Then, the induced q_0-integrated L_1-distance is defined as $\|f - f_0\|_1 \equiv \int \left[\int |f(z|c_0) - f_0(z|c_0)|\, dz\right] q_0(c_0)\, dc_0$, and the induced q_0-integrated KL divergence is $d_{KL}(f_0, f) \equiv \int \left[\int f_0(z|c_0) \log \frac{f_0(z|c_0)}{f(z|c_0)} dz\right] q_0(c_0)\, dc_0$. Note that in both definitions, the conditioning variables c_0 are integrated out with respect to the true q_0, so this setup does not require estimating q_0 and thus relaxes the assumption on the conditioning set.[30]

[29] Or Assumptions 3.1 (except 2-b,c) and 3.5 for unbalanced panels.

[30] Denote the joint densities $\tilde{f}_0(z, c_0) = f_0(z|c_0) \cdot q_0(c_0)$, $\tilde{f}(z, c_0) = f(z|c_0) \cdot q_0(c_0)$, where $\tilde{f}$ and $\tilde{f}_0$ share the same marginal density q_0, but different conditional densities f and f_0. Then, the induced q_0-integrated L_1-distance/KL divergence of f with respect to f_0 equals to the L_1-distance/KL divergence of $\tilde{f}$ with respect to $\tilde{f}_0$.

The main assumptions and theorem for correlated random coefficients models are stated as follows.

Assumption 3.11. *(Conditioning set)*

Let $\mathcal{C}$ be the support of c_{i0}. $\mathcal{C}$ is a compact subset of $\mathbb{R}^{d_{c_0}}$, and $q_0(c_0) > 0$ for all $c_0 \in \mathcal{C}$.

The compactness fosters uniform behavior along the conditioning variables. This assumption is stronger than Assumption 3.8 (1 and 2) for random coefficients models.

Assumption 3.12. *(Nonparametric Bayesian Prior: Correlated Random Coefficients)*

1. *Conditions on f_0^z:*
   (a) *For some $0 < M < \infty$, $0 < f_0^z(z|c_0) \leq M$ for all (z, c_0).*
   (b) *$\left|\int \left[\int f_0^z(z|c_0) \log f_0^z(z|c_0)\, dz\right] q_0(c_0)\, dc_0\right| < \infty$.*
   (c) *$\int \left[\int f_0^z(z|c_0) \log \frac{f_0^z(z|c_0)}{\varphi_\delta(z|c_0)} dz\right] q_0(c_0)\, dc_0 < \infty$, where $\varphi_\delta(z|c_0) = \inf_{\|z'-z\|_2<\delta} f_0^z(z|c_0)$, for some $\delta > 0$.*
   (d) *For some $\eta > 0$, $\int \left[\int \|z\|_2^{2(1+\eta)} f_0^z(z|c_0)\, dz\right] q_0(c_0)\, dc_0 < \infty$.*
   (e) *$f_0^z(\cdot|\cdot)$ is jointly continuous in (z, c_0).*
2. *Conditions on G_0^z: Let $d_\mu = d_z(d_{c_0} + 1)$ be the dimension of $vec(\mu)$.*
   (a) *G_0^z has full support on $\mathbb{R}^{d_\mu} \times \mathcal{S}$.*
   (b) *G_0^z is absolutely continuous.*
   (c) *For some c_1, c_2, $c_3 > 0$, $r > (d_\mu - 1)/2$, and $\kappa > d_z(d_z - 1)$, for sufficiently large $x > 0$, $G_0^z(\|\text{vec}(\mu)\|_2 > x) = O\left(x^{-2(r+1)}\right)$, $G_0^z(\Lambda_1 > x) = O\left(\exp\left(-c_1 x^{c_2}\right)\right)$, $G_0^z\left(\Lambda_{d_z} < \frac{1}{x}\right) = O\left(x^{-c_3}\right)$, $G_0^z\left(\frac{\Lambda_1}{\Lambda_{d_z}} > x\right) = O\left(x^{-\kappa}\right)$.*
3. *Conditions on the stick-breaking process: $V_k^z(c, \tilde{c}) = \tau^z \exp\left(-A_k^z \|c - \tilde{c}\|_2^2\right)$, $\tau^z > 0$ fixed.*
   (a) *The prior for A_k^z has full support on $\mathbb{R}^+$.*
   (b) *There exist β, $\gamma > 0$ and a sequence $\delta_N = O\left(N^{-5/2}(\log N)^2\right)$ such that $\mathbb{P}(A_k^z > \delta_N) \leq \exp\left(-N^\beta h^{(\beta+2)/\gamma} \log h\right)$.*
   (c) *There exists an increasing sequence $r_N \to \infty$ and $(r_N)^{d_{c0}} = o\left(N^{1-\gamma}(\log N)^{-(d_{c_0}+1)}\right)$ such that $\mathbb{P}(A_k^z > r_N) \leq \exp(-N)$.*

These conditions build on Pati *et al.* (2013) for posterior consistency under the conditional density topology and further extend it to multivariate conditional density estimation with infinite location-scale mixtures. The conditions on f_0^z and G_0^z can be viewed as conditional density analogs of the conditions in Assumption 3.9. Conditions 1, 2-a,b, and 3-a ensure the KL property, and conditions 2-c and 3-b,c address the sieve property. For G_0^z with a multivariate normal distribution on $\text{vec}(\mu)$ and an inverse Wishart distribution on Ω (or an inverse gamma distribution if z is a scalar), the tail condition on $\text{vec}(\mu)$ automatically holds, and Corollary 1 in Canale and De Blasi (2017) is satisfied as long as the degree of freedom of the inverse Wishart component $\nu_0^z > (2d_z + 1)(d_z - 1)$.

Proposition 3.13. *(Consistency: Correlated Random Coefficients)*

Suppose we have:

1. *Model: Assumption 3.1.*[31]
2. *Covariates:* $(x_{i,0:T}, w_{i,0:T})$ *satisfy Assumption 3.8 (3) and Assumption 3.11.*
3. *Common parameters:* ϑ_0 *is in the interior of supp* $\left(\Pi^{\vartheta}\right)$.
4. *Distributions of individual effects:*
   (a) f_0^z *and* Π^z *satisfy Assumption 3.12.*
   (b) *For cross-sectional heteroskedastic models, supp* $\left(f_0^{\sigma^2}\right)$ *is bounded above by some large* $\bar{\sigma}^2 > 0$.

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

The proof in Appendix B.4 parallels the random effects case except that now both the KL property and the sieve property are constructed on the q_0-induced measure.

3.4 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 ϑ and f, and then show the asymptotical convergence of the density forecasts to the oracle forecast (see Appendix B.5 for the detailed proof).

Proposition 3.14. *(Density Forecasts)*

Suppose we have:

1. *For random coefficients models, conditions in Proposition 3.10.*
2. *For correlated random coefficients models,*
   (a) *Conditions in Proposition 3.13.*
   (b) *There exists* $\underline{q} > 0$ *such that* $|q_0(c_0)| > \underline{q}$ *for all* $c_0 \in \mathcal{C}$.
3. *In addition, for cross-sectional heteroskedastic models, supp* $\left(f_0^{\sigma^2}\right)$ *is bounded below by some* $\underline{\sigma}^2 > 0$.

Then, for any i *and any* $\epsilon > 0$, *as* $N \to \infty$,

$$\mathbb{P}\left(\left\|f_{i,T+1}^{cond} - f_{i,T+1}^{oracle}\right\|_1 < \epsilon \middle| D\right) \to 1.$$

Remark 3.15. (i) Requirement 3 ensures that the likelihood would not explode in cross-sectional heteroskedastic models.
(ii) The asymptotic convergence of aggregate-level density forecasts can be derived by summing individual-specific forecasts over different subcategories.

[31] Or Assumptions 3.1 (except 2-b,c) and 3.5 for unbalanced panels.

4 Numerical Implementation

In this section, I propose a posterior sampling procedure for the general panel data model introduced in Subsection 2.1. The nonparametric Bayesian prior is specified in Subsection 2.3 and enjoys desirable theoretical properties as discussed in Section 3.[32]

4.1 Random Coefficients Model

For the random coefficients 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 (see Algorithm C.4 in the Appendix). The estimates and forecasts for the two samplers are almost indistinguishable, 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 component probabilities are constructed via a truncated stick-breaking process governed by the DP scale parameter α.

$$p_k \begin{cases} \sim \mathrm{SB}\left(1, \alpha\right), & k < K, \\ = 1 - \sum_{j=1}^{K-1} p_j, & k = K. \end{cases}$$

Note that due to the truncation approximation, the probability for component K is different from its infinite mixture counterpart in equation (2.8). Resembling the infinite mixture case, I denote the above truncated stick-breaking process as $p_k \sim \mathrm{TSB}\left(1, \alpha, K\right)$, where TSB stands for "truncated stick-breaking", the first two arguments are from the parameters of the Beta distribution, and the last argument is the truncated number of components.

Below, the algorithms are stated for cross-sectional heteroskedastic models, while the adjustments for cross-sectional homoskedastic scenarios are discussed in Remark 4.2 (ii). For individual heterogeneity $z = \lambda, l$, let γ_i^z be individual i's component affiliation, which can take values $\{1, \cdots, K^z\}$, J_k^z be the set of individuals in component k, i.e. $J_k^z = \{i : \gamma_i^z = k\}$, and n_k^z be the number of individuals in component k, i.e. $n_k = \#J_k$. Then, the (data-augmented) joint posterior for the model

[32] 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 C.1 for details of the baseline model with random effects.

parameters is given by

$$
\begin{aligned}
& p\left(\left\{\alpha^z, \left\{p_k^z, \mu_k^z, \Omega_k^z\right\}, \left\{\gamma_i^z, z_i\right\}\right\}, \beta \middle| D\right) \\
& = \prod_{i,t} p\left(y_{it} \middle| \lambda_i, l_i, \beta, w_{i,t-1}, x_{i,t-1}\right) \cdot \prod_{z,i} p\left(z_i \middle| \mu_{\gamma_i^z}^z, \Omega_{\gamma_i^z}^z\right) p\left(\gamma_i^z \middle| \left\{p_k^z\right\}\right) \\
& \quad \cdot \prod_{z,k} p\left(\mu_k^z, \Omega_k^z\right) p\left(p_k^z \middle| \alpha^z\right) \cdot p\left(\alpha^z\right) \cdot p\left(\beta\right),
\end{aligned} \tag{4.1}
$$

where $z = \lambda, l$, $k = 1, \cdots, K^z$, $i = 1, \cdots N$, and $t = 1, \cdots, T$. The first block links observations to model parameters $\{\lambda_i, l_i\}$ and β. The second block links the individual heterogeneity z_i to the underlying distribution f^z. The last block formulates the prior belief on (β, f).

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

Note that if the individual heterogeneity z_i were known, only step 5 would be sufficient to recover the common parameters. If the mixture structure of f^z were known (i.e. if $(p_k^z, \mu_k^z, \Omega_k^z)$ for all components were known), only steps 3 to 5 would be needed to first assign individuals to components and then infer z_i based on the specific component that individual i has been assigned to. In reality, neither z_i nor its distribution f^z is known, so I incorporate two more steps 1 and 2 to model the underlying distribution f^z.

Algorithm 4.1. *(Random Coefficients with Cross-sectional Heteroskedasticity)*[33]

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

1. *Component probabilities: For* $z = \lambda, l$,

   (a) *Draw* $\alpha^{z(s)}$ *from a gamma distribution* $p\left(\alpha^{z(s)} \middle| p_{K^z}^{z(s-1)}\right)$*:*

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

   (b) *For* $k = 1, \cdots, K^z$*, draw* $p_k^{z(s)}$ *from the truncated stick-breaking process* $p\left(\left\{p_k^{z(s)}\right\} \middle| \alpha^{z(s)}, \left\{n_k^{z(s-1)}\right\}\right)$*:*

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

2. *Component parameters: For* $z = \lambda, l$*, for* $k = 1, \cdots, K^z$*, draw* $\left(\mu_k^{z(s)}, \Omega_k^{z(s)}\right)$ *from a multivariate-*

[33] 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 C.3.

normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a scalar) $p\left(\mu_k^{z(s)}, \Omega_k^{z(s)} \left| \left\{ z_i^{(s-1)} \right\}_{i \in J_k^{z(s-1)}} \right.\right)$.

3. *Component memberships: For $z = \lambda, l$, for $i = 1, \cdots N$, draw $\gamma_i^{z(s)}$ from a multinomial distribution $p\left(\left\{\gamma_i^{z(s)}\right\} \left| \left\{ p_k^{z(s)}, \mu_k^{z(s)}, \Omega_k^{z(s)} \right\}, z_i^{(s-1)} \right.\right)$:*

$$\gamma_i^{z(s)} = k, \text{ with probability } p_{ik} \propto p_k^{z(s)} \phi\left(z_i^{(s-1)};\ \mu_k^{z(s)}, \Omega_k^{z(s)}\right), \quad \sum_{k=1}^{K^z} p_{ik} = 1.$$

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 $l_i^{(s)}$ via the random-walk Metropolis-Hastings approach,*

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

   where $\sigma^2(l) = \frac{\bar{\sigma}^2 - \underline{\sigma}^2}{1 + \bar{\sigma}^2 \exp(-l)} + \underline{\sigma}^2$. Then, calculate $\left(\sigma_i^2\right)^{(s)}$ based on $\sigma^2(l)$.
5. *Common parameters: Draw $\beta^{(s)}$ from a linear regression model with "known" variance $p\left(\beta^{(s)} \left| \left\{ \lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)} \right\}, D \right.\right)$.*

Remark 4.2. (i) With the above prior specification, all steps enjoy closed-form conditional posterior distributions except step 4-b for σ_i^2, which does not exhibit a well-known density form. Hence, I resort to the random-walk Metropolis-Hastings algorithm to sample σ_i^2. 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 keeps 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. See Algorithm C.1 in the Appendix for details.
(ii) In cross-sectional homoskedastic cases, the algorithm would need the following changes: (1) in steps 1 to 4, only λ_i is considered, and (2) in step 5, $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ are drawn from a linear regression model with "unknown" variance, $p\left(\beta^{(s)}, \sigma^{2(s)} \left| \left\{ \lambda_i^{(s)} \right\}, D \right.\right)$.

4.2 Correlated Random Coefficients Model

To account for the conditional structure in the correlated random coefficients model, I implement a multivariate MGLR_x prior as specified in Subsection 2.3, which can be viewed as the conditional

counterpart of the Gaussian-mixture prior. The conditioning set c_{i0} is characterized by equation (2.2) for balanced panels or equation (2.3) for unbalanced panels.

The major computational difference from the random coefficients model in the previous subsection is that now the component probabilities become flexible functions of c_{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

$$V_k(c, \tilde{c}) = \exp\left(-A_k \|c - \tilde{c}\|_2^2\right),$$

where $A_k = C_k B_k$. Let $\eta = d_{c_0} + 1$, then B_k^η follows the standard exponential distribution, i.e. $p\left(B_k^\eta\right) = \exp\left(-B_k^\eta\right)$, and $C_k = k^{-2(3\eta+2)} (\log k)^{-1/\eta}$. This prior structure satisfies Pati *et al.* (2013) Remark 5.12 that ensures strong consistency.[34] Furthermore, it is helpful to introduce a set of auxiliary stochastic functions $\xi_k(c_{i0})$, $k = 1, 2, \cdots$, such that

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

Note that the probit stick-breaking process defined in equation (2.10) can be recovered by marginalizing over $\{\xi_k(y_{i0})\}$. 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 $\boldsymbol{\zeta}_k = [\zeta_k(c_{10}), \zeta_k(c_{20}), \cdots, \zeta_k(c_{N0})]'$ and $\boldsymbol{\xi}_k = [\xi_k(c_{10}), \xi_k(c_{20}), \cdots, \xi_k(c_{N0})]'$, as well as an $N \times N$ matrix $\boldsymbol{V}_k$ with the ij-th element being $(\boldsymbol{V}_k)_{ij} = \exp\left(-A_k \|c_{i0} - c_{j0}\|_2^2\right)$. The next algorithm extends Algorithm 4.1 to the correlated random coefficients scenario. Step 1 for component probabilities has been changed, while the rest of the steps are in line with those in Algorithm 4.1.

Algorithm 4.3. *(Correlated Random Coefficients with Cross-sectional Heteroskedasticity)*[35]

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

1. *Component probabilities: For* $z = \lambda, l$,
   (a) *For* $k = 1, \cdots, K^z - 1$, *draw* $A_k^{z(s)}$ *via the random-walk Metropolis-Hastings approach,*[36]

$$\begin{aligned}
p\left(A_k^{z(s)} \middle| \boldsymbol{\zeta}_k^{z(s-1)}, \{y_{i0}\}\right) \propto &\left(A_k^{z(s)}\right)^{\eta-1} \exp\left(-\left(\frac{A_k^{z(s)}}{C_k}\right)^\eta\right) \\
&\cdot \phi\left(\boldsymbol{\zeta}_k^{z(s-1)};\ 0, \exp\left(-A_k^{z(s)} \|c_{i0} - c_{j0}\|_2^2\right)\right).
\end{aligned}$$

[34] Their p is the d_{c_0} in the notation of this paper, and their d, η_1, and η can be constructed as $d_{c_0} + 1$, $\frac{d_{c_0}}{d_{c_0}+1}$, and $\frac{1}{2\left(d_{c_0}+1\right)}$, respectively.

[35] See Remark 4.2 (ii) for cross-sectional homoskedastic models.

[36] The first term comes from the change of variables from B_k^η to A_k.

Then, calculate $\boldsymbol{V}_k^{z(s)}$, *where* $\left(\boldsymbol{V}_k^{z(s)}\right)_{ij} = \exp\left(-A_k^{z(s)} \left\| c_{i0} - c_{j0} \right\|_2^2\right)$.

(b) *For* $k = 1, \cdots, K^z - 1$, *and* $i = 1, \cdots, N$, *draw* $\xi_k^{z(s)}(c_{i0})$ *from a truncated normal distribution* $p\left(\xi_k^{z(s)}(c_{i0}) \left| \zeta_k^{z(s-1)}(c_{i0}), \gamma_i^{z(s-1)} \right.\right)$:[37]

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

(c) *For* $k = 1, \cdots, K^z - 1$, $\boldsymbol{\zeta}_k^{z(s)}$ *from a multivariate normal distribution* $p\left(\boldsymbol{\zeta}_k^{z(s)} \left| \boldsymbol{V}_k^{z(s)}, \boldsymbol{\xi}_k^{z(s)} \right.\right)$:

$$\boldsymbol{\zeta}_k^{z(s)} \sim N\left(m_k^{\zeta}, \Sigma_k^{\zeta}\right), \text{ where } \Sigma_k^{\zeta} = \left[\left(\boldsymbol{V}_k^{z(s)}\right)^{-1} + I_N\right]^{-1} \text{ and } m_k^{\zeta} = \Sigma_k^{\zeta} \boldsymbol{\xi}_k^{z(s)}.$$

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

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

2. *Component parameters: For* $z = \lambda, l$, *for* $k = 1, \cdots, K^z$,
   (a) *Draw* $vec\left(\mu_k^{z(s)}\right)$ *from a multivariate normal distribution* $p\left(\mu_k^{z(s)} \left| \Omega_k^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_k^{z(s-1)}} \right.\right)$.
   (b) *Draw* $\Omega_k^{z(s)}$ *from an inverse Wishart distribution (or an inverse gamma distribution if* z *is a scalar)* $p\left(\Omega_k^{z(s)} \left| \mu_k^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_k^{z(s-1)}} \right.\right)$.
3. *Component memberships: For* $z = \lambda, l$, *for* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \left| \left\{p_k^{z(s)}, \mu_k^{z(s)}, \Omega_k^{z(s)}\right\}, z_i^{(s-1)}, c_{i0} \right.\right)$:

$$\gamma_i^{z(s)} = k, \text{ with probability } p_{ik} \propto p_k^{z(s)}(c_{i0}) \, \phi\left(z_i^{(s-1)};\ \mu_k^{z(s)}\left[1, c_{i0}'\right]', \Omega_k^{z(s)}\right), \quad \sum_{k=1}^{K^z} p_{ik} = 1.$$

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* $l_i^{(s)}$ *via the random-walk Metropolis-Hastings approach* $p\left(l_i^{(s)} \left| \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A \right.\right)$, *then calculate* $\left(\sigma_i^2\right)^{(s)}$ *based on* $\sigma^2(l)$.
5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model with "known" variance*

[37] $\mathbf{1}(\cdot)$ is an indicator function.

$$p\left(\beta^{(s)} \left| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right.\right).$$

5 Simulation

In this section, I have conducted extensive Monte Carlo simulation experiments to examine the numerical performance of the proposed semiparametric Bayesian predictor. Subsection 5.1 describes the evaluation criteria for point forecasts and density forecasts. Subsection 5.2 introduces other alternative predictors. Subsection 5.3 considers the baseline setup with random effects. Subsection 5.4 extends to the general setup incorporating cross-sectional heterogeneity and correlated random coefficients.

5.1 Forecast Evaluation Methods

As mentioned in the model setup in Subsection 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 corresponds to 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,

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

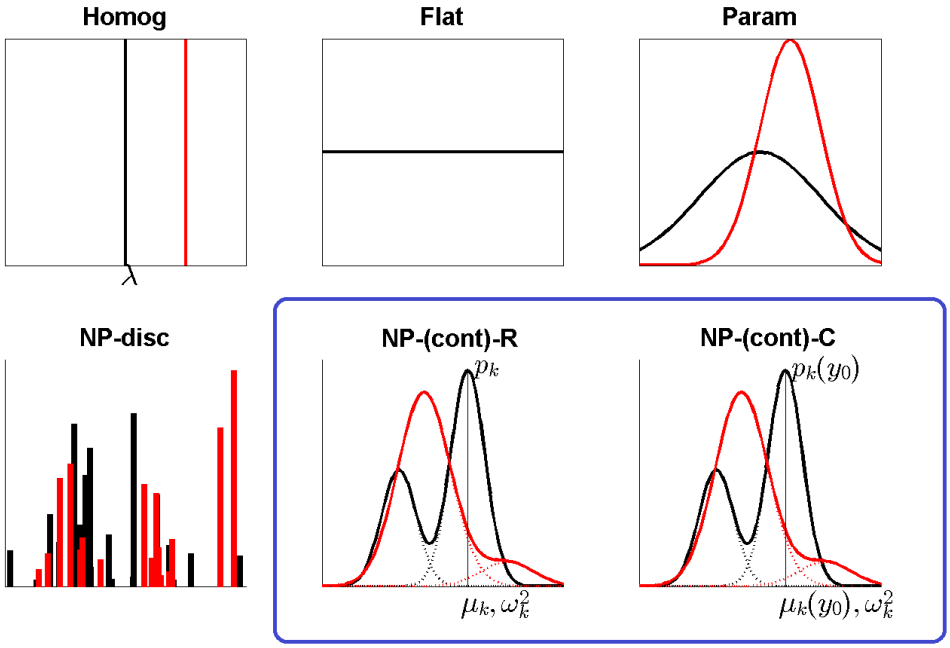
The Diebold and Mariano (1995) test is further implemented to assess whether the difference in the MSE is significant.

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. In addition, $\exp\left(LPS_A - LPS_B\right)$

Figure 5.1: Alternative Predictors



The black and red lines represent two draws from each prior.

gives the odds of future realizations based on predictor A versus predictor B. I also perform the Amisano and Giacomini (2007) test to examine the significance in the LPS difference.

5.2 Alternative Predictors

In the simulation experiments, I compare the proposed semiparametric Bayesian predictor with alternatives. Different predictors can be interpreted as different priors on the distribution of λ_i. As these priors are distributions over distributions, Figure 5.1 plots two draws from each prior – one in red and the other in black.[38]

The homogeneous prior (Homog) implies an extreme kind of pooling, which assumes that all firms share the same level of skill λ^*. It can be viewed as a Bayesian counterpart of the pooled OLS estimator. Because λ^* is unknown beforehand, the corresponding subgraph plots two vertical lines representing two degenerate distributions with different locations. More rigorously, this prior is defined as $\lambda_i \sim \delta_{\lambda^*}$, where δ_{λ^*} is the Dirac delta function representing a degenerate distribution $\mathbb{P}(\lambda_i = \lambda^*) = 1$. The unknown λ^* becomes another common parameter, similar to β, so I adopt a multivariate-normal-inverse-gamma prior on $\left([\beta, \lambda^*]', \sigma^2\right)$.

The flat prior (Flat) is specified as $p(\lambda_i) \propto 1$, an uninformative prior with the posterior mode being the MLE estimate. Roughly speaking, given the common parameters, there is no pooling from

[38] For easier illustration, here I consider the baseline model with univariate λ_i and homoskedasticity.

the cross-section, so we learn firm i's skill λ_i only using its own history.

The parametric prior (Param) pools the information from the cross-section via a parametric skill distribution, such as a Gaussian distribution with unknown mean and variance. The corresponding subgraph contains two curves with different means and variances. More explicitly, we have $\lambda_i \sim N\left(\mu, \omega^2\right)$, where a normal-inverse-gamma hyperprior is further imposed on $\left(\mu, \omega^2\right)$. This prior can be thought of as a limit case of the DPM prior when the scale parameter $\alpha \to 0$, so there is only one component, and $\left(\mu, \omega^2\right)$ are directly drawn from the base distribution G_0. The choice of the 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).

The nonparametric discrete prior (NP-disc) is modeled by a DP where λ_i follows a flexible nonparametric distribution, but on a discrete support. This paper focuses on continuous f, which may be more sensible for the skill of young firms as well as other similar empirical studies. In this sense, the NP-disc predictor helps examine how much can be gained or lost from the continuity assumption and from the additional layer of mixture.

In addition, NP-R denotes the proposed nonparametric prior for random effects/coefficients models, and NP-C for correlated random effects/coefficients models. Both of them are flexible priors on continuous distributions, and NP-C allows λ_i to depend on the initial condition of the firms.

The nonparametric predictors would reduce the estimation bias due to their flexibility while increasing the estimation variance due to their complexity. It is not transparent ex-ante which predictor performs better – the parsimonious parametric ones or the flexible nonparametric ones. Therefore, it is worthwhile to implement the Monte Carlo experiments and assess which predictor produces more accurate forecasts under which circumstances.

5.3 Baseline Model

Let us first consider the baseline model with random effects. The specifications are summarized in Table 5.1.

β_0 is set to 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 $\sigma_0^2/\mathbb{V}\left(\lambda_i\right) = 1/4$. The initial condition y_{i0} is drawn from a standard normal distribution, which complies with the tail condition in Assumption 3.8. Choices of $N = 1000$ and $T = 6$ are comparable with the young firm dynamics application.

There are four experiments with different true distributions of λ_i, $f_0\left(\cdot\right)$. As this subsection focuses on the simplest baseline model with random effects, λ_i is independent of y_{i0} in all four experiments. The first experiment features a degenerate λ_i distribution, where all firms enjoy the same skill level. Note that it does not satisfy Assumption 3.9 (1-a), which requires the true λ_i distribution to be continuous. The purpose of this distribution is to learn how poorly things can

Table 5.1: 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	$y_{i0} \sim N\left(0, 1\right)$
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	$\lambda_i \sim \left(0.35N\left(0, 1\right) + 0.65N\left(10, 1\right)\right) / \sqrt{1 + 10^2 \cdot 0.35 \cdot 0.65}$

go under the misspecification that the true λ_i distribution is completely off from the prior support. The second and third experiments are based on skewed and fat tail distributions, which reflect more realistic scenarios in empirical studies. The last experiment portrays a bimodal distribution with asymmetric weights on the two components.

I simulate 1,000 panel datasets for each setup and report the average statistics of these 1,000 repetitions. Forecasting performance, especially the relative rankings and magnitudes, is highly stable across repetitions. In each repetition, I generate 40,000 MCMC draws with the first 20,000 being discarded as burn-in. Based on graphical and statistical tests, the MCMC draws appear to converge to a stationary distribution. Both the Brook-Draper diagnostic and the Raftery-Lewis diagnostic yield desirable MCMC accuracy. See Figures D.1 to D.4 for trace plots, prior/posterior distributions, rolling means, and autocorrelation graphs of β, σ^2, α, and λ_1.

Table 5.2 shows the forecasting comparison among alternative predictors. For each experiment, 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 far exceeds all other competitors. 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 for the correlated random effects model)

Table 5.2: 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***

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. Both performance statistics are further averaged over 1,000 Monte Carlo samples. For the oracle predictor, the table reports the exact values of MSE and LPS*N (i.e. 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.

promotes better forecasts. In the empirical application to young firm dynamics, it would be more reasonable to assume continuous distributions of individual heterogeneity 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 6.2).

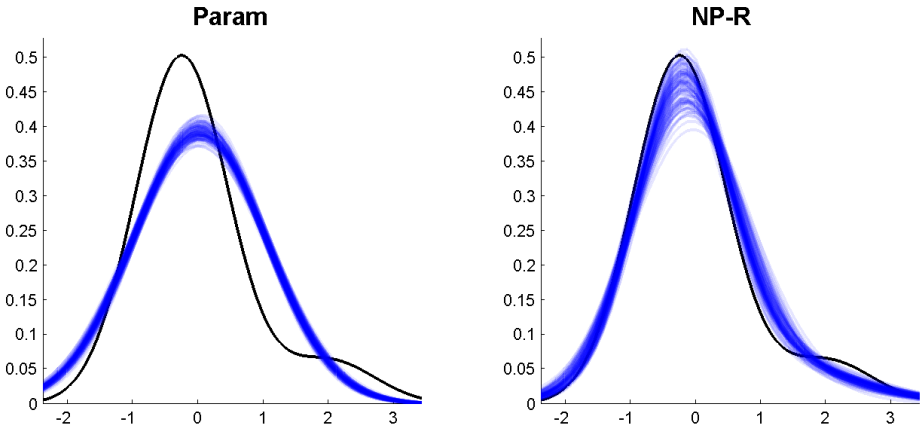
To investigate why we obtain better forecasts, Figure 5.2 demonstrates the posterior distribution of the λ_i distribution (i.e. a distribution over distributions) for experiments Skewed, Fat Tail, and Bimodal. In each case, the subgraphs are constructed from the estimation results of one of the 1,000 repetitions, with the left subgraph given by the Param estimator and the right one 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, $\Pi\left(f \,|y_{1:N,0:T}\right)$.

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 nicely captures the M-shape. 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 revealed in Proposition 3.14.

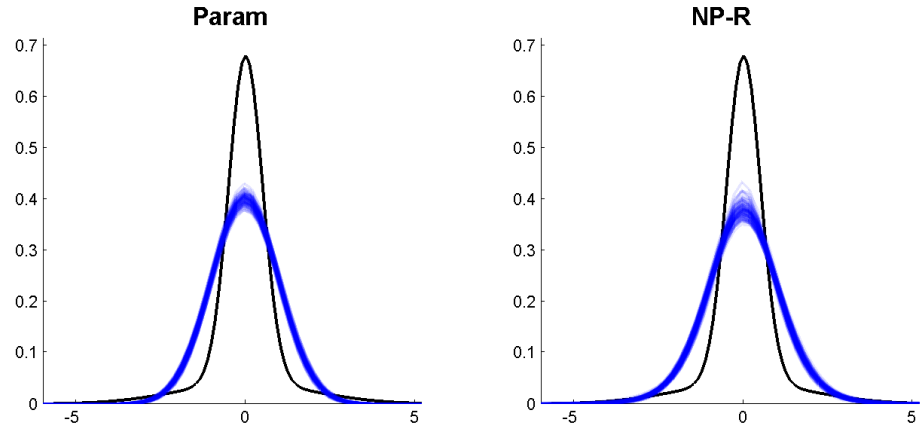
I have also considered various robustness checks. In terms of the setup, I have run different cross-sectional dimensions $N = 100,\ 500,\ 1000,\ 10^5$, different time spans $T = 6,\ 10,\ 20,\ 50$, different persistences $\beta = 0.2,\ 0.5,\ 0.8,\ 0.95$, different sizes of the i.i.d. shocks $\sigma^2 = 1/4$ and 1 (affecting the signal-to-noise ratio), and different underlying λ_i distributions (including standard normal). In

Figure 5.2: f_0 vs $\Pi(f\,|y_{1:N,0:T})$: Baseline Model

(a) Skewed



(b) Fat Tail



(c) Bimodal

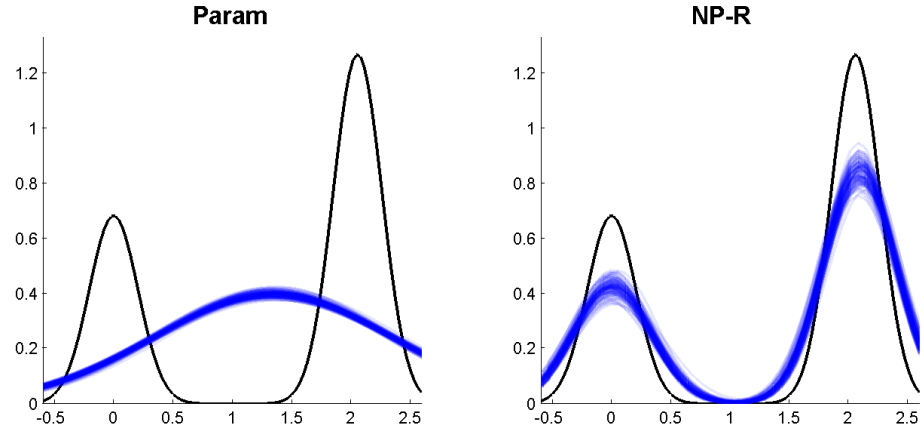


Table 5.3: Simulation Setup: General Model

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 coefficients	$\lambda_i \vert y_{i0} \sim e^{-2y_{i0}} N\left(y_{i0} v, 0.1^2 vv'\right) + \left(1 - e^{-2y_{i0}}\right) N\left(y_{i0}^4 v, 0.2^2 vv'\right)$, where $v = [1, 2, -1]'$
Cross-sectional heteroskedasticity	$\sigma_i^2 \vert y_{i0} \sim \left[0.454 \left(y_{i0} + 0.5\right)^2 \cdot \text{IG}\left(51, 50\right) + 0.05\right] \cdot \mathbf{1}\left(\sigma_i^2 \leq 10^6\right)$
Sample size	$N = 1000,\ T = 6$

general, the NP-R predictor is the overall best for density forecasts 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 but is only marginally better than the NP-R predictor. Intuitively, in the language of young firm dynamics, 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 uncertainty due to heterogenous skill. That is, NP-R works better than the alternatives 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. Furthermore, as the cross-sectional dimension N increases, the blue band in Figure 5.2 gets closer to the true f_0 and eventually completely overlaps it (see Figure D.5), 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 a power law tail for clustering behaviors, as well as DPM with skew normal components which better accommodates asymmetric data generating processes. They provide some improvement in the corresponding situations, but call for extra computation efforts.

5.4 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 5.3.

In terms of multidimensional individual heterogeneity, λ_i is now 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

Table 5.4: Prior Structures

Predictor		λ_i prior	σ_i^2 prior
Heterosk	NP-C	$f^{\lambda} \sim \text{MGLR}_{\text{x}}$	$f^{l} \sim \text{MGLR}_{\text{x}}$
Homog		Point mass	Point mass
Homosk	NP-C	$f^{\lambda} \sim \text{MGLR}_{\text{x}}$	Point mass
Heterosk	Flat	Uninformative	Uninformative
	Param	N	IG
	NP-disc	$f^{\lambda} \sim \text{DP}$	$f^{l} \sim \text{DP}$
	NP-R	$f^{\lambda} \sim \text{DPM}$	$f^{l} \sim \text{DPM}$

See Appendix C.5 for details of Heterosk-Param.

such conditional distribution exhibits rapid changes in its shape, which considerably restricts local sample size. The original conditional distribution in their papers is one-dimensional, and I expand it to accommodate the three-dimensional λ_i via a linear transformation of the original. In Figure 5.3 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 involved as well and evolves with the conditioning variable 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 that the σ_i^2 distribution is continuous with a bounded support above zero (see Propositions 3.13 and 3.14), and that the signal-to-noise ratio is not far from 1. Their joint and conditional distributions are depicted in Figure 5.3 panel (b).

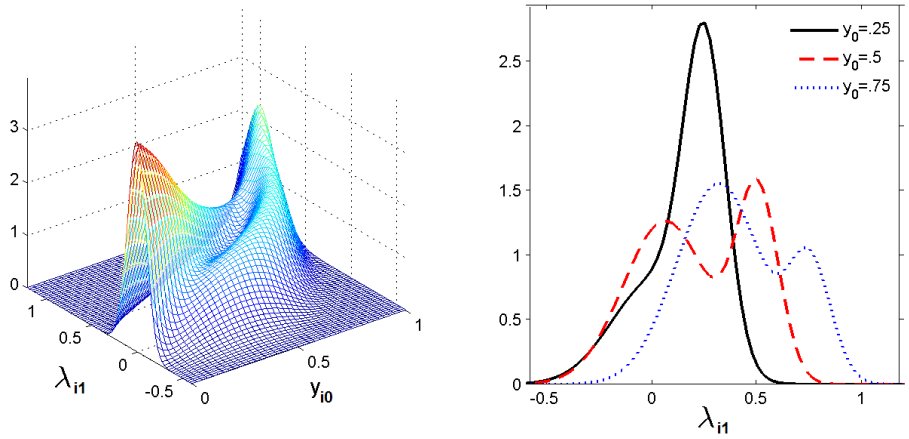
Due to cross-sectional heteroskedasticity and correlated random coefficients, the prior structures become more complicated. Table 5.4 describes the prior setups of λ_i and l_i, with the predictor labels being consistent with the definitions in Subsection 5.2. Note that I further add the Homosk-NP-C predictor in order to examine whether it is practically relevant to model heteroskedasticity.

Table 5.5 assesses the forecasting performance of these predictors. Considering point forecasts, Heterosk-NP-R and Heterosk-Param constitute the first tier, Heterosk-NP-disc and Heterosk-NP-C can be viewed as the second tier, Homosk-NP-C is the third tier, and Homog and Heterosk-Flat are markedly inferior. It is not surprising that more parsimonious estimators 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_is and shock sizes σ_i^2s.

Figure 5.3: DGP: General Model

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



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

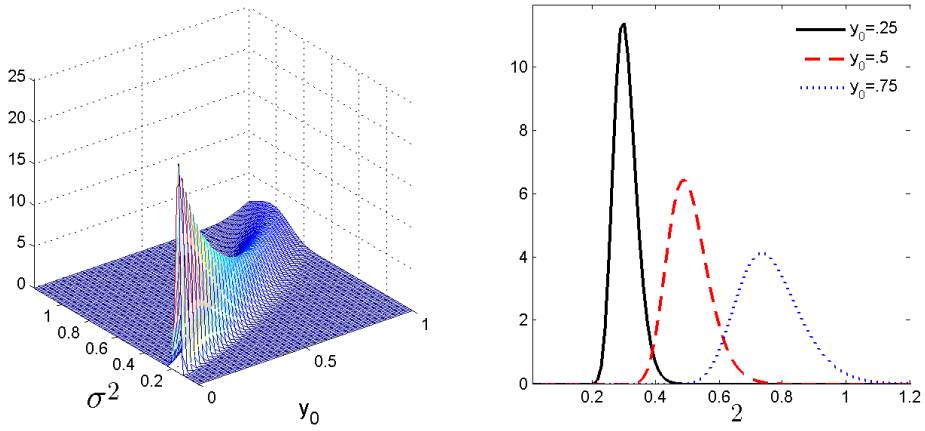


Table 5.5: Forecast Evaluation: General Model

		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*

See the description of Table 5.2. Here the tests are conducted with respect to Heterosk-NP-C.

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.

6 Empirical Application: Young Firm Dynamics

6.1 Background and Data

To see 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 (Akcigit and Kerr, 2016; Robb and Seamans, 2014; Zarutskie and Yang, 2015). In this empirical application, I examine this type of firm-specific latent heterogeneity from a density forecasting perspective.

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 innovation and performance, and spreads over the recent recession. The restricted-access Kauffman Firm Survey (KFS) is the ideal candidate for such purposes, as it offers the largest panel of startups (4,928 firms founded in 2004, nationally representative sample) and the longest time span (2004-2011, one baseline survey and seven follow-up annual surveys), together with detailed information on young firms. See Robb *et al.* (2009) for further description of the survey design.[39]

[39]Here I do not impose KFS sample 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.

Table 6.1: 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

6.2 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 (2016) and Zarutskie and Yang (2015), firm performance is measured by employment. From an economic point of view, young firms make a significant contribution to employment and job creation (Haltiwanger *et al.*, 2012), and their struggle during the recent recession may partly account for the recent jobless recovery. 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 the employment growth rate, as the latter significantly reduces the cross-sectional sample size due to the rank requirement for unbalanced panels.

Below, I focus on the following model specification,[40]

$$\log \text{emp}_{it} = \beta \log \text{emp}_{i,t-1} + \lambda_{1i} + \lambda_{2i} \text{R\&D}_{i,t-1} + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right),$$

where $\text{R\&D}_{it}$ 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 with other innovation intensity gauges.[41]

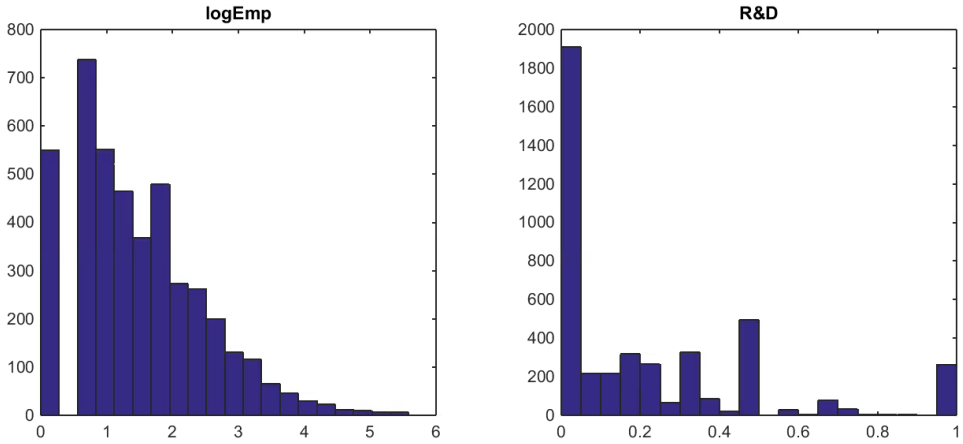
The panel used for estimation spans from 2004 to 2010 with time dimension $T = 6$.[42] The data for 2011 are reserved for pseudo out-of-sample forecast evaluation. The sample is constructed as follows. First, for any (i, t) combination where R&D employment is greater than the total employment, there is an incompatibility issue, so I set $\text{R\&D}_{it} = NA$, which only affects 0.68% of the observations. Then, I only keep firms with long enough observations according to Assumption 3.5, which ensures identification in unbalanced panels. This results in cross-sectional dimension $N = 654$. 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 6.1, and the corresponding histograms are plotted in Figure 6.1, where both distributions are right skewed and may have more than one peak. Therefore, we anticipate that the proposed predictors with nonparametric priors would perform well in this scenario.

[40] See Appendix D.2 for other setups.

[41] I have also explored other measures of firm performance (e.g. the log of revenue) and 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 relative rankings of density forecasts are generally robust across measures.

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

Figure 6.1: Histograms for Observables



6.3 Results

The alternative priors are similar to those in the Monte Carlo simulation except for one additional prior, Heterosk-NP-C/R, where λ_i can be correlated with y_{i0} while σ_i^2 is independent with respect to y_{i0}, by adopting an MGLR_x prior on λ_i and a DPM prior on $l_i = \log \frac{\bar{\sigma}^2(\sigma_i^2 - \underline{\sigma}^2)}{\bar{\sigma}^2 - \sigma_i^2}$,.[43] 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.

The first two columns in Table 6.2 characterize the posterior estimates of the common parameter β. In most of the cases, the posterior means are around $0.5 \sim 0.6$, which suggests that the young firm performance exhibits some degree of persistence, 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. NP-R also gives a large estimate of β. The reason is similar – if the true data generating process is correlated random effects/coefficients, the random effects/coefficients model would miss the effects of the initial condition and misinterpret them as the persistence of the system. In all scenarios, the posterior standard deviations are relatively small, which indicates that the posterior distributions are very tight.[44]

The last two columns in Table 6.2 compare the forecasting performance. The Heterosk-NP-C/R predictor is the benchmark for all comparisons. In terms of point forecasts, most of the estimators are comparable according to MSE, with only Flat performing significantly poorly. Intuitively, shrinkage

[43] 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}. I will leave this to future exploration.

[44] Comparing with the literature, the closest one is Zarutskie and Yang (2015) using traditional panel data methods, where the estimated persistence of log employment is 0.824 and 0.816 without firm fixed effects (Table 2) which is close to Homog, and 0.228 with firm fixed effects estimated via OLS (Table 4) which is smaller than all the estimates here.

Table 6.2: Parameter Estimation and Forecast Evaluation: Young Firm Dynamics

		β		Forecast	
		Mean	Std	MSE	LPS*N
Heterosk	NP-C/R	0.52	0.01	**0.20**	**-228**
Homog		0.89	0.02	8%*	-74***
Homosk	NP-C	0.51	0.03	9%	-52***
Heterosk	Flat	0.50	0.00	102%***	-309***
	Param	0.56	0.03	7%	-52***
	NP-disc	0.84	0.04	2%	-20***
	NP-R	0.74	0.04	3%	-16***
	NP-C	0.53	0.01	0.1%	-5**

See the description of Table 5.2 for forecast evaluation. Here Heterosk-NP-C/R is the benchmark for both normalization and significance tests.

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 to 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).

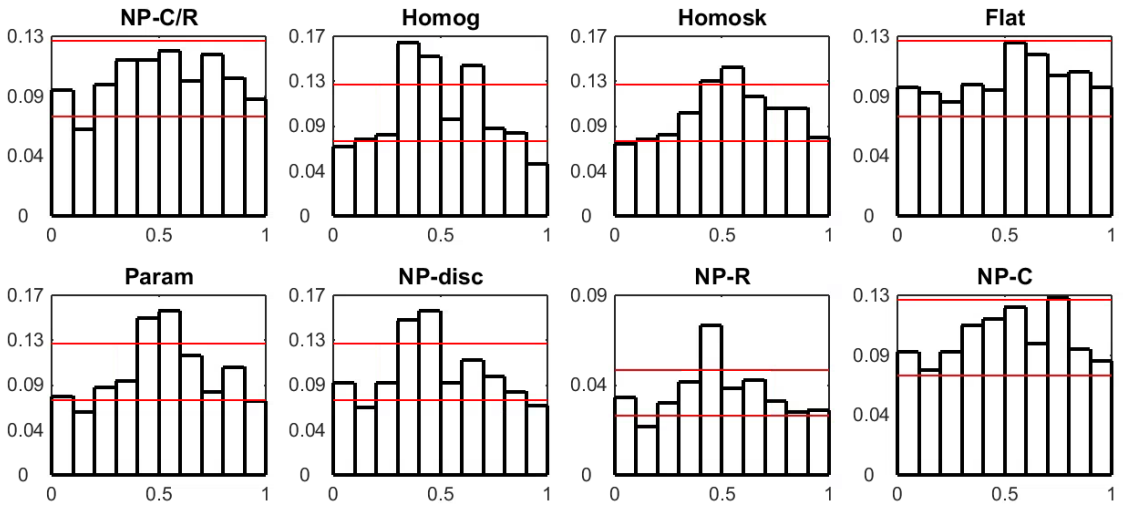
For density forecasts measured by LPS, the overall best is the Heterosk-NP-C/R predictor. The main message is similar to the Monte Carlo simulation of the general model in Subsection 5.4. In summary, it is crucial to account for individual effects in both coefficients λ_is and shock sizes σ_i^2s through a flexible nonparametric prior that acknowledges continuity and correlated random effects/coefficients when the underlying individual heterogeneity is likely to possess these features. Intuitively, the odds given by the exponential of the difference in LPS indicate that the future realizations are on average 12% more likely in Heterosk-NP-C/R versus Homog, 60% more likely in Heterosk-NP-C/R versus Heterosk-Flat, etc. 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.[45]

Figure 6.2 provides the histograms of the probability integral transformation (PIT). While LPS characterizes the relative ranks of predictors, PIT supplements LPS and can be viewed as an absolute evaluation on how well 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 true data generating process, the probability integral transforms are i.i.d. $U(0,1)$ and the histogram is close to a flat line.[46] In each subgraph, the two red lines indicate the confidence interval. We can see that, in NP-C/R, NP-C, and Flat,

[45]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.

[46]See Diebold *et al.* (1998) for details of PIT and Amisano and Geweke (2017) for formal PIT tests.

Figure 6.2: PIT



Red lines indicate the confidence interval.

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 attribute the subtlety of correlated random coefficients to the estimated variance, which leads to more diffused predictive distributions.

Figure 6.3 shows the predictive distributions of 10 randomly selected firms. 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 their center location, variance, and skewness.

Figure 6.4 further aggregates the predictive distributions over sectors based on the two-digit NAICS codes (Table 6.3). 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. The reason is that NP-C/R tailors to each individual firm while Homog prescribes a general model to all the firms, so NP-C/R yields more precise predictive distributions. Second, NP-C/R predictive distributions have longer right tails, whereas Homog ones are distributed in the standard bell shape. The long right tails in NP-C/R concur with the general intuition that good ideas are scarce. Finally, there is substantial heterogeneity in density forecasts across sectors. For sectors with relatively large average employment, e.g. construction (sector 23), Homog pushes the forecasts down and 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

Figure 6.3: Predictive Distributions: 10 Randomly Selected Firms

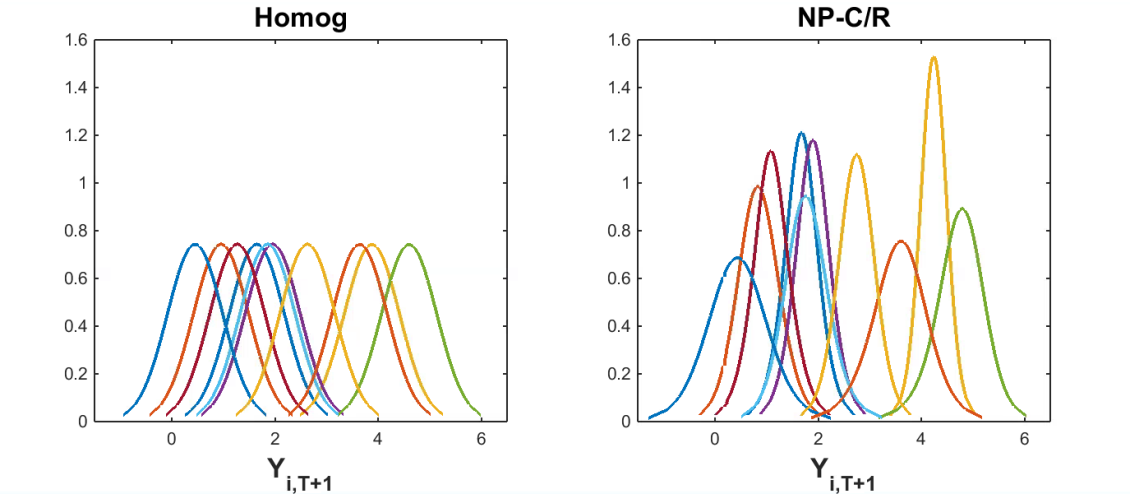


Table 6.3: Two-digit NAICS Codes

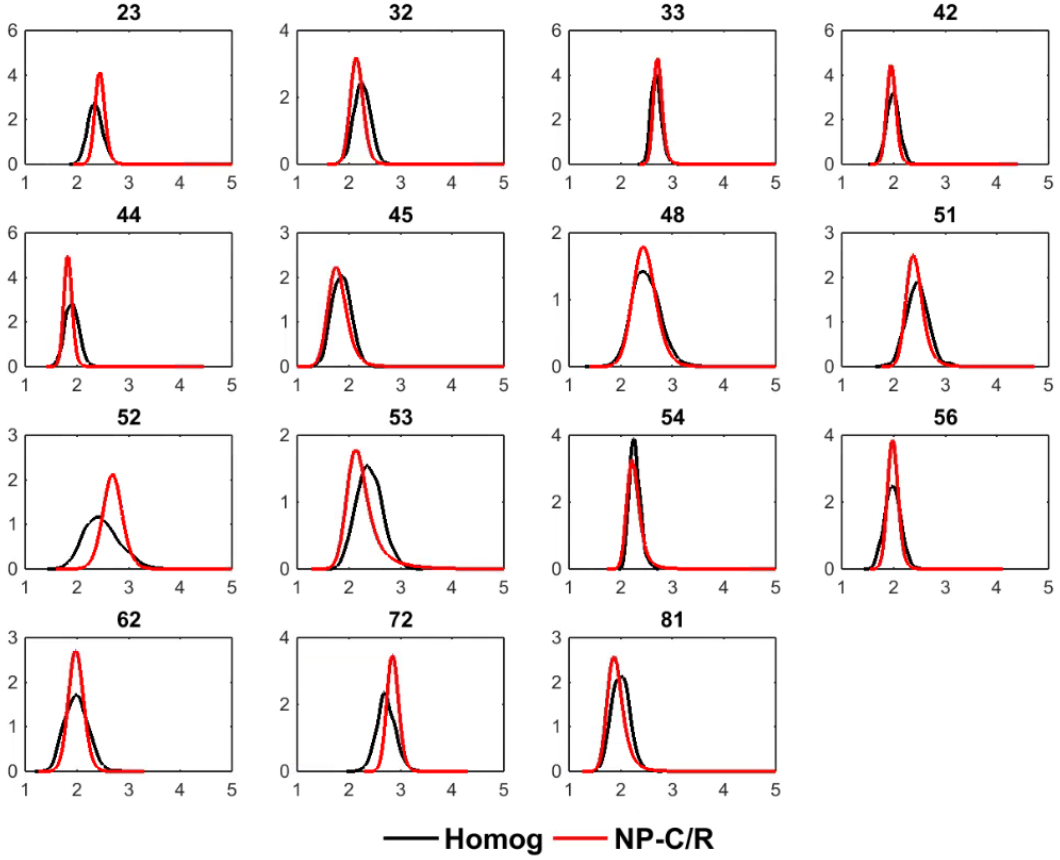
Code	Sector	Code	Sector
11	Agriculture, Forestry, Fishing and Hunting	52	Finance and Insurance
21	Mining, Quarrying, and Oil and Gas Extraction	53	Real Estate and Rental and Leasing
22	Utilities	54	Professional, Scientific, and Technical Services
23	Construction	56	Administrative and Support and Waste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

firm-specific heterogeneity.

The latent heterogeneity structure is presented in Figure 6.5, which plots the joint distributions of the estimated individual effects and the conditional variable. 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. Furthermore, $\lambda_{i,\text{level}}$, $\lambda_{i,\text{RD}}$, 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.[47]

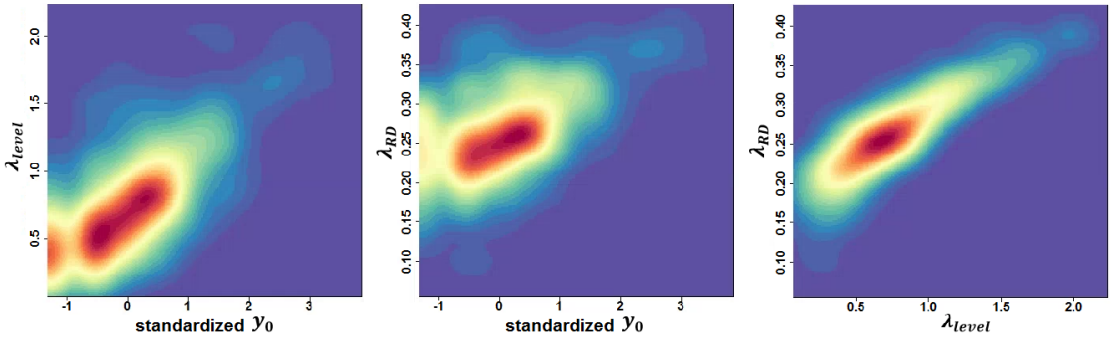
[47]The model here mainly serves the forecasting purpose, so we need to be careful with any causal interpretation.

Figure 6.4: Predictive Distributions: Aggregated by Sectors



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

Figure 6.5: Joint Distributions of $\hat{\lambda}_i$ and Condition Variable



7 Concluding Remarks

This paper proposes a semiparametric Bayesian predictor, which performs well in density forecasts of individuals in a panel data setup. It considers the underlying distribution of individual effects and pools the information from the whole cross-section in a flexible and efficient way. Monte Carlo simulations and an empirical application to young firm dynamics show that the keys for the better density forecasts are, in order of importance, nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients.

Moving forward, I plan to extend my research in the following directions. Theoretically, I will continue the Bayesian asymptotic discussion with rates of convergence, which will provide more insight into how N, T, d_w, and shock size affect the performance of the proposed semiparametric Bayesian predictor. 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 us 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 derived from economic theories 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, which helps accommodate firms' endogenous exit choice. Such extensions would be numerically feasible, but require further theoretical work. A natural next step would be extending the theoretical discussion to the family of "generalized linear models".

References

Akcigit, U. and Kerr, W. R. (2016). Growth through heterogeneous innovations. *Journal of Political Economy*, forthcoming.

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 Geweke, J. (2017). Prediction using several macroeconomic models. *The Review of Economics and Statistics*, **99** (5), 912–925.

— and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, **25** (2), 177–190.

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 Bonhomme, S. (2012). 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.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, **27** (2), 536–561.

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.

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.

Canale, A. and De Blasi, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, **23** (1), 379–404.

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.

Chib, S. (2008). Panel data modeling and inference: a Bayesian primer. In *The econometrics of panel data*, Springer, pp. 479–515.

— and Carlin, B. P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, **9** (1), 17–26.

Chung, Y. and Dunson, D. B. (2012). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*.

Compiani, G. and Kitamura, Y. (2016). Using mixtures in econometric models: a brief review and some new results. *The Econometrics Journal*, **19** (3), C95–C127.

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. (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.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity.

— and White, H. (2012). Some extensions of a lemma of Kotlarski. *Econometric Theory*, **28** (4), 925–932.

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.

— and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, **35** (2), 697–723.

— and — (2017). *Fundamentals of Nonparametric Bayesian Inference*, vol. 44. Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer-Verlag.

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. (2017a). Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics*, **32** (3), 575–599.

— and — (2017b). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics*, **35** (1), 1–16.

Hall, B. H. and Rosenberg, N. (2010). *Handbook of the Economics of Innovation*, vol. 1. Elsevier.

Haltiwanger, J., Jarmin, R. S. and Miranda, J. (2012). Who creates jobs? Small versus large versus young. *Review of Economics and Statistics*, **95** (2), 347–361.

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*. Cambridge university press.

Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics*, **200** (2), 154–168.

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.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21** (1), 93–105.

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.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, **65** (2), 139 – 165.

Liu, L., Moon, H. R. and Schorfheide, F. (2017). 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).

Llera, A. and Beckmann, C. (2016). Estimating an Inverse Gamma distribution. *arXiv preprint arXiv:1605.01019*.

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 Pati, D. (2017). Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, **33** (4), 980–1012.

— 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. (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.

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.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pp. 639–650.

Shin, M. (2014). Bayesian GMM.

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of econometrics*, **95** (2), 443–462.

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.

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.

A 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, $U(0,1)$.

$N\left(\mu,\sigma^2\right)$ stands for a **Gaussian/normal 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 $N(\mu,\Sigma)$, $\phi(x;\mu,\Sigma)$, and $\Phi(x;\mu,\Sigma)$ are for the distribution with the mean vector μ and the covariance matrix Σ.

The **gamma distribution** is denoted as $\mathrm{Ga}(a,b)$ with pdf 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 $\mathrm{IG}(a,b)$ with pdf 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 $d\times d$ positive definite matrix following an inverse Wishart distribution $\mathrm{IW}(\Psi,\nu)$, then 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 an inverse gamma distribution with $a=\nu/2,\ b=\Psi/2$.

Let G be a distribution drawn from the **Dirichlet Process (DP)**. Denote $G\sim DP(\alpha,G_0)$, if for any partition $(A_1,\cdots,A_K)$, $(G(A_1),\cdots,G(A_K))\sim \mathrm{Dir}(\alpha G_0(A_1),\cdots,\alpha G_0(A_K))$. $\mathrm{Dir}(\cdot)$ stands for the Dirichlet distribution, which is a multivariate generalization of the Beta distribution. An alternative view of DP is given by the stick-breaking process, $G=\sum_{k=1}^{\infty}p_k\mathbf{1}(\theta=\theta_k)$, where $\theta_k\sim G_0$ and $p_k\sim \mathrm{SB}(1,\alpha)$. $\mathbf{1}(\cdot)$ is an **indicator function** that equals 1 if the condition in the parenthesis is satisfied and equals 0 otherwise.

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

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.

$\|\cdot\|_p$ represents the L_p**-norm**, e.g. the **Euclidean norm** for a n-dimensional vector $z=[z_1,z_2,\cdots,z_n]'$ is given by $\|z\|_2=\sqrt{z_1^2+\cdots+z_n^2}$, and the L_1**-norm** for an integrable function is given by $\|f\|_1=\int|f(x)|\,dx$.

$\mathrm{supp}(\cdot)$ denotes the **support** of a probability measure.

$\mathrm{vec}(\cdot)$ denotes **matrix vectorization**, and $\otimes$ is the **Kronecker product**.

B Explanations and Proofs

B.1 Intuition: MGLR$_x$ Prior

Here we give some intuition why the MGLR$_x$ Prior is general enough to accommodate a broad class of conditional distributions.

Define a generic variable z which can represent either λ or l. By Bayes' theorem,

$$f\left(z|c_0\right) = \frac{f\left(z, c_0\right)}{f\left(c_0\right)}.$$

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

$$f\left(z, c_0\right) \approx \sum_{k=1}^{\infty} \tilde{p}_k \phi\left(\left[z', c_0'\right]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right),$$

where $\tilde{\mu}_k$ is a $(d_z + d_{c_0})$-element column vector, and $\tilde{\Omega}_k$ is a $(d_z + d_{c_0}) \times (d_z + d_{c_0})$ covariance matrix.

$$\begin{aligned}
\tilde{\mu}_k &= \left[\tilde{\mu}_{k,z}', \tilde{\mu}_{k,c_0}'\right]', \\
\tilde{\Omega}_k &= \left[\begin{array}{cc} \tilde{\Omega}_{k,zz} & \tilde{\Omega}_{k,zc_0} \\ \tilde{\Omega}_{k,c_0z} & \tilde{\Omega}_{k,c_0c_0} \end{array}\right].
\end{aligned}$$

Applying Bayes' theorem again to the normal kernel for each component k,

$$\phi\left(\left[z', c_0'\right]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right) = \phi\left(c_0;\ \tilde{\mu}_{k,c_0}, \tilde{\Omega}_{k,c_0c_0}\right) \phi\left(z;\ \mu_k\left[1, c_0'\right]', \Omega_k\right),$$

where $\mu_k = \left[\tilde{\mu}_{k,z} - \tilde{\Omega}_{k,zc_0}\tilde{\Omega}_{k,c_0c_0}^{-1}\tilde{\mu}_{k,c_0}\right]$, $\Omega_k = \tilde{\Omega}_{k,zz} - \tilde{\Omega}_{k,zc_0}\tilde{\Omega}_{k,c_0c_0}^{-1}\tilde{\Omega}_{k,zc_0}'$. Combining all the steps above, the conditional distribution can be approximated as

$$\begin{aligned}
f\left(z|c_0\right) &\approx \sum_{k=1}^{\infty} \frac{\tilde{p}_k \phi\left(c_0;\ \tilde{\mu}_{k,c_0}, \tilde{\Omega}_{k,c_0c_0}\right) \phi\left(z;\ \mu_k\left[1, c_0'\right]', \Omega_k\right)}{f\left(y_0\right)} \\
&= \sum_{k=1}^{\infty} p_k\left(c_0\right) \phi\left(z;\ \mu_k\left[1, c_0'\right]', \Omega_k\right),
\end{aligned}$$

The last line is given by collecting marginals of c_0 into $p_k\left(c_0\right) = \frac{\tilde{p}_k \phi\left(c_0;\, \tilde{\mu}_{k,c_0}, \tilde{\Omega}_{k,c_0c_0}\right)}{f(c_0)}$.

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 c_0, and thus allows for more relaxed assumptions on it.

B.2 Identification

Proof. **(Proposition 3.3)**

Part 2 for cross-sectional heteroskedasticity σ_i^2 is new. Part 3 for additive individual-heterogeneity λ_i follows Liu *et al.* (2017), which is based on the early works such as Arellano and Bover (1995) and Arellano and Bonhomme (2012).

1. Identify common parameters β

First, let us perform orthogonal forward differencing, i.e. for $t = 1, \cdots, T - d_w$,

$$\tilde{y}_{it} = y_{it} - w_{i,t-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}' \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} y_{is}, \tag{B.1}$$

$$\tilde{x}_{i,t-1} = x_{i,t-1} - w_{i,t-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}' \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} x_{i,s-1}. \tag{B.2}$$

Then, β is identified given Assumption 3.1 (2-d) and the following moment conditions

$$\mathbb{E} \sum_{t} \tilde{x}_{i,t-1} \left(\tilde{y}_{it} - \tilde{x}_{i,t-1}' \beta \right) = 0.$$

2. Identify the distribution of shock sizes f^{σ^2}

After orthogonal forward differencing, define

$$\begin{aligned} \tilde{u}_{it} &= \tilde{y}_{it} - \beta' \tilde{x}_{i,t-1}, \\ \hat{\sigma}_i^2 &= \sum_{t=1}^{T-d_w} \tilde{u}_{it}^2 = \sigma_i^2 k_i^2. \end{aligned}$$

where $k_i^2 \sim \chi^2 (T - d_w)$ follows an i.i.d. chi-squared distribution with $(T - d_w)$ degrees of freedom.

Note that Fourier transform (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[48]

$$M_x(\xi) = \int x^{i\xi} f(x) \, dx,$$

which exists for all $\xi \in \mathbb{R}$.

Considering that $\sigma_i^2 | c_{i0}$ and k_i^2 are independent, we have

$$M_{\hat{\sigma}^2}(\xi | c) = M_{\sigma^2}(\xi | c) \, M_{k^2}(\xi).$$

Note that the non-vanishing characteristic function of σ^2 implies non-vanishing Mellin transform

[48] See the discussion on page 16 of Galambos and Simonelli (2004) for the generality of this specification.

$M_{\sigma^2}(\xi|c)$ (almost everywhere), so it is legitimate to take the logarithm of both sides,

$$\log M_{\hat{\sigma}^2}(\xi|c) = \log M_{\sigma^2}(\xi|c) + \log M_{k^2}(\xi).$$

Taking the second derivative with respect to ξ, we get

$$\frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\sigma^2}(\xi|c) = \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\hat{\sigma}^2}(\xi|c) - \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{k^2}(\xi).$$

The Mellin transform of chi-squared distribution $M_{k^2}(\xi)$ is a known functional form. In addition, we have

$$\begin{aligned}
\log M_{\sigma^2}(0|c) &= \log M_{\hat{\sigma}^2}(0|c) - \log M_{k^2}(0) = 0, \\
\frac{\partial}{\partial\xi}\log M_{\sigma^2}(0|c) &= \frac{\partial}{\partial\xi}\log M_{\hat{\sigma}^2}(0|c) - \frac{\partial}{\partial\xi}\log M_{k^2}(0) \\
&= i\left(\mathbb{E}\left(\log\hat{\sigma}^2\middle|\, c\right) - \mathbb{E}\left(\log k^2\middle|\, c\right)\right).
\end{aligned}$$

Based on Pav (2015),

$$\mathbb{E}\left(\log k^2\middle|\, c\right) = \log 2 + \psi\left(\frac{T-d_w}{2}\right),$$

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}. See Theorem 1.19 in Galambos and Simonelli (2004) for the uniqueness.

3. 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 3.1 (2-a), F_{Λ} and F_U are non-vanishing almost everywhere. Then, we obtain

$$\log F_{\Lambda}\left(W'\xi|c\right) = \log F_{\mathring{Y}}(\xi|c) - \log F_U(\xi|c),$$

where $F_{\mathring{Y}}$ is constructed from the observables and the common parameters identified in part 1, and F_U is based on the f^{σ^2} identified in part 2. Let $\zeta = W'\xi$ and $A_W = (W'W)^{-1}W'$, then the second

derivative of $\log F_\Lambda(\zeta|c)$ is characterized by

$$\frac{\partial^2}{\partial\zeta\partial\zeta'}\log F_\Lambda(\zeta|c) = A_W\left(\frac{\partial^2}{\partial\xi\partial\xi'}\left(\log F_{\mathring{Y}}(\xi|c) - \log F_U(\xi|c)\right)\right)A_W'.$$

Moreover,

$$\begin{aligned}\log F_\Lambda(0|c) &= 0,\\ \frac{\partial}{\partial\zeta}\log F_\Lambda(0|c) &= iA_W\mathbb{E}\left(\mathring{Y}\middle| c\right),\end{aligned}$$

so we can pin down $\log\Lambda(\zeta|c)$ and f^λ.

The proof for unbalanced panels follows in a similar manner. □

B.3 Posterior Consistency: Random Coefficients Model

Proof. **(Proposition 3.10)**

The proof follows **?**, which is based on the early work by Barron *et al.* (1999) and Ghosal and van der Vaart (2007). Now we significantly extend the discussion to take care of the deconvolution and dynamic panel data structure.

1. Random coefficients: cross-sectional homoskedasticity

The posterior probability of the alternative region can be decomposed as

$$\begin{aligned}&\Pi\left((\vartheta,f):\ \|\vartheta-\vartheta_0\|_2\geq\delta \text{ or } \|f-f_0\|_1\geq\epsilon|\,D\right)\\ \leq&\Pi^\vartheta\left(\|\vartheta-\vartheta_0\|_2\geq\delta|\,D\right)+\Pi^f\left(\{\|f-f_0\|_1\geq\epsilon\}\bigcap\mathcal{F}_N^c\middle|\,D\right)\\ &+\Pi\left(\{\|f-f_0\|_1\geq\epsilon\}\bigcap\mathcal{F}_N,\|\vartheta-\vartheta_0\|_2<\delta\middle|\,D\right).\end{aligned}$$

It suffices to show that (a) for all $\delta>0$, $\Pi^\vartheta\left(\|\vartheta-\vartheta_0\|_2\geq\delta|\,D\right)\to0$, (b) for all $\epsilon>0$, $\Pi^f\left(\{\|f-f_0\|_1\geq\epsilon\}\bigcap\mathcal{F}_N^c|\,D\right)\to0$, and (c) for all $\epsilon>0$, $\Pi\left(\{\|f-f_0\|_1\geq\epsilon\}\bigcap\mathcal{F}_N,\|\vartheta-\vartheta_0\|_2<\delta(\epsilon)|\,D\right)\to0$. We let δ depend on ϵ in part (c) because part (a) holds for all $\delta>0$.

(a) For all $\delta>0$, $\Pi^\vartheta\left(\|\vartheta-\vartheta_0\|_2\geq\delta|\,D\right)\to0$.

After orthogonal forward differencing in equations (B.1) and (B.2), the posterior of $\left(\beta,\sigma^2\right)$ is

given by

$$
\begin{aligned}
p\left(\beta,\sigma^2|D\right) &\propto \phi\left(\beta;m^\beta,\psi^\beta\sigma^2\right) f_{\mathrm{IG}}\left(\sigma^2;a^{\sigma^2},b^{\sigma^2}\right) d\Pi^\vartheta\left(\beta,\sigma^2\right),\\
\psi^\beta &= \left(\sum_{i,t}\tilde{x}_{i,t-1}\tilde{x}_{i,t-1}'\right)^{-1},\\
m^\beta &= \psi^\beta\left(\sum_{i,t}\tilde{x}_{i,t-1}\tilde{y}_{it}\right),\\
a^{\sigma^2} &= \frac{N\left(T-d_w\right)}{2}\\
b^{\sigma^2} &= \frac{1}{2}\left(\sum_{i,t}\tilde{y}_{it}^2 - m^{\beta\prime}\left(\psi^\beta\right)^{-1}m^\beta\right).
\end{aligned}
$$

Then, the traditional posterior consistency argument implies that $\left(\beta,\sigma^2\right)\big|\, D$ converges to $\left(\beta_0,\sigma_0^2\right)$, given Assumption 3.1 (2-c) ($\mathbb{E}\big[\sum_t \tilde{x}_{i,t-1}\tilde{x}_{i,t-1}'\big]$ has full rank) and Proposition 3.10 condition 3 (ϑ_0 is in the interior of $\mathrm{supp}\left(\Pi^\vartheta\right)$).

(b) For all $\epsilon>0$, $\Pi^f\left(\left\{\|f-f_0\|_1\geq\epsilon\right\}\bigcap\mathcal{F}_N^c|\, D\right)\to 0$.

Based on Lemma 1 in Canale and De Blasi (2017), Assumption 3.9 (1, 2-a) ensures that the KL property holds for the distribution of λ, i.e. for all $\epsilon>0$,

$$
\Pi^f\left(f\in\mathcal{F}:\ \int f_0\left(\lambda\right)\log\frac{f_0\left(\lambda\right)}{f\left(\lambda\right)}d\lambda<\epsilon\right)>0. \tag{B.3}
$$

Now, we need to establish an altered KL property specified on the observables. First, the individual-specific likelihood function is characterized as

$$
\begin{aligned}
&g\left(y_{i,0:T},x_{i,0:T-1}^*,w_{i,0:T-1}^I|D_A\right)\\
=&\prod_t p\left(x_{i,t-1}^{P*}|y_{i,t-1},c_{i,0:t-2}\right)p\left(c_{i0}^*|D_A\right)\int\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)f\left(\lambda_i\right)d\lambda_i,
\end{aligned} \tag{B.4}
$$

and $g_0\left(y_{i,0:T},x_{i,0:T-1}^*,w_{i,0:T-1}^I|D_A\right)$ corresponds to the true data generating process $\left(\beta_0,\sigma_0^2,f_0\right)$. Then, we would like to prove that for all $\epsilon>0$,

$$
\Pi\left(\begin{array}{l}
f\in\mathcal{F},\ \left(\beta,\sigma^2\right)\in\mathbb{R}^{d_x}\times\mathbb{R}^+:\\
\int g_0\left(y_{i,0:T},x_{i,0:T-1}^*,w_{i,0:T-1}^I|D_A\right)\log\dfrac{g_0\left(y_{i,0:T},x_{i,0:T-1}^*,w_{i,0:T-1}^I|D_A\right)}{g\left(y_{i,0:T},x_{i,0:T-1}^*,w_{i,0:T-1}^I|D_A\right)}dy_{i,0:T}dx_{i,0:T-1}^*dw_{i,0:T-1}^I<\epsilon
\end{array}\right)>0, \tag{B.5}
$$

The KL divergence of g with respect to g_0 can be further decomposed as

$$
\begin{aligned}
&\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)}{g\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}\\
=&\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)f\left(\lambda_i\right)d\lambda_i}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}\\
=&\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}\\
&+\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)f\left(\lambda_i\right)d\lambda_i}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1},
\end{aligned}
$$

where the first equality is given by crossing out common factors in the numerator and denominator.

For the first term, define $h\left(x\right)=x\log x$, $a\left(\lambda_i\right)=\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)$, $A=\int a\left(\lambda_i\right)d\lambda_i$, $b\left(\lambda_i\right)=\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)$, $B=\int b\left(\lambda_i\right)d\lambda_i$. We can rewrite the integral over λ_i as

$$
\begin{aligned}
&\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}\\
=&A\cdot\log\frac{A}{B}=B\cdot g\left(\frac{A}{B}\right)=B\cdot g\left(\int\frac{b\left(\lambda_i\right)}{B}\cdot\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}d\lambda_i\right)\\
\leq&\int b\left(\lambda_i\right)g\left(\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}\right)d\lambda_i\\
=&\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)\log\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}d\lambda_i, \qquad \text{(B.6)}
\end{aligned}
$$

where the inequality is given by Jensen's inequality. Then, further integrating the above expression over $\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}\right)$, we have

$$
\begin{aligned}
&\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)f_0\left(\lambda_i\right)d\lambda_i}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}\\
\leq&\int\prod_t p\left(x^{P*}_{i,t-1}|y_{i,t-1},c_{i,0:t-2}\right)p\left(c^*_{i0}|D_A\right)\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)\log\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}dy_{i,0:T}dx^*_{i,0:T-1}dw^*_{i,0:T-1}\\
&\cdot\int f_0\left(\lambda_i\right)\log\frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)}d\lambda_i,
\end{aligned}
$$

where the inequality follows the above expression (B.6). According to the KL property of the distribution of λ in equation (B.3), for all $\epsilon'>0$, there exists $G^f_{\epsilon'}=\left\{f\in\mathcal{F}:\ \int f_0\left(\lambda\right)\log\frac{f_0(\lambda)}{f(\lambda)}d\lambda<\epsilon'\right\}$ such that f_0 is in the interior of $G^f_{\epsilon'}$, $\Pi^f\left(G^f_{\epsilon'}\right)>0$, and the first term is less than ϵ' for all $f\in G^f_{\epsilon'}$.

For the second term, we first employ the convexity of KL divergence,

$$
\begin{aligned}
&\int g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1}|D_A\right)\log\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)f\left(\lambda_i\right)d\lambda_i}dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}\\
\leq&\int\prod_t p\left(x^{P*}_{i,t-1}\left|y_{i,t-1},c_{i,0:t-2}\right.\right)p\left(c^*_{i0}|D_A\right)\frac{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)d\lambda_i}{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}\\
&\cdot\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)\log\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)}{\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)}d\lambda_i dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}.\\
=&\int\prod_t p\left(x^{P*}_{i,t-1}\left|y_{i,t-1},c_{i,0:t-2}\right.\right)p\left(c^*_{i0}|D_A\right)\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f_0\left(\lambda_i\right)\\
&\cdot\left[\int\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)}{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}\log\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)}{\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)}d\lambda_i\right]\\
&\cdot d\lambda_i dy_{i,0:T}dx^*_{i,0:T-1}dw^I_{i,0:T-1}. \qquad \text{(B.7)}
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\int\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)}{\int\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)f\left(\lambda_i\right)d\lambda_i}\log\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)}{\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)}d\lambda_i\\
=&\int\frac{\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f\left(\lambda_i\right)}{\int\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f\left(\lambda_i\right)d\lambda_i}\log\frac{\prod_t\phi\left(y_{it};\beta_0'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma_0^2\right)}{\prod_t\phi\left(y_{it};\beta'x_{i,t-1}+\lambda_i'w_{i,t-1},\sigma^2\right)}d\lambda_i, \qquad \text{(B.8)}
\end{aligned}
$$

where

$$
\begin{aligned}
m\left(\beta\right)&=\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)^{-1}\sum_t w_{i,t-1}\left(y_{it}-\beta'x_{i,t-1}\right),\\
\Sigma\left(\sigma^2\right)&=\sigma^2\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)^{-1}.
\end{aligned}
$$

The log of the ratio of normal distributions has an analytical form,

$$\begin{aligned}
&\log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} \\
=&\frac{T}{2}\left(\log\sigma^2 - \log\sigma_0^2\right) + \frac{1}{2}\sum_t \left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right) \\
&+ \sum_t \frac{\left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 - \left(y_{it} - \beta_0' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2}{2\sigma_0^2} \\
=&\frac{T}{2}\left(\log\sigma^2 - \log\sigma_0^2\right) + \frac{1}{2}\sum_t \left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right) \\
&+ \sum_t \frac{\left(\beta' x_{i,t-1}\right)^2 - \left(\beta_0' x_{i,t-1}\right)^2 - 2\left(y_{it} - \lambda_i' w_{i,t-1}\right)\left(\beta - \beta_0\right)' x_{i,t-1}}{2\sigma_0^2}.
\end{aligned}$$

For all $\epsilon' > 0$, there exists $G_{\epsilon'}^{\sigma^2} = \left\{\sigma^2 \in \sigma_0^2\left[1, 1+\eta\right)\right\}$ such that σ_0^2 is on the boundary of $G_{\epsilon'}^{\sigma^2}$, $\Pi^{\sigma^2}\left(G_{\epsilon'}^{\sigma^2}\right) > 0$, and the sum of the first two terms is less than $\epsilon' = \frac{T}{2}\log\left(1+\eta\right)$. Given that T is finite and that $w_{i,0:T-1}$ is bounded due to Assumption 3.8 (1), the last term is of order

$$O\left(\left(\sigma_0^2\right)^{-1}\left\|\beta - \beta_0\right\|_2 \sum_t \left(y_{it}^2 + \left\|x_{i,t-1}\right\|_2^2 + \left\|\lambda_i\right\|_2^2\right)\right). \tag{B.9}$$

The term $\left(\sigma_0^2\right)^{-1}$ can be ignored in cross-sectional homoskedastic cases, but I keep it here so that the derivation is more comparable to cross-sectional heteroskedastic cases. Considering that $f\left(\lambda_i\right)$ is characterized by an infinite mixture of multivariate normals, we have

$$\begin{aligned}
&\phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_0^2\right)\right) f\left(\lambda_i\right) \\
=&\phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_0^2\right)\right) \sum_k p_k \phi\left(\lambda_i; \mu_k, \Omega_k\right) \\
=&\sum_k w_k\left(\beta_0, \sigma_0^2\right) \phi\left(\lambda_i; m_k\left(\beta_0, \sigma_0^2\right), \Sigma_k\left(\sigma_0^2\right)\right),
\end{aligned}$$

where

$$\begin{aligned}
m_k\left(\beta\right) &= \Sigma_k\left(\sigma^2\right)\left(\Sigma\left(\sigma^2\right)^{-1} m\left(\beta\right) + \Omega_k^{-1}\mu_k\right), \\
\Sigma_k\left(\sigma^2\right) &= \left(\Sigma\left(\sigma^2\right)^{-1} + \Omega_k^{-1}\right)^{-1}, \\
w_k\left(\beta, \sigma^2\right) &= p_k \frac{\left|\Sigma_k\left(\sigma^2\right)\right|}{\left(2\pi\right)^{d_w/2}\left|\Omega_k\right|\left|\Sigma\left(\sigma^2\right)\right|} \exp\left(-\frac{1}{2}\begin{pmatrix} \mu_k'\Omega_k^{-1}\mu_k + m\left(\beta\right)'\Sigma\left(\sigma^2\right)^{-1} m\left(\beta\right) \\ -m_k\left(\beta, \sigma^2\right)'\Sigma_k\left(\sigma^2\right)^{-1} m_k\left(\beta, \sigma^2\right) \end{pmatrix}\right).
\end{aligned}$$

Therefore,

$$h(\lambda_i) = \int \frac{\phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) f(\lambda_i)}{\int \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) f(\lambda_i)\, d\lambda_i}$$
$$= \sum_k \frac{w_k\left(\beta_0, \sigma_0^2\right)}{\sum_k w_k\left(\beta_0, \sigma_0^2\right)} \phi\left(\lambda_i; m_k\left(\beta_0, \sigma_0^2\right), \Sigma_k\left(\sigma_0^2\right)\right) \tag{B.10}$$

is an infinite mixture of normals as well, and there exists $G_{\epsilon'}^{f*} \in G_{\epsilon'}^{f}$ such that for all $f \in G_{\epsilon'}^{f*}$,

$$\int h(\lambda_i) \|\lambda_i\|_2^2 \, d\lambda_i = O\left(\|m(\beta_0)\|_2^2\right) = O\left(\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2\right)\right). \tag{B.11}$$

Combining expressions (B.9), (B.10), and (B.11), we see that (B.8) is of the order

$$O\left(\left(\sigma_0^2\right)^{-1} \|\beta - \beta_0\|_2 \sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2\right)\right).$$

Now let us proceed with the integration in expression (B.7). First, collecting terms related to $y_{iT}|c_{i,T-1}$,

$$\int \phi\left(y_{iT}; \beta_0' x_{i,T-1} + \lambda_i' w_{i,T-1}, \sigma_0^2\right) O\left(\left(\sigma_0^2\right)^{-1} \|\beta - \beta_0\|_2 y_{iT}^2\right) dy_{iT}$$
$$= O\left(\left(\sigma_0^2\right)^{-1} \|\beta - \beta_0\|_2 \left(\|x_{i,T-1}\|_2^2 + \|\lambda_i\|_2^2\right)\right). \tag{B.12}$$

Next, collecting terms related to $x_{i,T-1}^{P*}|y_{i,T-1}, c_{i,0:T-2}$,

$$\int p\left(x_{i,t-1}^{P*}|y_{i,t-1}, c_{i,0:t-2}\right) O\left(\left(\sigma_0^2\right)^{-1} \|\beta - \beta_0\|_2 \left\|x_{i,T-1}^{P*}\right\|_2^2\right).$$

Assumption 3.8 (3) ensures that for all $\epsilon' > 0$, there exists $C > 0$ such that

$$\int_{\left\|x_{i,T-1}^{P*}\right\|_2 \geq C} \left\|x_{i,T-1}^{P*}\right\|_2^2 p\left(x_{i,T-1}^{P*}|y_{i,T-1}, c_{i,0:T-2}\right) < \epsilon'.$$

When $\left\|x_{i,T-1}^{P*}\right\|_2 < C$, for all $\epsilon' > 0$, there exists $G_{\epsilon',T}^{\beta}$ such that β_0 is in the interior of $G_{\epsilon',T}^{\beta}$, $\Pi^{\beta}\left(G_{\epsilon',T}^{\beta}\right) > 0$, and the integration with respect to $x_{i,T-1}^{P*}|y_{i,T-1}, c_{i,0:T-2}$ is less than ϵ'.

Now the remaining terms in expression (B.12) are of order

$$O\left(\left(\sigma_0^2\right)^{-1} \|\beta - \beta_0\|_2 \left(y_{i,T-1}^2 + \left\|x_{i,T-1}^{O}\right\|_2^2 + \|\lambda_i\|_2^2\right)\right).$$

As T is finite, continuing with $t = T-2, T-3, \cdots, 2$, we can employ the tail conditions of $x_{i,t-1}^{P*}|y_{i,t-1}, c_{i,0:t-2}$ to obtain $G_{\epsilon',t}^{\beta}$. Furthermore, when $t = 1$, $G_{\epsilon',1}^{\beta}$ is constructed via the tail

conditions of λ_i, x_{i0}^P, and $x_{i,0:T-1}^O$ in Assumption 3.9 (1-e) and Assumption 3.8 (2). Hence, the relevant set of β is charaterized by $G_{\epsilon'}^{\beta} = \bigcap_t G_{\epsilon',t}^{\beta}$, and we achieve the altered KL property specified on the observables in expression (B.5).

Following Barron *et al.* (1999) Lemma 4, the altered KL property in expression (B.5) ensures that for all $\epsilon' > 0$,

$$\mathbb{P}_0^{\infty}\left(\int \prod_{i\leq N} \frac{g\left(y_{i,0:T}, x_{i,0:T-1}^*, w_{i,0:T-1}^I | D_A\right)}{g_0\left(y_{i,0:T}, x_{i,0:T-1}^*, w_{i,0:T-1}^I | D_A\right)} d\Pi\left(\vartheta, f\right) \leq \exp\left(-N\epsilon'\right), \text{ infinitely often} \middle| D_A\right) = 0, \tag{B.13}$$

where $\mathbb{P}_0^{\infty}$ is characterized by the true data generating process when $N \to \infty$. Based on Assumption 3.9 (2-b), we can obtain $\Pi\left(\mathcal{F}_N^c\right) = O\left(\exp\left(-\beta N\right)\right)$ for some $\beta > 0$. Therefore, $\Pi^f\left(\left\{\|f - f_0\|_1 \geq \epsilon\right\} \bigcap \mathcal{F}_N^c | D\right) \to 0$ in $\mathbb{P}_0^{\infty}$-probability.

(c) For all $\epsilon > 0$, $\Pi\left(\left\{\|f - f_0\|_1 \geq \epsilon\right\} \bigcap \mathcal{F}_N, \|\vartheta - \vartheta_0\|_2 < \delta\left(\epsilon\right) | D\right) \to 0$.[49]

Note that

$$\begin{aligned}
&\|g - g_0\|_1 \\
=&\int \left| \prod_t p\left(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}^* | D_A\right) \int \begin{pmatrix} \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) f\left(\lambda_i\right) \\ -\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) \end{pmatrix} d\lambda_i \right| \\
&\cdot dy_{i,0:T} dx_{i,0:T-1}^* dw_{i,0:T-1}^I.
\end{aligned}$$

Let

$$\begin{aligned}
A \equiv& \int \prod_t p\left(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}^* | D_A\right) \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) dy_{i,0:T} dx_{i,0:T-1}^* dw_{i,0:T-1}^I \\
&\cdot \int \left|f\left(\lambda_i\right) - f_0\left(\lambda_i\right)\right| d\lambda_i \\
=& \left\|f - f_0\right\|_1, \\
B \equiv& \int \prod_t p\left(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}^* | D_A\right) \sum_t \left[\prod_{\tau=1}^{t-1} \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)\right. \\
&\cdot \prod_{\tau=t+1}^{T} \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) \left| \begin{matrix} \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) \\ -\phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) \end{matrix} \right| \Bigg] \\
&\cdot dy_{i,0:T} dx_{i,0:T-1}^* dw_{i,0:T-1}^I \cdot f_0\left(\lambda_i\right) d\lambda_i.
\end{aligned}$$

Same as the iterated integral argument for the first term in part (b) based on the tail conditions in Assumption 3.8, we can establish that for all $\epsilon' > 0$, there exists $\delta > 0$, such that $B < \epsilon'$ as long as

[49] We let δ depend on ϵ in part (c) because part (a) holds for all $\delta > 0$.

$\|\vartheta - \vartheta_0\|_2 < \delta$. Note that

$$\|f - f_0\|_1 - \epsilon' < A - B \leq \|g - g_0\|_1 \leq A + B < \|f - f_0\|_1 + \epsilon',$$

then the distance between g and g_0 is comparable to the distance between f and f_0.

More rigorously, let $\epsilon_1 = \epsilon/9$ and $F_1 = \{f \in \mathcal{F} : \|f - f_0\|_1 < \epsilon = 9\epsilon_1\}$. For small $\eta \in \left(0, \frac{1}{9}\right)$, define $\delta_\eta(\epsilon)$ such that $B < \eta\epsilon$ as long as $\|\vartheta - \vartheta_0\|_2 < \delta_\eta(\epsilon)$, then when $f \in F_1^c$,

$$\|g - g_0\|_1 > \|f - f_0\|_1 - \eta\epsilon > 8\epsilon_1.$$

Let $\mathcal{G}$ be the space induced by $f \in \mathcal{F}$ and $\|\vartheta - \vartheta_0\|_2 < \delta_\eta(\epsilon)$ according to the likelihood function in equation (B.4), then the covering number

$$\mathcal{N}(\epsilon_1, G) \leq \mathcal{N}(\epsilon_1(1 - 9\eta), F),$$

where $G \in \mathcal{G}$ induced by $F \in \mathcal{F}$.

Further define

$$R_N(\vartheta, f) = \prod_{i \leq N} \frac{g\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1} | D_A\right)}{g_0\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1} | D_A\right)},$$

and H_N as the event where

$$\int R_N(\vartheta, f)\, d\Pi(\vartheta, f) \geq \exp\left(-\gamma_0 N \epsilon_1^2\right),\ \gamma_0 = \gamma(1 - 9\eta)^2,$$

for the γ in Assumption 3.9 (2-b). Equation (B.13) implies that $\mathbb{P}_0^N(H_N | D_A) \to 1$, and hence

$$\begin{aligned}
&\mathbb{P}_0^N\left[\Pi\left(F_1^c \bigcap \mathcal{F}_N, \|\vartheta - \vartheta_0\|_2 < \delta_\eta(\epsilon) \middle| D\right) \middle| D_A\right] \\
=&\mathbb{P}_0^N\left[\Pi\left(F_1^c \bigcap \mathcal{F}_N, \|\vartheta - \vartheta_0\|_2 < \delta_\eta(\epsilon) \middle| D\right) \mathbf{1}(H_N) \middle| D_A\right] + o_p(1).
\end{aligned}$$

Note that based on Ghosal and van der Vaart (2007) Corollary 1, for any set $\mathcal{Q}$ with $\inf_{g \in \mathcal{Q}} \|g - g_0\|_1 \geq 8\epsilon_1$,[50] for any $\gamma_1, \gamma_2 > 0$, there exists a test φ_N such that

$$\mathbb{E}_{g_0}^N(\varphi_N | D_A) \leq \sqrt{\frac{\gamma_2}{\gamma_1}} \mathcal{N}(\epsilon_1, \mathcal{Q}) \exp\left(-N\epsilon_1^2\right) \text{ and } \sup_{g \in \mathcal{Q}} \mathbb{E}_g^N(1 - \varphi_N | D_A) \leq \sqrt{\frac{\gamma_1}{\gamma_2}} \exp\left(-N\epsilon_1^2\right).$$

Let G_1 be the induced set by F_1, and $\mathcal{G}_{N,j}$ be the induced set by $\mathcal{F}_{N,j}$. Therefore, we can construct

[50] The original Ghosal and van der Vaart (2007) Corollary 1 considers the Hellinger distance, which is defined as $d_H(g, g_0) = \sqrt{\int \left(\sqrt{g} - \sqrt{g_0}\right)^2}$. Note that $d_H^2(g, g_0) \leq \|g - g_0\|_1 \leq 2 d_H(g, g_0)$, so $\inf_{g \in \mathcal{Q}} d_H(g, g_0) \geq 4\epsilon_1$.

tests $\{\varphi_{N,j}\}$, such that

$$\begin{aligned}
&\mathbb{P}_0^N\left[\Pi\left(F_1^c\bigcap\mathcal{F}_{N,j},\|\vartheta-\vartheta_0\|_2<\delta_\eta(\epsilon)\middle|D\right)\mathbf{1}(H_N)\middle|D_A\right]\\
\leq&\mathbb{P}_0^N(\varphi_{N,j}|D_A)+\mathbb{P}_0^N\left[(1-\varphi_{N,j})\int R_N(\vartheta,f)\,\mathbf{1}\left(F_1^c\bigcap\mathcal{F}_{N,j},\|\vartheta-\vartheta_0\|_2<\delta_\eta(\epsilon)\right)d\Pi(\vartheta,f)\middle|D_A\right]\exp\left(\gamma_0N\epsilon_1^2\right)\\
\leq&\mathbb{E}_{g_0}^N(\varphi_{N,j}|D_A)+\sup_{g\in G_1^c\bigcap\mathcal{G}_{N,j}}\mathbb{E}_g^N(1-\varphi_{N,j}|D_A)\cdot\Pi(\mathcal{F}_{N,j},\|\vartheta-\vartheta_0\|_2<\delta_\eta(\epsilon))\exp\left(\gamma_0N\epsilon_1^2\right)\\
\leq&\mathbb{E}_{g_0}^N(\varphi_{N,j}|D_A)+\sup_{g\in G_1^c\bigcap\mathcal{G}_{N,j}}\mathbb{E}_g^N(1-\varphi_{N,j}|D_A)\cdot\Pi^f(\mathcal{F}_{N,j})\exp\left(\gamma_0N\epsilon_1^2\right)\\
\leq&\sqrt{\frac{\gamma_{2,j}}{\gamma_{1,j}}}\mathcal{N}(\epsilon_1,\mathcal{G}_{N,j})\exp\left(-N\epsilon_1^2\right)+\sqrt{\frac{\gamma_{1,j}}{\gamma_{2,j}}}\Pi^f(\mathcal{F}_{N,j})\exp\left(-N\epsilon_1^2(1-\gamma_0)\right)\\
=&\sqrt{\mathcal{N}(\epsilon_1,\mathcal{G}_{N,j})\Pi^f(\mathcal{F}_{N,j})}\left(\exp\left(-N\epsilon_1^2\right)+\exp\left(-N\epsilon_1^2(1-\gamma_0)\right)\right).
\end{aligned}$$

The last line is given by plugging in $\gamma_{1,j}=\mathcal{N}(\epsilon_1,\mathcal{G}_{N,j})$ and $\gamma_{2,j}=\Pi^f(\mathcal{F}_{N,j})$. Finally, as $\mathcal{N}(\epsilon_1,\mathcal{G}_{N,j})$ less than $\mathcal{N}\left(\frac{\epsilon_1}{1+\eta},\mathcal{F}_{N,j}\right)$,

$$\begin{aligned}
&\mathbb{P}_0^N\left[\Pi\left(F_1^c\bigcap\mathcal{F}_N,\|\vartheta-\vartheta_0\|_2<\delta_\eta(\epsilon)\middle|D\right)\mathbf{1}(H_N)\middle|D_A\right]\\
\leq&O\left(\sum_j\sqrt{\mathcal{N}(\epsilon_1,\mathcal{G}_{N,j})\Pi^f(\mathcal{F}_{N,j})}\exp\left(-N\epsilon_1^2(1-\gamma_0)\right)\right)\\
\leq&O\left(\sum_j\sqrt{\mathcal{N}(\epsilon_1(1-9\eta),\mathcal{F}_{N,j})\Pi^f(\mathcal{F}_{N,j})}\exp\left(-N\epsilon_1^2(1-\gamma_0)\right)\right)\\
=&o\left(\exp\left((1-\gamma)(1-9\eta)^2N\epsilon_1^2\right)\exp\left(-N\epsilon_1^2(1-\gamma_0)\right)\right)\\
=&o\left(\exp\left(-N\epsilon_1^2\left(1-(1-9\eta)^2\right)\right)\right)\\
\to&0.
\end{aligned}$$

The fourth line follows the summability condition of covering numbers as in Lemma 3.7 condition 2-b, which can be deduced from Assumption 3.9 (2-b).

2. Random coefficients: cross-sectional heteroskedasticity

(a) For all $\delta>0$, $\Pi^\vartheta(\|\vartheta-\vartheta_0\|_2\geq\delta|D)\to0$.

After orthogonal forward differencing, the posterior of β is given by

$$p(\beta|D) \propto \int \phi\left(\beta; m^\beta, \psi^\beta \sigma_i^2\right) f^{\sigma^2}\left(\sigma_i^2\right) d\Pi^{\beta, f^{\sigma^2}}\left(\beta, f^{\sigma^2}\right) d\sigma_i^2,$$
$$\psi^\beta = \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}\right)^{-1},$$
$$m^\beta = \psi^\beta \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{y}_{it}\right).$$

Note that $\int \phi\left(\beta; m^\beta, \psi^\beta \sigma_i^2\right) f^{\sigma^2}\left(\sigma_i^2\right) d\Pi^{f^{\sigma^2}}\left(f^{\sigma^2}\right) d\sigma_i^2$ is a scale mixture of normals, and its variance is $\psi^\beta \mathbb{E}\sigma_i^2 \leq \psi^\beta \bar{\sigma}^2$, $\bar{\sigma}^2$ is the upper bound specified in Proposition 3.10 condition 4-b. Therefore, $\beta|\, D$ converges to β_0 given Assumption 3.1 (2-c) and Proposition 3.10 condition 3.

(b) For all $\epsilon > 0$, $\Pi^f\left(\{\|f - f_0\|_1 \geq \epsilon\} \bigcap \mathcal{F}_N^c |\, D\right) \to 0$.

As λ and σ^2 are independent, we have

$$d_{KL}(f_0, f) = d_{KL}\left(f_0^\lambda f_0^{\sigma^2}, f^\lambda f^{\sigma^2}\right) = d_{KL}\left(f_0^\lambda, f^\lambda\right) + d_{KL}\left(f_0^{\sigma^2}, f^{\sigma^2}\right).$$

In addition, since the KL divergence is invariant under variable transformations,

$$d_{KL}\left(f_0^{\sigma^2}, f^{\sigma^2}\right) = d_{KL}\left(f_0^l, f^l\right).$$

Assumption 3.9 (1, 2-a) ensures that the KL property holds for f.

Now the individual-specific likelihood function is

$$\begin{aligned} & g\left(y_{i,0:T}, x^*_{i,0:T-1}, w^I_{i,0:T-1} | D_A\right) \\ = & \prod_t p\left(x^{P*}_{i,t-1} | y_{i,t-1}, c_{i,0:t-2}\right) p\left(c^*_{i0} | D_A\right) \int \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_i^2\right) f\left(\lambda_i, \sigma_i^2\right) d\lambda_i d\sigma_i^2, \end{aligned} \tag{B.14}$$

and we want to prove the altered KL property specified on the observables in expression (B.5). As in the proof of part (1-b), similar convexity reasoning and tail conditions can be applied to bound the KL divergency of g with respect to g_0. Note that all bounds are proportional to $\left(\sigma_i^2\right)^{-1}$, which is further integrated out via $f_0^{\sigma^2}\left(\sigma_i^2\right)$. This integration exists due to the integrability of $f_0^l(l_i)$.

(c) For all $\epsilon > 0$, $\Pi\left(\{\|f - f_0\|_1 \geq \epsilon\} \bigcap \mathcal{F}_N, \|\vartheta - \vartheta_0\|_2 < \delta(\epsilon) |\, D\right) \to 0$.[51]

We can show that the distance between g and g_0 is comparable to the distance between f and f_0 in a similar manner to part (1-c). The rest of the proof remains the same. □

[51] We let δ depend on ϵ in part (c) because part (a) holds for all $\delta > 0$.

B.4 Posterior Consistency: Correlated Random Coefficients Model

Proof. **(Proposition 3.13)**

The proof builds on Pati *et al.* (2013)'s study on univariate conditional density estimation and introduces two major extensions: (1) multivariate conditional density estimation based on location-scale mixture, and (2) deconvolution and dynamic panel data structure.

Part (a) for common parameters is the same as the random coefficients cases.

Parts (b) and (c) for the underlying distribution of individual heterogeneity need more careful treatment. First, we replace $f(\cdot)$ with its conditional counterpart $f(\cdot|c_{i0})$ in the individual-specific likelihoods in equations (B.4) and (B.14).

Second, for $z = \lambda$ (and l), Assumption 3.12 condition 1 (on f_0^z), conditions 2-a,b (on G_0^z), and condition 3-a (on stick breaking process) ensure the induced q_0-integrated KL property on the conditional distribution of z_i, i.e. for all $\epsilon > 0$,

$$\Pi^{f^z}\left(f^z \in \mathcal{F}^z : \int\left[\int f_0(z|c_0)\log\frac{f_0(z|c_0)}{f(z|c_0)}dz\right]q_0(c_0)\,dc_0 < \epsilon\right) > 0.$$

Pati *et al.* (2013) Theorem 5.3 proved it for univariate z. For multivariate z, we work with the spectral norm for the positive definite component covariance matrices and consider $\|\Omega\|_2 \in [\underline{\sigma}, \bar{\sigma}]$ as the approximating compact set in the proof of Pati *et al.* (2013) Lemma 5.5, Theorem 5.6, and Corollary 5.7.

Third, Assumption 3.12 condition 2-c (on G_0^z) and conditions 3-b,c (on stick breaking process) address the sieve property stated in Lemma 3.7 (2). Now the covering number is based on the induced q_0-integrated L_1-distance $\|f^z - f_0^z\|_1 \equiv \int\left[\int |f^z(z|c_0) - f_0^z(z|c_0)|\,dz\right]q_0(c_0)\,dc_0$. Condition 2-c resembles the random coefficients cases while expands component means to include coefficients on c_{i0}. Comparing to Pati *et al.* (2013) Theorem 5.10, condition 2-c here imposes weaker tail conditions on G_0^z and hence is able to accommodate multivariate normal inverse Wishart components. Conditions 3-b,c on stick breaking process directly follow Pati *et al.* (2013) Remark 5.12 and Lemma 5.15. The rest of the proof parallels the random coefficients cases. □

B.5 Density Forecasts

Proof. **(Proposition 3.14)**

1. Random coefficients: cross-sectional homoskedasticity

In this part, I am going to prove that for any i and any $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\left\|f_{i,T+1}^{cond} - f_{i,T+1}^{oracle}\right\|_1 < \epsilon \middle| D\right) \to 1.$$

Following the definitions in Section 2.2, we have

$$
\begin{aligned}
&\int \left| f_{i,T+1}^{cond}\left(y|\vartheta, f\right) - f_{i,T+1}^{oracle}\left(y\right) \right| dy \\
&= \int \left| \int p\left(y|\, h_i, \vartheta, w_{iT}, x_{iT}\right) p\left(h_i\,|\vartheta, f, D_i, D_A\right) dh_i - \int p\left(y|\, h_i, \vartheta_0, w_{iT}, x_{iT}\right) p\left(h_i\,|\vartheta_0, f_0, D_i, D_A\right) dh_i \right| dy \\
&= \int \left| \frac{\int p\left(y|\, h_i, \vartheta, w_{iT}, x_{iT}\right) \prod_t p\left(y_{it}|\, h_i, \vartheta, w_{i,t-1}, x_{i,t-1}\right) f\left(h_i\right) dh_i}{\int \prod_t p\left(y_{it}|\, h_i, \vartheta, w_{i,t-1}, x_{i,t-1}\right) f\left(h_i\right) dh_i} \right. \\
&\quad \left. - \frac{\int p\left(y|\, h_i, \vartheta_0, w_{iT}, x_{iT}\right) \prod_t p\left(y_{it}|\, h_i, \vartheta_0, w_{i,t-1}, x_{i,t-1}\right) f_0\left(h_i\right) dh_i}{\int \prod_t p\left(y_{it}|\, h_i, \vartheta_0, w_{i,t-1}, x_{i,t-1}\right) f_0\left(h_i\right) dh_i} \right| dy \qquad \text{(B.15)} \\
&= \int \left| \frac{\int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right) f\left(\lambda_i\right) d\lambda_i}{\int \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right) f\left(\lambda_i\right) d\lambda_i} \right. \\
&\quad \left. - \frac{\int \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_0^2\right) \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_0^2\right)\right) f_0\left(\lambda_i\right) d\lambda_i} \right| dy.
\end{aligned}
$$

To obtain the last equality, we first rewrite $\prod_t p\left(y_{it}\left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right)$ as a distribution of λ_i

$$
\prod_t p\left(y_{it}\left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) = C\left(\beta, \sigma^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right),
$$

where

$$
\begin{aligned}
m\left(\beta\right) &= \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1} \sum_t w_{i,t-1}\left(y_{it} - \beta' x_{i,t-1}\right), \\
\Sigma\left(\sigma^2\right) &= \sigma^2 \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1}, \\
C_A\left(\beta, \sigma^2\right) &= \left(\left(2\pi\sigma^2\right)^{\frac{T-d_w}{2}} \left|\sum_t w_{i,t-1} w_{i,t-1}'\right|\right)^{-1} \\
&\quad \cdot \exp\left(-\frac{1}{2}\left(\frac{\sum_t \left(y_{it} - \beta' x_{i,t-1}\right)^2}{\sigma^2} - m\left(\beta\right)' \left(\Sigma\left(\sigma^2\right)\right)^{-1} m\left(\beta\right)\right)\right),
\end{aligned}
$$

and then cross out the common factor in the numerator and denominator. Set

$$
\begin{aligned}
A &= \int \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right) f\left(\lambda_i\right) d\lambda_i, \\
B\left(y\right) &= \int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right) f\left(\lambda_i\right) d\lambda_i.
\end{aligned}
$$

with A_0 and $B_0\left(y\right)$ being the counterparts for the oracle predictor. Note that both A and $B\left(y\right)$ are

positive by definition. Then, we want to make sure the following expression is arbitrarily small,

$$\int \left| \frac{B(y)}{A} - \frac{B_0(y)}{A_0} \right| dy \leq \frac{\int B_0(y)\, dy \cdot |A - A_0|}{A_0 A} + \frac{\int |B(y) - B_0(y)|\, dy}{A},$$

and it is sufficient to establish the following four statements (with probability approaching one).

(a) $|A - A_0| < \epsilon'$

$$\begin{aligned} & |A - A_0| \\ \leq & \left| \int \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| \\ & + \left| \int \left(\phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) \right) f_0(\lambda_i)\, d\lambda_i \right|. \end{aligned}$$

For the first term,

$$\begin{aligned} & \left| \int \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| \\ \leq & \int \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left|f(\lambda_i) - f_0(\lambda_i)\right| d\lambda_i \\ \leq & \frac{\left|\sum_t w_{i,t-1} w'_{i,t-1}\right|}{\left(2\pi\sigma^2\right)^{d_w/2}} \cdot \left\|f - f_0\right\|_1. \end{aligned} \tag{B.16}$$

It is less than $\epsilon'/2$ with probability approaching one due to the posterior consistency of f and that $\phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right)$ is a bounded function in λ_i (given that σ^2 converges to σ_0^2). For the second term,

$$\begin{aligned} & \left| \int \left(\phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) \right) f_0(\lambda_i)\, d\lambda_i \right| \\ \leq & M \int \left| \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) \right| d\lambda_i \\ \leq & M \sqrt{2 d_{KL}\left(\phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right), \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right)\right)} \\ \leq & M \sqrt{d_w \left(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln \frac{\sigma^2}{\sigma_0^2} \right) + \sigma_0^{-2} (\beta - \beta_0)' \left[\sum_t x_{i,t-1} w'_{i,t-1} \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \sum_t w_{i,t-1} x'_{i,t-1} \right] (\beta - \beta_0)}. \end{aligned}$$

where the second inequality follows Pinsker's inequality that bounds the L_1-distance by the KL divergence. As $\left(\beta, \sigma^2\right)$ enjoys posterior consistency, the last expression can be arbitrarily small. Therefore, the second term is less than $\epsilon'/2$ with probability approaching one.

(b) $\int |B(y) - B_0(y)|\, dy < \epsilon'$

$$\begin{aligned}
&\int |B(y) - B_0(y)|\, dy \\
\leq &\int \left| \int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| dy \\
&+ \int \left| \int \left(\begin{array}{l} \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \\ - \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_0^2\right) \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) \end{array} \right) f_0(\lambda_i)\, d\lambda_i \right| dy.
\end{aligned}$$

Similar to part (a), the first term is small due to the posterior consistency of f and σ^2,

$$\begin{aligned}
&\int \left| \int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| dy \\
\leq &\int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left|f(\lambda_i) - f_0(\lambda_i)\right| d\lambda_i dy \\
= &\int \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \left|f(\lambda_i) - f_0(\lambda_i)\right| d\lambda_i,
\end{aligned}$$

which is the same as expression (B.16) in part (a). Pinsker's inequality together with the posterior consistency of $\left(\beta, \sigma^2\right)$ ensures a small second term,

$$\begin{aligned}
&\int \left| \int \left(\begin{array}{l} \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) \\ - \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_0^2\right) \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) \end{array} \right) f_0(\lambda_i)\, d\lambda_i \right| dy \\
\leq &\int \left|\phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) - \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_0^2\right)\right| \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right) f_0(\lambda_i)\, d\lambda_i dy \\
&+ \int \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) \left|\phi\left(\lambda_i; m(\beta), \Sigma\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_0^2\right)\right)\right| f_0(\lambda_i)\, d\lambda_i dy \\
\leq &M \sqrt{\frac{\sigma^2}{\sigma_0^2} - 1 - \ln \frac{\sigma^2}{\sigma_0^2} + \sigma_0^{-2} (\beta - \beta_0)' x_{iT} x_{iT}' (\beta - \beta_0)} \\
&+ M \sqrt{d_w \left(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln \frac{\sigma^2}{\sigma_0^2} \right) + \sigma_0^{-2} (\beta - \beta_0)' \left[\sum_t x_{i,t-1} w_{i,t-1}' \left(\sum_t w_{i,t-1} w_{i,t-1}' \right)^{-1} \sum_t w_{i,t-1} x_{i,t-1}' \right] (\beta - \beta_0)}.
\end{aligned}$$

(c) There exists $\underline{A} > 0$ such that $A_0 > \underline{A}$.

Let μ_0^λ and V_0^λ be the mean and variance of λ_i based on the true distribution f_0. The moment condition in Assumption 3.9 (1-e) ensures the existence of both μ_0^λ and V_0^λ. Following Chebyshev's inequality, we have

$$\mathbb{P}_{f_0}\left(\sqrt{\left(\lambda_i - \mu_0^\lambda\right)' \left(V_0^\lambda\right)^{-1} \left(\lambda_i - \mu_0^\lambda\right)} > k \right) \leq \frac{d_w}{k^2}.$$

Define $K^{\lambda}=\left\{\lambda_i:\ \sqrt{\left(\lambda_i-\mu_0^{\lambda}\right)'\left(V_0^{\lambda}\right)^{-1}\left(\lambda_i-\mu_0^{\lambda}\right)}\leq k\right\}$. Then,

$$\begin{aligned}A_0&=\int\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f_0\left(\lambda_i\right)d\lambda_i\\&\geq\int_{\lambda_i\in K^{\lambda}}\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f_0\left(\lambda_i\right)d\lambda_i\\&\geq\left(1-\frac{d_w}{k^2}\right)\min_{\lambda_i\in K^{\lambda}}\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right).\end{aligned}$$

Intuitively, since $\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)$ and $f_0\left(\lambda_i\right)$ share the same support on $\mathbb{R}^{d_w}$, 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, A is also bounded below with probability approaching one.

(d) $\int B_0\left(y\right)dy<\infty$

$$\begin{aligned}&\int B_0\left(y\right)dy\\&=\int\phi\left(y;\beta_0'x_{iT}+\lambda_i'w_{iT},\sigma_0^2\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f_0\left(\lambda_i\right)d\lambda_i dy\\&=\int\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right)f_0\left(\lambda_i\right)d\lambda_i\\&\leq\frac{\left|\sum_t w_{i,t-1}w_{i,t-1}'\right|}{\left(2\pi\sigma_0^2\right)^{d_w/2}}\int f_0\left(\lambda_i\right)d\lambda_i\\&=\frac{\left|\sum_t w_{i,t-1}w_{i,t-1}'\right|}{\left(2\pi\sigma_0^2\right)^{d_w/2}}.\end{aligned}$$

2. Random coefficients: cross-sectional heteroskedasticity

Now A and $B\left(y\right)$ become

$$\begin{aligned}A&=\int C\left(\beta,\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i,\\B\left(y\right)&=\int\phi\left(y;\beta'x_{iT}+\lambda_i'w_{iT},\sigma_i^2\right)C\left(\beta,\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2.\end{aligned}$$

Consider Proposition 3.14 condition 3 ($\text{supp}\left(f_0^{\sigma^2}\right)$ is bounded below by some $\underline{\sigma}^2>0$), the above statements can be derived as follows.

(a) $|A - A_0| < \epsilon'$

$$
\begin{aligned}
& |A - A_0| \\
\leq & \left| \int C\left(\beta, \sigma_i^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) \left(f^{\lambda}\left(\lambda_i\right) f^{\sigma^2}\left(\sigma_i^2\right) - f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) \right) d\lambda_i d\sigma_i^2 \right| \\
& + \left| \int C\left(\beta, \sigma_i^2\right) \left(\phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) - \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_i^2\right)\right) \right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \right| \\
& + \left| \int \left(C\left(\beta, \sigma_i^2\right) - C\left(\beta_0, \sigma_i^2\right) \right) \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \right|.
\end{aligned}
$$

The first term

$$
\begin{aligned}
& \left| \int C\left(\beta, \sigma_i^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) \left(f^{\lambda}\left(\lambda_i\right) f^{\sigma^2}\left(\sigma_i^2\right) - f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) \right) d\lambda_i d\sigma_i^2 \right| \\
\leq & \int C\left(\beta, \sigma_i^2\right) \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) \left| f^{\lambda}\left(\lambda_i\right) f^{\sigma^2}\left(\sigma_i^2\right) - f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) \right| d\lambda_i d\sigma_i^2 \qquad \text{(B.17)} \\
\leq & \left(2\pi \underline{\sigma}^2\right)^{-T/2} \cdot \left\| f - f_0 \right\|_1 .
\end{aligned}
$$

The second term

$$
\begin{aligned}
& \left| \int C\left(\beta, \sigma_i^2\right) \left(\phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) - \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_i^2\right)\right) \right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \right| \\
\leq & M \int C\left(\beta, \sigma_i^2\right) \left| \phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma_i^2\right)\right) - \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_i^2\right)\right) \right| f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \\
= & M \int C\left(\beta, \sigma_i^2\right) \sqrt{\sigma_i^{-2} \left(\beta - \beta_0\right)' V_2 \left(\beta - \beta_0\right)} f_0^{\sigma^2}\left(\sigma_i^2\right) d\sigma_i^2 \\
\leq & M_2 \left(\underline{\sigma}^2\right)^{-\frac{T - d_w + 1}{2}} \sqrt{\left(\beta - \beta_0\right)' V_2 \left(\beta - \beta_0\right)} \int f_0^{\sigma^2}\left(\sigma_i^2\right) d\sigma_i^2 \\
= & M_2 \left(\underline{\sigma}^2\right)^{-\frac{T - d_w + 1}{2}} \sqrt{\left(\beta - \beta_0\right)' V_2 \left(\beta - \beta_0\right)},
\end{aligned}
$$

where

$$
\begin{aligned}
M_2 &= M \left(\left(2\pi\right)^{\frac{T - d_w}{2}} \left| \sum_t w_{i,t-1} w_{i,t-1}' \right| \right)^{-1}, \\
V_2 &= \sum_t x_{i,t-1} w_{i,t-1}' \left(\sum_t w_{i,t-1} w_{i,t-1}' \right)^{-1} \sum_t w_{i,t-1} x_{i,t-1}'.
\end{aligned}
$$

The third term

$$
\begin{aligned}
&\left|\int \left(C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right)f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2\right| \\
=&\int\left|C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right|\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right)f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2 \qquad \text{(B.18)} \\
\leq& M\int\left|C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right|\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right)f_0^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2 \\
\leq& M\int\left|C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right|f_0^{\sigma^2}\left(\sigma_i^2\right)d\sigma_i^2 \\
\leq& M_2\int\left(\sigma_i^2\right)^{-\frac{T-d_w}{2}}\frac{\left|C_3\left(\beta_0\right)-C_3\left(\beta\right)\right|}{2\sigma_i^2}f_0^{\sigma^2}\left(\sigma_i^2\right)d\sigma_i^2 \\
\leq& \frac{1}{2}M_2\left(\underline{\sigma}^2\right)^{-\frac{T-d_w+2}{2}}\left|C_3\left(\beta_0\right)-C_3\left(\beta\right)\right|.
\end{aligned}
$$

where $C_3\left(\beta\right)=\sum_t\left(y_{it}-\beta' x_{i,t-1}\right)^2-m\left(\beta\right)\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)m\left(\beta\right)$ is continuous in β.

(b) $\int\left|B\left(y\right)-B_0\left(y\right)\right|dy<\epsilon'$

$$
\begin{aligned}
&\int\left|B\left(y\right)-B_0\left(y\right)\right|dy \\
\leq&\int\left|\int\phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)C\left(\beta,\sigma_i^2\right)\left(f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)-f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)\right)d\lambda_i d\sigma_i^2\right|dy \\
&+\int\left|\int\left(\begin{array}{l}\phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right) \\ -\phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta_0\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right)\end{array}\right)C\left(\beta,\sigma_i^2\right)f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2\right|dy \\
&+\int\left|\int\phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\left(C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right)f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)d\lambda_i d\sigma_i^2\right|dy.
\end{aligned}
$$

The first term

$$
\begin{aligned}
&\int\left|\int\phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)C\left(\beta,\sigma_i^2\right)\left(f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)-f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)\right)d\lambda_i d\sigma_i^2\right|dy \\
\leq&\int\phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)C\left(\beta,\sigma_i^2\right)\left|f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)-f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)\right|d\lambda_i d\sigma_i^2 dy \\
=&\int\phi\left(\lambda_i;m\left(\beta\right),\Sigma\left(\sigma_i^2\right)\right)C\left(\beta,\sigma_i^2\right)\left|f^{\lambda}\left(\lambda_i\right)f^{\sigma^2}\left(\sigma_i^2\right)-f_0^{\lambda}\left(\lambda_i\right)f_0^{\sigma^2}\left(\sigma_i^2\right)\right|d\lambda_i d\sigma_i^2,
\end{aligned}
$$

which is the same as expression (B.17) in part (a). The second term

$$\int \left| \int \begin{pmatrix} \phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right) \\ -\phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta_0\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right) \end{pmatrix} C\left(\beta,\sigma_i^2\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \right| dy$$

$$\leq \int \left| \begin{matrix} \phi\left(y;\beta' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right) \\ -\phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\phi\left(\lambda_i;m_i^{\lambda}\left(\beta_0\right),\Sigma_i^{\lambda}\left(\sigma_i^2\right)\right) \end{matrix} \right| C\left(\beta,\sigma_i^2\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 dy$$

$$= M \int C\left(\beta,\sigma_i^2\right)\left(\sqrt{\sigma_i^{-2}\left(\beta-\beta_0\right)' x_{iT} x_{iT}'\left(\beta-\beta_0\right)}+\sqrt{\sigma_i^{-2}\left(\beta-\beta_0\right)' V_2\left(\beta-\beta_0\right)}\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\sigma_i^2$$

$$\leq M_2\left(\underline{\sigma}^2\right)^{-\frac{T-d_w+1}{2}}\left(\sqrt{\left(\beta-\beta_0\right)' x_{iT} x_{iT}'\left(\beta-\beta_0\right)}+\sqrt{\left(\beta-\beta_0\right)' V_2\left(\beta-\beta_0\right)}\right)\int f_0^{\sigma^2}\left(\sigma_i^2\right) d\sigma_i^2$$

$$= M_2\left(\underline{\sigma}^2\right)^{-\frac{T-d_w+1}{2}}\left(\sqrt{\left(\beta-\beta_0\right)' x_{iT} x_{iT}'\left(\beta-\beta_0\right)}+\sqrt{\left(\beta-\beta_0\right)' V_2\left(\beta-\beta_0\right)}\right).$$

The third term

$$\int \left| \int \phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\left(C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \right| dy$$

$$\leq \int \phi\left(y;\beta_0' x_{iT}+\lambda_i' w_{iT},\sigma_i^2\right)\left|C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right|\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 dy$$

$$= \int \left|C\left(\beta,\sigma_i^2\right)-C\left(\beta_0,\sigma_i^2\right)\right|\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2$$

which is the same as expression (B.18) in part (a).

(c) There exists $\underline{A}>0$ such that $A_0>\underline{A}$.

Let $l_i=\log\left(\sigma_i^2-\underline{\sigma}^2\right)$, μ_0^l and V_0^l be the mean and variance of l_i based on the true distribution f_0^l, and $K^l=\left\{l_i:\ \frac{|l_i-\mu_0^l|}{\sqrt{V_0^l}}\leq k\right\}$. Then,

$$\begin{aligned} A_0 &= \int C\left(\beta_0,\sigma_i^2\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \\ &= \int C\left(\beta_0,\exp l_i+\underline{\sigma}^2\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\exp l_i+\underline{\sigma}^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{l}\left(l_i\right) d\lambda_i dl_i \\ &\geq \int_{\lambda_i\in K^{\lambda},l_i\in K^l} C\left(\beta_0,\exp l_i+\underline{\sigma}^2\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\exp l_i+\underline{\sigma}^2\right)\right) f_0^{\lambda}\left(\lambda_i\right) f_0^{l}\left(l_i\right) d\lambda_i dl_i \\ &\geq \left(1-\frac{d_w}{k^2}\right)\left(1-\frac{1}{k^2}\right)\min_{\lambda_i\in K^{\lambda},l_i\in K^l} C\left(\beta_0,\exp l_i+\underline{\sigma}^2\right)\phi\left(\lambda_i;m\left(\beta_0\right),\Sigma\left(\sigma_0^2\right)\right), \end{aligned}$$

where the second line is given by the change of variables, and the last line follows Chebyshev's inequality on both λ_i and l_i.

(d) $\int B_0(y)\,dy < \infty$

$$\begin{aligned}
&\int B_0(y)\,dy \\
&= \int \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_i^2\right) C\left(\beta_0, \sigma_i^2\right) \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}(\lambda_i)\, f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 dy \\
&= \int C\left(\beta_0, \sigma_i^2\right) \phi\left(\lambda_i; m(\beta_0), \Sigma\left(\sigma_i^2\right)\right) f_0^{\lambda}(\lambda_i)\, f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \\
&\leq \left(2\pi \underline{\sigma}^2\right)^{-T/2} \int f_0^{\lambda}(\lambda_i)\, f_0^{\sigma^2}\left(\sigma_i^2\right) d\lambda_i d\sigma_i^2 \\
&= \left(2\pi \underline{\sigma}^2\right)^{-T/2}.
\end{aligned}$$

3. Correlated random coefficients

Now we replace $f(h_i)$ with $f(h_i|c_{i0})$ in expression (B.15) that characterizes $\left\| f_{i,T+1}^{cond} - f_{i,T+1}^{oracle} \right\|_1$. Given Proposition 3.14 condition 2-b ($q_0(c_0)$ is bounded below by some $\underline{q} > 0$), we have

$$\underline{q}\left[\int |f(z|c_0) - f_0(z|c_0)|\, dz\right] < \int \left[\int |f(z|c_0) - f_0(z|c_0)|\, dz\right] q_0(c_0)\, dc_0 < \epsilon,$$

so

$$\int |f(z|c_0) - f_0(z|c_0)|\, dz < \epsilon/\underline{q}.$$

Therefore, we achieve the convergence of conditional distribution for any c_0 and ensure that the first term in part (a) is sufficiently small. The rest of the proof parallels the random coefficients scenarios. □

C Algorithms

C.1 Hyperparameters

Let us take the baseline model with random effects as an example, and the priors and hyperparameters for more complicated models can be constructed in a similar way. The prior for the common parameters takes a conjugate norma-inverse-gamma form,

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

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.

$$a_0^{\sigma^2} = 2, \tag{C.1}$$

$$b_0^{\sigma^2} = \hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right) \cdot \left(a_0^{\sigma^2} - 1\right) = \hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right), \tag{C.2}$$

$$m_0^{\beta} = 0.5, \tag{C.3}$$

$$\psi_0^{\beta} = \frac{1}{b_0^{\sigma^2} / \left(a_0^{\sigma^2} - 1\right)} = \frac{1}{\hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right)}. \tag{C.4}$$

In equation (C.2) here and equation (C.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 (C.2) ensures that on average the prior and the data have a similar scale. Equation (C.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 (C.1) characterizes a rather less informative prior on σ^2 with infinite variance, and Equation (C.4) assumes that the prior variance of β is equal to 1 on average.

The hyperpriors for the DPM prior are specified as:

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

Similarly, the hyperparameters are chosen to be:

$$a_0^{\lambda} = 2,\ b_0^{\lambda} = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right) \cdot \left(a_0^{\lambda} - 1\right) = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right), \tag{C.5}$$

$$m_0^{\lambda} = 0,\ \psi_0^{\lambda} = 1,$$

$$a_0^{\alpha} = 2,\ b_0^{\alpha} = 2. \tag{C.6}$$

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 (C.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:

$$\left\| f^{\lambda,K} - f^{\lambda} \right\|_1 \sim 4N \exp\left(-\frac{K-1}{\alpha}\right) \leq 2.10 \times 10^{-18},$$

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.

C.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), who adaptively adjust the random walk step size in order to keep acceptance rates around certain desirable percentage.

Algorithm C.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 C.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).

C.3 Details on Posterior Samplers

C.3.1 Step 2: Component Parameters

Random Coefficients Model For $z = \lambda, l$ and $k = 1, \cdots, K^z$, draw $\left(\mu_k^{z(s)}, \Omega_k^{z(s)}\right)$ from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a

scalar) $p\left(\mu_k^{z(s)}, \Omega_k^{z(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_k^{z(s-1)}}\right)$:

$$\begin{aligned}
\left(\mu_k^{z(s)}, \Omega_k^{z(s)}\right) &\sim N\left(m_k^z, \psi_k^z \Omega_k^{z(s)}\right) \mathrm{IW}\left(\Psi_k^z, \nu_k^z\right), \\
\hat{m}_k^z &= \frac{1}{n_k^{z(s-1)}} \sum_{i \in J_k^{z(s-1)}} z_i^{(s-1)}, \\
\psi_k^z &= \left(\left(\psi_0^z\right)^{-1} + n_k^{z(s-1)}\right)^{-1}, \\
m_k^z &= \psi_k^z \left(\left(\psi_0^z\right)^{-1} m_0^z + \sum_{i \in J_k^{z(s-1)}} z_i^{(s-1)}\right), \\
\nu_k^z &= \nu_0^z + n_k^{z(s-1)}, \\
\Psi_k^z &= \Psi_0^z + \sum_{i \in J_k^{z(s-1)}} \left(z_i^{(s-1)}\right)^2 + m_0^{z\prime} \left(\psi_0^z\right)^{-1} m_0^z - m_k^{z\prime} \left(\psi_k^z\right)^{-1} m_k^z.
\end{aligned}$$

Correlated Random Coefficients Model Due to the complexity arising from the conditional structure, I break the updating procedure for $\left(\mu_k^{z(s)}, \Omega_k^{z(s)}\right)$ into two steps. For $z = \lambda, l$, and $k = 1, \cdots, K^z$,

(a) Draw $\mathrm{vec}\left(\mu_k^{z(s)}\right)$ from a multivariate normal distribution $p\left(\mu_k^{z(s)} \middle| \Omega_k^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_k^{z(s-1)}}\right)$:

$$\begin{aligned}
\mathrm{vec}\left(\mu_k^{z(s)}\right) &\sim N\left(\mathrm{vec}\left(m_k^z\right), \psi_k^z\right), \\
\hat{m}_k^{z,zc} &= \sum_{i \in J_k^{z(s-1)}} z_i^{(s-1)} \left[1, c_{i0}'\right], \\
\hat{m}_k^{z,cc} &= \sum_{i \in J_k^{z(s-1)}} \left[1, c_{i0}'\right]' \left[1, c_{i0}'\right], \\
\hat{m}_k^z &= \hat{m}_k^{z,zc} \left(\hat{m}_k^{z,cc}\right)^{-1}, \\
\psi_k^z &= \left[\left(\psi_0^z\right)^{-1} + \hat{m}_k^{z,cc} \otimes \left(\Omega_k^{z(s-1)}\right)^{-1}\right]^{-1}, \\
\mathrm{vec}\left(m_k^z\right) &= \psi_k^z \left[\left(\psi_0^z\right)^{-1} \mathrm{vec}\left(m_0^z\right) + \left(\hat{m}_k^{z,cc} \otimes \left(\Omega_k^{z(s-1)}\right)^{-1}\right) \mathrm{vec}\left(\hat{m}_k^z\right)\right].
\end{aligned}$$

(b) Draw $\Omega_k^{z(s)}$ from an inverse Wishart distribution (or an inverse gamma distribution if z is a

scalar) $p\left(\Omega_k^{z(s)} \middle| \mu_k^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_k^{z(s-1)}}\right)$:

$$\begin{aligned}
\Omega_k^{z(s)} &\sim \mathrm{IW}\left(\Psi_k^z, \nu_k^z\right), \\
\nu_k^z &= \nu_0^z + n_k^{z(s-1)}, \\
\Psi_k^z &= \Psi_0^z + \sum_{i \in J_k^{z(s-1)}} \left(z_i^{(s-1)} - \mu_k^{z(s)}\left[1, c_{i0}'\right]'\right)\left(z_i^{(s-1)} - \mu_k^{z(s)}\left[1, c_{i0}'\right]'\right)'.
\end{aligned}$$

C.3.2 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) $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A\right)$:

$$\begin{aligned}
\lambda_i^{(s)} &\sim N\left(m_i^\lambda, \Sigma_i^\lambda\right), \\
\Sigma_i^\lambda &= \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=t_{0i}}^{t_{1i}} w_{i,t-1} w_{i,t-1}'\right)^{-1}, \\
m_i^\lambda &= \Sigma_i^\lambda \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} \tilde{\mu}_i^\lambda + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=t_{0i}}^{t_{1i}} w_{i,t-1}\left(y_{it} - \beta^{(s-1)\prime} x_{i,t-1}\right)\right),
\end{aligned}$$

where the conditional "prior" mean is characterized by

$$\tilde{\mu}_i^\lambda = \begin{cases} \mu_{\gamma_i^\lambda}^{\lambda(s)}, & \text{for the random coefficients model,} \\ \mu_{\gamma_i^\lambda}^{\lambda(s)}\left[1, c_{i0}'\right]', & \text{for the correlated random coefficients model.} \end{cases}$$

C.3.3 Step 5: Common parameters

Cross-sectional Homoskedasticity Draw $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ from a linear regression model with "unknown" variance, $p\left(\beta^{(s)}, \sigma^{2(s)} \middle| \left\{\lambda_i^{(s)}\right\}, D\right)$:

$$\begin{aligned}
\left(\beta^{(s)}, \sigma^{2(s)}\right) &\sim N\left(m^{\beta}, \psi^{\beta}\sigma^{2(s)}\right) \mathrm{IG}\left(a^{\sigma^2}, b^{\sigma^2}\right), \\
\psi^{\beta} &= \left(\left(\psi_0^{\beta}\right)^{-1} + \sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}x_{i,t-1}'\right)^{-1}, \\
m^{\beta} &= \psi^{\beta}\left(\left(\psi_0^{\beta}\right)^{-1} m_0^{\beta} + \sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right), \\
a^{\sigma^2} &= a_0^{\sigma^2} + \frac{NT}{2} \\
b^{\sigma^2} &= b_0^{\sigma^2} + \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)^2 + m_0^{\beta\prime}\left(\psi_0^{\beta}\right)^{-1} m_0^{\beta} - m^{\beta\prime}\left(\psi^{\beta}\right)^{-1} m^{\beta}\right).
\end{aligned}$$

Cross-sectional Heteroskedasticity Draw $\beta^{(s)}$ from a linear regression model with "known" variance, $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right)$:

$$\begin{aligned}
\beta^{(s)} &\sim N\left(m^{\beta}, \Sigma^{\beta}\right), \\
\Sigma^{\beta} &= \left(\left(\Sigma_0^{\beta}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1}\sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}x_{i,t-1}'\right)^{-1}, \\
m^{\beta} &= \Sigma^{\beta}\left(\left(\Sigma_0^{\beta}\right)^{-1} m_0^{\beta} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1}\sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right).
\end{aligned}$$

Remark C.3. For unbalanced panels, the summations and products in steps 4 and 5 (Subsections C.3.2 and C.3.3) are instead over $t = t_{0i}, \cdots, t_{1i}$, the observed periods for individual i.

C.4 Slice-Retrospective Sampler

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 (2.7) can be rewritten as

$$z \sim \sum_{k=1}^{\infty} \mathbf{1}\left(u_i^z < p_{ik}^z\right) f^z\left(z; \theta_k^z\right),$$

where $z = \lambda, l$. By marginalizing over u_i^z, we can recover equation (2.7). 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

$$K^{z,P} = \min\left\{k : \left(1 - \sum_{j=1}^{k} p_j^z\right) < \min_{1 \le i \le N} u_i^z\right\}.$$

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 C.4. *(Slice-Retrospective: Random Coefficients with Cross-sectional Heteroskedasticity)*
For each iteration $s = 1, \cdots, n_{sim}$, steps 1-3 in Algorithm 4.1 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 $k = 1, \cdots, K^{z,A}$, draw p_k^{z*} from the stick-breaking process $p\left(\{p_k^{z*}\} \middle| \alpha^{z(s-1)}, \left\{n_k^{z(s-1)}\right\}\right)$:*

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

   (c) *Component parameters: for $k = 1, \cdots, K^{z,A}$, draw θ_k^{z*} from $p\left(\theta_k^{z*} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_k^{z(s-1)}}\right)$ as in Algorithm 4.1 step 2.*
   (d) *Label switching: jointly update $\left\{p_k^{z(s)}, \theta_k^{z(s)}, \gamma_i^{z*}\right\}_{k=1}^{K^{z,A}}$ based on $\left\{p_k^{z*}, \theta_k^{z*}, \gamma_i^{z(s-1)}\right\}_{k=1}^{K^{z,A}}$ 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 (θ_k^z) and component probabilities (p_k^z) unchanged;*

ii. *randomly select two adjacent components, switch their component labels (γ_i^z) and component "stick lengths" (ζ_k^z), while leaving component parameters (θ_k^z) unchanged;*

iii. *randomly select two non-empty components, switch their component labels (γ_i^z) and component parameters (θ_k^z), as well as update their component probabilities (p_k^z).*

Then, adjust $K^{z,A}$ accordingly.

2. *Auxiliary variables: for $i = 1, \cdots, N$, draw $u_i^{z(s)}$ from a uniform distribution $p\left(u_i^{z(s)} \left| \left\{p_k^{z(s)}\right\}, \gamma_i^{z*}\right.\right)$:*

$$u_i^{z(s)} \sim U\left(0, p_{\gamma_i^{z*}}^{z(s)}\right).$$

3. *DP scale parameter:*

   (a) *Draw the latent variable $\xi^{z(s)}$ from a beta distribution $p\left(\xi^{z(s)} \left| \alpha^{z(s-1)}, N\right.\right)$:*

$$\xi^{z(s)} \sim Beta\left(\alpha^{z(s-1)} + 1, N\right).$$

   (b) *Draw $\alpha^{z(s)}$ from a mixture of two gamma distributions $p\left(\alpha^{z(s)} \left| \xi^{z(s)}, K^{z,A}, N\right.\right)$* :Parametric Prior for Heteroskedastic σ_i^2

$$\alpha^{z(s)} \sim p^{\alpha^z} Ga\left(a^{\alpha^z} + K^{z,A}, b^{\alpha^z} - \log \xi^{z(s)}\right) + \left(1 - p^{\alpha^z}\right) Ga\left(a^{\alpha^z} + K^{z,A} - 1, b^{\alpha^z} - \log \xi^{z(s)}\right),$$

$$p^{\alpha^z} = \frac{a^{\alpha^z} + K^{z,A} - 1}{N\left(b^{\alpha^z} - \log \xi^{z(s)}\right)}.$$

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 $k = K^{z,A} + 1, \cdots, K^{z,P}$, draw $\theta_k^{z(s)}$ 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 $p\left(\left\{\gamma_i^{z(s)}\right\} \left| \left\{p_k^{z(s)}, \mu_k^{z(s)}, \Omega_k^{z(s)}\right\}, u_i^{z(s)}, z_i^{(s-1)}\right.\right)$:*

$$\gamma_i^{z(s)} = k, \text{ with probability } p_{ik}^z, \ k = 1, \cdots, K^{z,P},$$

$$p_{ik}^z \propto p_k^{z(s)} \phi\left(z_i^{(s-1)}; \mu_k^{z(s)}, \Omega_k^{z(s)}\right) \mathbf{1}\left(u_i^{z(s)} < p_k^{z(s)}\right), \quad \sum_{k=1}^{K^{z,P}} p_{ik}^z = 1.$$

The remaining part of the algorithm resembles steps 4 and 5 in Algorithm 4.1.

Remark C.5. Note that:
(i) Steps 1-b,c,d are sampling from "marginal" posterior of $(p_k^z, \theta_k^z, \gamma_i^z)$ for the active components with the auxiliary variables u_i^zs 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 4.1 due to the unrestricted number of components in the current sampler.
(iv) Steps 4-a-ii and 4-b that update the potential components are very similar to steps 1-b and 1-c that update the active components—just take J_k^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.

C.5 Parametric Specification of Heteroskedasticity

For Heterosk-Param, we adopt an inverse gamma prior for σ_i^2,

$$\sigma_i^2 \sim \mathrm{IG}\left(a, b\right).$$

The conjugate priors for shape parameter a and scale parameter b are based on Llera and Beckmann (2016) Sections 2.3.1 and 2.3.2:

$$\begin{aligned} b &\sim \mathrm{Ga}\left(a_0^b, b_0^b\right), \\ p\left(a|b, a_0^a, b_0^a, c_0^a\right) &\propto \frac{\left(a_0^a\right)^{-1-a}\left(b\right)^{ac_0^a}}{\Gamma(a)^{b_0^a}}. \end{aligned} \tag{C.7}$$

Following Llera and Beckmann (2016), the hyperparameters are chosen as $a_0^a = 1$, $b_0^a = c_0^a = a_0^b = b_0^b = 0.01$, which specifies relatively uninformative priors for a and b. The corresponding segment of the posterior sampler is given as follows.

Algorithm C.6. *(Parametric Specification: Cross-sectional Heteroskedasticity)*

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

1. *Shape parameter: Draw* $a^{(s)}$ *via the random-walk Metropolis-Hastings approach,*

$$p\left(a^{(s)} \left| b^{(s-1)}, \left\{\sigma_i^{2(s-1)}\right\}\right.\right) = p\left(a^{(s)}|b^{(s-1)}, a_1^a, b_1^a, c_0^a\right),$$

which is characterized by the same kernel form as expression (C.7) with

$$\begin{aligned}\log(a_1^a) &= \log(a_0^a) + \sum_{i=1}^{N} \log(\sigma_i^{2(s-1)}),\\ b_1^a &= b_0^a + N,\\ c_1^a &= c_0^a + N.\end{aligned}$$

2. *Scale parameter: Draw $b^{(s)}$ from a gamma distribution, $p\left(b^{(s)} \left| a^{(s)}, \left\{\sigma_i^{2(s-1)}\right\}\right.\right)$:*

$$\begin{aligned}b^{(s)} &\sim Ga\left(a_1^b, b_1^b\right),\\ a_1^b &= a_0^b + Na^{(s)},\\ b_1^b &= b_0^b + \sum_{i=1}^{N}\left(\sigma_i^{2(s-1)}\right)^{-1}.\end{aligned}$$

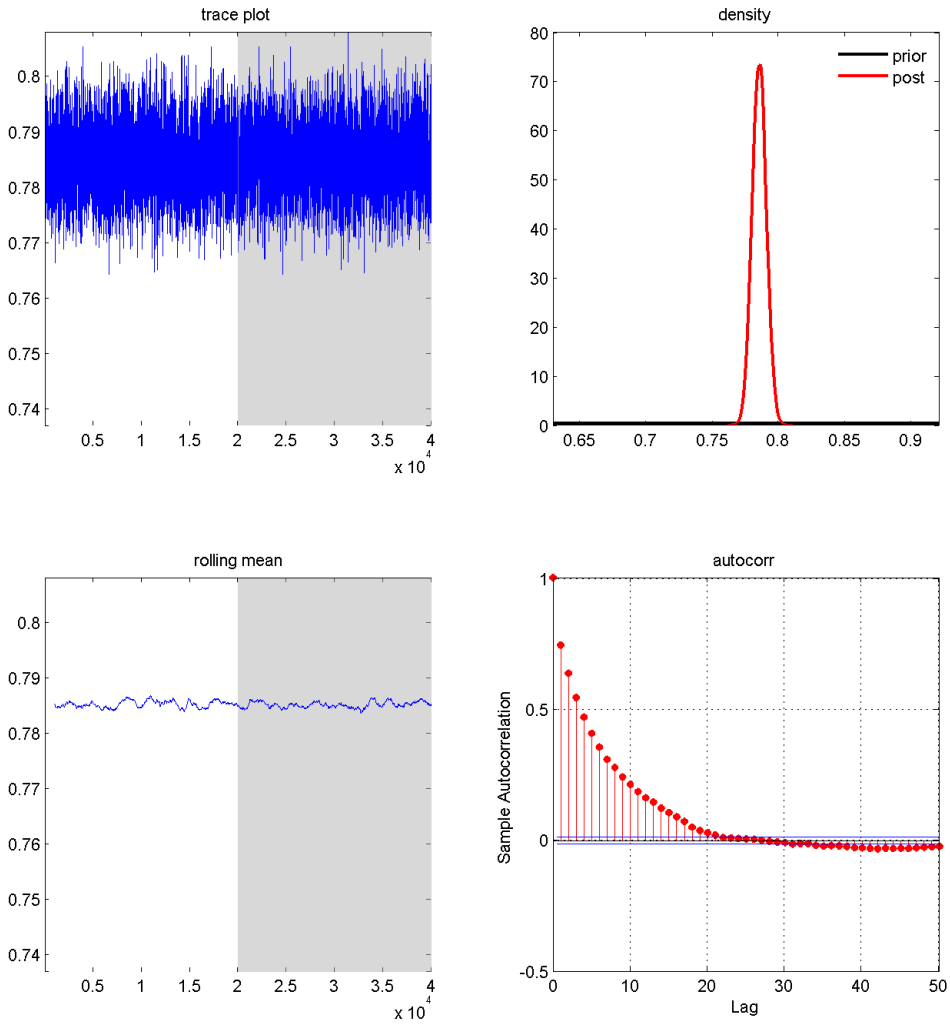
3. *Heteroskedasticity: For $i = 1, \cdots, N$, draw $\sigma_i^{2(s)}$ from an inverse gamma distribution, $p\left(\sigma_i^{2(s)} \left| a^{(s)}, b^{(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A\right.\right)$:*

$$\begin{aligned}\sigma_i^{2(s)} &\sim IG\left(a_i, b_i\right),\\ a_i &= a^{(s)} + T/2,\\ b_i &= b^{(s)} + \frac{1}{2}\sum_{t=1}^{T}\left(y_{it} - \beta^{(s-1)\prime} x_{i,t-1} - \lambda_i^{(s)\prime} w_{i,t-1}\right)^2.\end{aligned}$$

D Simulations and Empirical Application

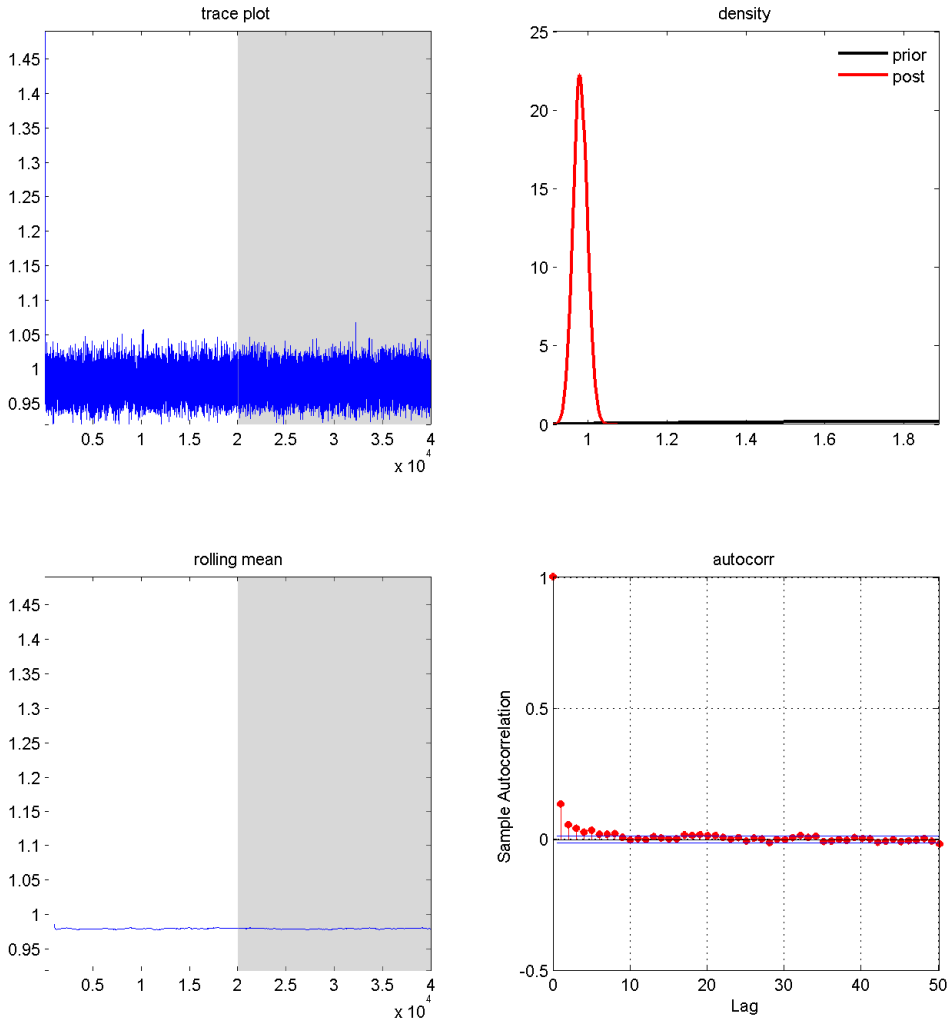
D.1 Simulations

Figure D.1: Convergence Diagnostics: β



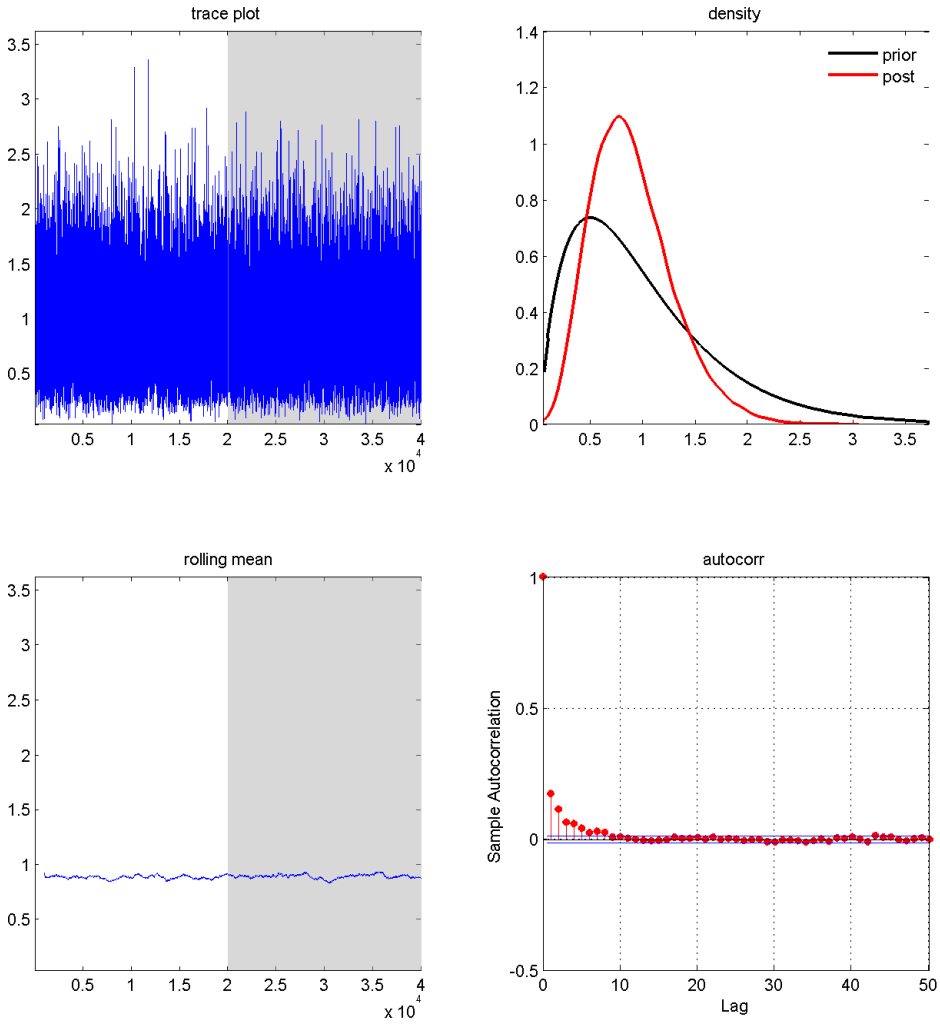
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure D.2: Convergence Diagnostics: σ^2



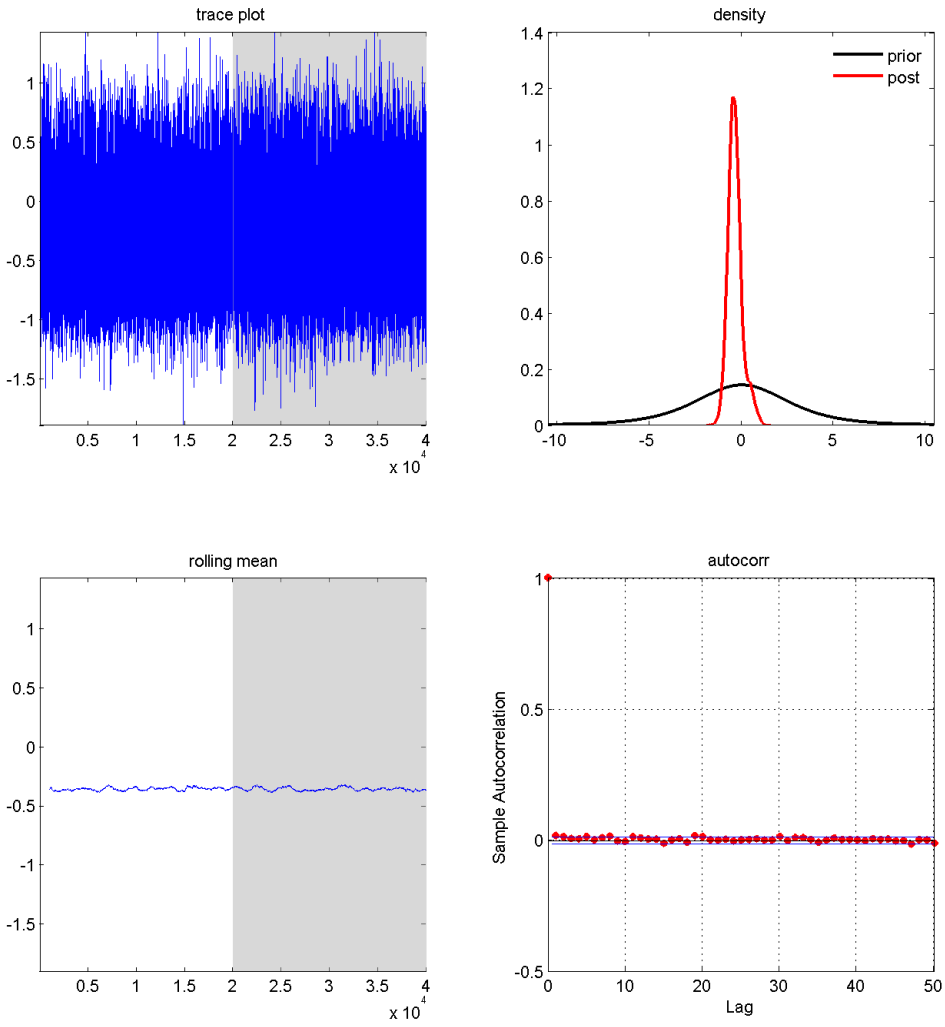
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure D.3: Convergence Diagnostics: α



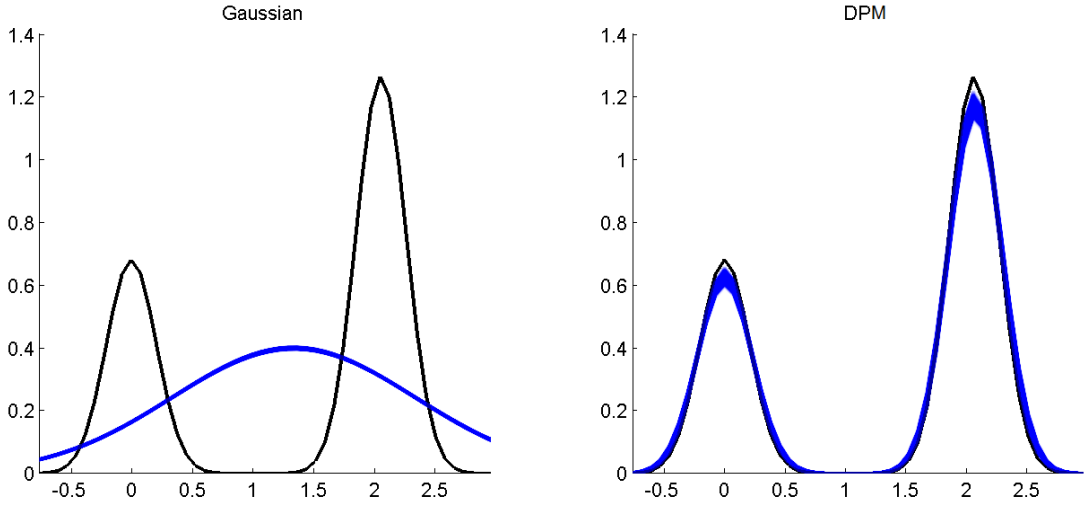
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure D.4: Convergence Diagnostics: λ_1



For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure D.5: f_0 vs $\Pi\left(f\left|y_{1:N,0:T}\right.\right)$: 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, $\Pi\left(f\left|y_{1:N,0:T}\right.\right)$.

D.2 Empirical application

Other model specifications Following the young firm dynamics literature, for the key variables with potential heterogeneous effects ($w_{i,t-1}$), I also examined the following two setups beyond the R&D setup in Section 6:[52]

(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.

Details on sample construction After the second step (based on Assumption 3.5 for unbalanced panels), we have the cross-sectional dimension $N = 859$ for the baseline specification, $N = 794$ with recession, and $N = 677$ with R&D. 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, and we are left with $N = 654$ firms.

Additional results Common parameter β: In most cases, the posterior means are around $0.4 \sim 0.6$.

Point Forecasts: Most of the estimators are comparable according to MSE, with only Flat performing poorly in all three setups.

[52] I do not jointly incorporate recession and R&D because such specification largely restricts the cross-sectional sample size due to the rank requirement for unbalanced panels.

Density Forecasts: The overall best is the Heterosk-NP-C/R predictor in the R&D setup. Comparing setups, the one with recession produces the worst density forecasts (and point forecasts as well), so the recession dummy with heterogeneous effects does not contribute much to forecasting and may even incur overfitting.

Table D.1: 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

Table D.2: 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**

See the description of Table 5.2. Here Heterosk-NP-C/R is the benchmark for both normalization and significance tests.