CAEPR
Working Paper
#2020-003

Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective

Laura Liu
Indiana University

February 16, 2020

This paper can be downloaded without charge from the Social Science Research Network electronic library at https://papers.ssrn.com/sol3/abstract_id=3545886

The Center for Applied Economics and Policy Research resides in the Department of Economics at Indiana University Bloomington. CAEPR can be found on the Internet at: <http://www.indiana.edu/~caepr>. CAEPR can be reached via email at caepr@indiana.edu or via phone at 812-855-4050.

©2020 by Laura Liu. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective*

Laura Liu[†]

February 16, 2020

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 empirical 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 Firm 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, Evan Chan, Benjamin Connault, Hyungsik R. Moon, Alexandre Poirier, 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, Libera Università di Bolzano, University of Michigan, Université de Montréal, Emory University, Tilburg University, Erasmus University Rotterdam, and Tinbergen Institute, 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, First Italian Workshop of Econometrics and Empirical Economics: Panel Data Models and Applications, the 2018 International Association for Applied Econometrics Conference, and the 2019 North American Winter Meeting of the Econometric Society. 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.

[†]Indiana University, lauraliu@iu.edu.

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 for income dynamics (Hirano, 2002; Gu and Koenker, 2017b) and bank balance sheet data for bank stress tests (Liu *et al.*, 2019b). This paper constructs individual-specific density forecasts using a dynamic linear panel data model with common and heterogeneous coefficients 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:

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

where $i = 1, \dots, N$, and $t = 1, \dots, T + 1$. y_{it} is the observed firm performance such as the log of employment, 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 $T + 1$, $y_{i,T+1}$.

The panel considered in this paper features a large cross-sectional dimension N but short time series T . For instance, the number of observations for each young firm is restricted by its age. Good estimates of the unobserved skill λ_i s 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 , even if N goes to infinity.

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 asymmetry, 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 $(\beta, \sigma^2) = (0, 1)$, the firm-specific posterior is characterized by Bayes' theorem,

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

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. Note that this is only an intuitive explanation why the skill distribution f is crucial. In the actual implementation, the inferences of the correlated random effect distribution f , common parameters (β, σ^2) , and firm-specific skill λ_i are all done simultaneously.

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 baseline model in (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 (β, σ^2) 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 common parameters and the distribution of the individual heterogeneity. Then, I prove that estimates of both 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 shocks, (2) an unknown common shock size in cross-sectional homoskedastic cases, (3) unknown individual-specific shock sizes in cross-sectional heteroskedastic cases, (4) strictly exogenous and predetermined variables (including lagged dependent variables) as covariates, and (5) correlated random coefficients addressed by flexible conditional density estimation. Based on the posterior consistency of the estimates, the discrepancy between the proposed density predictor and the oracle is arbitrarily small asymptotically. The oracle predictor is an (infeasible) benchmark defined as the individual-specific posterior predictive distribution, assuming known common parameters and a known distribution of the heterogeneous parameters.

Second, I develop a posterior sampling algorithm specifically addressing nonparametric density estimation of the unobserved individual effects. For a random coefficients model, which is a special case where the individual effects are independent of the conditioning variables, the f part becomes an 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 coefficients 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 improvement in density forecasts relative to alter-

native 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.* (2019b) and Gu and Koenker (2017a,b). Liu *et al.* (2019b) 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 cross-sectional heteroskedasticity where σ_i^2 is an unobserved random quantity, while Liu *et al.* (2019b) incorporate cross-sectional and time-varying heteroskedasticity via a deterministic function of observed conditioning variables.

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 sources 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 coefficients 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 semiparametric 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 could be 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.

Earlier works regarding full Bayesian analyses with parametric priors on λ_i can be found in Lancaster (2002) (orthogonal reparametrization and a flat prior); Chamberlain and Hirano (1999), Chib and Carlin (1999), and Sims (2000) (Gaussian prior); and Chib (2008) (student-t and finite mixture priors). There have also been empirical works on the DPM model with panel data, but they mostly focus on empirical studies rather than theoretical analyses. For example, Hirano (2002) and Jensen *et al.* (2015) use linear panel models with setups 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.

In the frequentist literature, Li and Vuong (1998), Delaigle *et al.* (2008), Evdokimov (2010), and Hu (2017), among others, have studied a 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 models. However, the frequentist approach misses estimation uncertainty, which matters in density forecasts, as mentioned previously.

Second, this paper also relates to the literature on nonparametric Bayesian methods in density estimation problems (Ghosh and Ramamoorthi, 2003; Hjort *et al.*, 2010; Ghosal and van der Vaart, 2017). In particular, for unconditional density estimation, a recent paper by Canale and De Blasi (2017) relaxed the tail conditions to accommodate multivariate location-scale mixtures. For 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 and establish the strong consistency for a multivariate conditional density estimator featuring infinite location-scale mixtures with a probit stick-breaking process. Then, I further take into account the deconvolution and dynamic panel data structure, as well as obtain the convergence of the proposed predictor to the oracle predictor in strong topology.

Last but not least, the empirical application in this paper also links to the young firm dynamics literature. Akcigit and Kerr (2018) document 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. The empirical analysis of this paper builds on these findings. Besides more accurate density forecasts, I 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 specifies the general panel data model, density forecasts, and nonparametric Bayesian priors; Section 3 establishes the posterior consistency of the estimates and the convergence of the density forecasts to the oracle; Section 4 conducts Monte Carlo simulations; Section 5 presents the empirical application to young firm dynamics; and Section 6 concludes. Notations, proofs, algorithms, and additional 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

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

where $i = 1, \dots, N$, and $t = 1, \dots, T + h$. Similar to the baseline setup in (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 notational simplicity, and 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 key sources of individual heterogeneity. If $w_{i,t-1} = 1$, λ_i is reduced to an individual-specific intercept, i.e. firm i 's skill level in the baseline model (1). More generally, $w_{i,t-1}$ can contain individual-specific variables (e.g. firm-specific R&D) and aggregate variables (e.g. a recession dummy). I focus on the former case below for notational simplicity. In the latter case, all theoretical analyses would be further conditioned on the aggregate observations.

$x_{i,t-1}$ is a vector of observed covariates that have homogeneous effects on the outcomes, and β is the corresponding vector of common parameters. I decompose $x_{i,t-1} = [x_{i,t-1}^O, x_{i,t-1}^P]'$, where $x_{i,t-1}^O$ is strictly exogenous and $x_{i,t-1}^P$ is predetermined. One example of $x_{i,t-1}^P$ is the lagged outcome $y_{i,t-1}$ capturing the persistence. Both $x_{i,t-1}^O$ and $x_{i,t-1}^P$ can include other control variables, such as firm characteristics and general economic conditions. Let $x_{i,t-1}^{P*}$ denote the subgroup of $x_{i,t-1}^P$ excluding lagged outcomes, then $x_{i,t-1} = [x_{i,t-1}^O, x_{i,t-1}^{P*}, y_{i,t-1}]'$ with $\beta = [\beta^O, \beta^{P*}, \rho]'$. Here, the distinction between homogeneous effects $\beta' x_{i,t-1}$ and heterogeneous effects $\lambda_i' w_{i,t-1}$ helps model the key latent heterogeneities while avoiding the curse of dimensionality. Combining information from the covariates, the conditioning set at period t is defined as

$$c_{i,t-1} = (x_{i,0:t-1}^P, x_{i,0:T}^O, w_{i,0:T}). \quad (3)$$

We further denote $D = (\{D_i\}_{i=1}^N)$, where $D_i = c_{iT}$, as the data used for estimation and the conditioning set for posterior inference.

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

and the underlying distribution of h_i by f . For instance, $\vartheta = \beta$, $h_i = (\lambda_i, \sigma_i^2)$ under cross-sectional heteroskedasticity. In many empirical applications, such as the young firm example, the size of risk may vary over the cross-section, so cross-sectional heteroskedasticity could contribute to better density forecasts.

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 $c_{i0} = (x_{i0}^P, x_{i,0:T}^O, w_{i,0:T})$. The other is the correlated random coefficients model, where h_i and c_{i0} could be correlated with each other. This paper considers both models while focusing on the latter—although the former is more parsimonious and easier to implement, the latter is more realistic for young firm dynamics as well as many other empirical setups. In practice, it is not necessary to incorporate all initial values of the predetermined variables and the whole series of the strictly exogenous variables into c_{i0} . 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.

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. Both definitions rely on the conditional predictor,

$$f_{i,T+1}^{cond}(y|\vartheta, f, D_i) = \int \underbrace{p(y|h_i, \vartheta, w_{iT}, x_{iT})}_{\text{future shock}} \cdot \underbrace{p(h_i|\vartheta, f, D_i)}_{\text{individual heterogeneity}} dh_i, \quad (4)$$

which provides the density forecasts of $y_{i,T+1}$ conditional on the common parameters ϑ , underlying distribution f , and individual i 's data D_i . 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

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

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 formulate 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 (4),

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

In practice, (ϑ, f) are all unknown and need to be estimated, thus introducing another source of uncertainty. For the common parameters ϑ , I adopt a conjugate prior (e.g. multivariate normal for cross-sectional heteroskedastic cases) in order to stay close to the linear regression framework. For the distribution of individual heterogeneity f , I resort to the nonparametric Bayesian prior (specified in the next subsection) to flexibly model this 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) ,

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

The conditional predictor reflects uncertainties due to future shocks 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) , 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 while being relatively easy to implement. The specific functional form depends on whether f is characterized by a random coefficients model or a correlated random coefficients model.

In cross-sectional heteroskedastic cases, I incorporate another flexible prior on the distribution of σ_i^2 . Define $l_i = \log \frac{\bar{\sigma}^2(\sigma_i^2 - \underline{\sigma}^2)}{\bar{\sigma}^2 - \sigma_i^2}$, where $\underline{\sigma}^2$ ($\bar{\sigma}^2$) is some small (large) positive number. This transformation ensures that the support of $f_0^{\sigma^2}$ is bounded by $[\underline{\sigma}^2, \bar{\sigma}^2]$ and thus satisfies the requirement for the asymptotic convergence, whereas support of l_i is unbounded so a similar prior structures can be applied to both λ_i and l_i . We assume λ_i and σ_i^2 are conditionally independent with respect to each other conditioning on c_{i0} , so their mixture structures are modeled separately and parallelly. For a concise exposition, I define a generic variable z that can represent either λ or l , and include z as a superscript to indicate which part a specific parameter belongs to. When there is no confusion, the superscript z and the subscript i are suppressed.

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.

A typical choice in the nonparametric Bayesian literature is the Dirichlet Process Mixture (DPM) prior. 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 normals,

$$z_i \sim \sum_{k=1}^{\infty} p_k N(\mu_k, \Omega_k). \quad (5)$$

Different draws from the DPM prior are characterized by different combinations of $\{p_k, \mu_k, \Omega_k\}$, which lead to different shapes of f . That is why the DPM prior is flexible enough to approximate a wide range of continuous distributions. The component parameters (μ_k, Ω_k) are drawn from the base distribution G_0 , which is chosen to be a conjugate multivariate-normal-inverse-Wishart distribution, or a normal-inverse-gamma distribution for scalar z_i . The component probability p_k is constructed via a stick-breaking process governed by the scale parameter α .

$$(\mu_k, \Omega_k) \sim G_0, \text{ and } p_k \sim \zeta_k \prod_{j < k} (1 - \zeta_j), \text{ where } \zeta_k \sim \text{Beta}(1, \alpha). \quad (6)$$

The 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. One virtue of the nonparametric Bayesian framework is its ability to flexibly elicit the tuning parameter, such as α , from the data. Namely, we can set up a relatively flexible hyperprior for $\alpha \sim \text{Ga}(a_0^\alpha, b_0^\alpha)$, and update it based on the observations, which “automatically” chooses the complexity of the mixture structure.

2.3.2 Correlated Random Coefficients Model

To accommodate the correlated random coefficients model where the individual heterogeneity $z_i (= \lambda_i \text{ or } l_i)$ can be correlated with the conditioning variables c_{i0} , it is necessary to consider a nonparametric Bayesian prior that is compatible with the much harder conditional density estimation problem. One issue is associated with the uncountable collection of conditional densities, and Pati *et al.* (2013) circumvent it by linking the properties of the conditional 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} ,

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

Similar to the DPM prior, the component parameters can be directly drawn from the base distribution, $(\mu_k, \Omega_k) \sim G_0$. G_0 is again specified as a conjugate matrix-variate-normal-inverse-Wishart form

(or a multivariate-normal-inverse-gamma distribution for scalar z_i). Now the mixture probabilities are characterized by a probit stick-breaking process

$$p_k(c_{i0}) = \Phi(\zeta_k(c_{i0})) \prod_{j < k} (1 - \Phi(\zeta_j(c_{i0}))), \quad (7)$$

where stochastic function ζ_k is drawn from Gaussian process $\zeta_k \sim GP(0, V_k)$ for $k = 1, 2, \dots$. Rodríguez and Dunson (2011) demonstrate the flexibility and computational simplicity of the probit stick-breaking prior.

This setup has three key features: component means are linear in c_{i0} ; component covariances are independent of c_{i0} ; and mixture probabilities are flexible functions of c_{i0} . This framework is relatively parsimonious for finite sample implementation and, at the same time, general enough to accommodate a broad class of conditional distributions. Intuitively, it is similar to approximating the conditional density via Bayes' theorem but does not explicitly model the distribution of the conditioning variables c_{i0} . The infinite mixture structure and flexible mixture probabilities could absorb dependency on c_{i0} , so we would not need any further dependency of component means and covariances on c_{i0} beyond the MGLR_x specification (see details in the Appendix).

3 Theoretical Properties

Generally speaking, it is desirable to ensure that the prior belief does not dominate the posterior inference asymptotically. For 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, the asymptotic properties act as a frequentist justification for the Bayesian method—as the sample size increases, the updated posterior recovers the unknown true data generating process (DGP). Also, 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—the null set for the prior can be topologically large, and hence the true model can easily fall beyond the scope of the prior (Freedman, 1963, 1965). Therefore, it is crucial to find reasonable conditions on the joint behavior of the prior and the true density to establish the posterior consistency argument.

3.1 Identification

Although identification may not be necessary to ensure the convergence of the density forecasts to the oracle predictor, identification is essential to ensure the posterior consistency of the estimates so that the proposed method could be general to problems beyond forecasting, e.g. heterogeneous treatment effect. Here, I present the identification result in terms of the correlated random coeffi-

cients model with cross-sectional heteroskedasticity, where random coefficients and cross-sectional homoskedasticity can be viewed as special cases.

Assumption 1. (Identification: General Model)

1. *Model setup: Consider the panel data model in (2),*

- (a) $(c_{i0}, \lambda_i, \sigma_i^2)$ are i.i.d. across i .
- (b) For all t , conditional on $(y_{it}, c_{i,t-1})$, x_{it}^{P*} is independent of (λ_i, σ_i^2) .
- (c) $(x_{i,0:T}^O, w_{i,0:T})$ are independent of (λ_i, σ_i^2) .
- (d) Conditioning on c_{i0} , λ_i and σ_i^2 are independent of each other.
- (e) Let $u_{it} = \sigma_i v_{it}$. $v_{it} \sim N(0, 1)$ is i.i.d. across i and t and independent of $(c_{i,t-1}, \lambda_i, \sigma_i^2)$.

2. *Identification:*

- (a) The characteristic functions for $\lambda_i|c_{i0}$ and $\sigma_i^2|c_{i0}$ are non-vanishing almost everywhere.
- (b) For all i , $w_{i,0:T-1}$ has full rank d_w almost everywhere.
- (c) After orthogonal forward differencing,

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

Then, the matrix $\mathbb{E}[\sum_{t=1}^{T-d_w} \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}]$ has full rank d_x .

Despite the conditional independence in condition 1-d, λ_i and σ_i^2 can potentially relate to each other through c_{i0} . The setup could be further extended, such as relaxing the conditional independence between λ_i and σ_i^2 and allowing for more general v_{it} distributions (discussed in the Appendix).

Theorem 2. (Identification: General Model) *Under Assumption 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.*

The argument is similar to Arellano and Bover (1995) and Arellano and Bonhomme (2012), except for the treatment of cross-sectional heteroskedasticity. Here σ_i^2 is an unobserved random quantity. First, the identification of common parameters β in panel data models is standard in the literature (Baltagi, 1995; Arellano and Honoré, 2001; Arellano, 2003; Hsiao, 2014). For example, the rank condition helps identify β via orthogonal forward differencing. Second, as λ_i is additively separable from the shocks, I follow the traditional proof based on characteristic functions (i.e. the Fourier transform) to identify f^λ . Finally, 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} .

3.2 Posterior Consistency

Previous nonparametric Bayesian literature focuses on density estimation problems (see Related Literature) without deconvolution and dynamic panel data structures. In this subsection, I first

provide general sufficient conditions that ensure posterior consistency of the estimated common parameters ϑ and the estimated (conditional) distribution of individual effects f in a general class of semiparametric setup, and then I specify and verify these conditions in cases of homoskedastic/heteroskedastic (correlated) random coefficients models.

General Semiparametric Model. Let Θ be the space of the common parameters ϑ , $\mathcal{F}$ be the space of the underlying distribution of individual heterogeneity f , $\Pi(\cdot, \cdot)$ be a joint prior on $\Theta \times \mathcal{F}$ with marginal priors being $\Pi^\vartheta(\cdot)$ and $\Pi^f(\cdot)$, and $\Pi(\cdot, \cdot | D)$ be the corresponding joint posterior. The individual specific likelihood takes a general “convolution” form

$$g(D_i | \vartheta, f) = \begin{cases} \int p(D_i | \vartheta, h_i) f(h_i) dh_i, & \text{if } f \text{ is an unconditional distribution,} \\ \int p(D_i \setminus c_{i0} | \vartheta, h_i) f(h_i | c_{i0}) q_0(c_{i0}) dh_i, & \text{if } f \text{ is a conditional distribution,} \end{cases} \quad (8)$$

where $D_i \setminus c_{i0}$ denotes the set difference, and $q_0(c_{i0})$ is the true marginal density of c_{i0} .

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 weak topology, and convergence of pdfs is stronger than convergence in distribution or weak convergence. Also, strong consistency implies the convergence of quantiles.

When f is a conditional distribution, it is helpful to link the properties of the conditional density to the corresponding joint density without explicitly modeling the marginal density of c_{i0} , which circumvents the difficulty associated with an uncountable set of conditional densities (Pati *et al.*, 2013). 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 the conditioning variables c_0 are integrated out with respect to the true q_0 , so this setup does not require estimating q_0 .

Definition 3. The posterior achieves *strong consistency* at (ϑ_0, f_0) if for all $\epsilon, \delta > 0$, as $N \rightarrow \infty$,

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

in probability with respect to the true DGP.

Theorem 4. (Posterior Consistency: General Semiparametric Model) *The posterior is strongly consistent at (ϑ_0, f_0) under three sets of sufficient conditions:*

1. *Individual-specific likelihood:*

(a) *Kullback-Leibler (KL) property: For all $\epsilon > 0$,*

$$\Pi(\vartheta \in \Theta, f \in \mathcal{F} : d_{KL}(g(D_i | \vartheta_0, f_0), g(D_i | \vartheta, f)) < \epsilon) > 0,$$

where $d_{KL}(g_0, g) = \int g_0 \log \frac{g_0}{g}$ is the KL divergence of g from g_0 .

(b) There exists a $\delta_0 > 0$ such that for all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_0$ and $f \in \mathcal{F}$, $\|g(D_i | \vartheta_1, f) - g(D_i | \vartheta_2, f)\|_1 \leq C_g \|\vartheta_1 - \vartheta_2\|_2$, for some $C_g > 0$ not depending on f .

2. Common parameters: There exists an exponentially consistent sequence of tests $\varphi_N(D)$ for testing $H_0 : \vartheta = \vartheta_0$, against $H_1 : \vartheta \in \Theta^c$, i.e. there exists a constant $C_\varphi > 0$ such that

$$(a) \mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) = O(e^{-C_\varphi N}), \text{ and } (b) \sup_{\vartheta \in \Theta^c, f \in \mathcal{F}} \mathbb{E}_{\vartheta, f} [1 - \varphi_N(D)] = O(e^{-C_\varphi N}),$$

where $\Theta^c \subset \Theta$ and $\vartheta \notin \Theta^c$.

3. Distribution of individual heterogeneity: Its prior satisfies a sieve property, i.e. 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 all $\epsilon > 0$,

(a) For some $\beta > 0$, $\Pi(\mathcal{F}_N^c) = O(\exp(-\beta N))$.

(b) For some $\gamma > 0$, $\sum_j \sqrt{\mathcal{N}(\epsilon, \mathcal{F}_{N,j}) \Pi(\mathcal{F}_{N,j})} = o(\exp((1-\gamma)N\epsilon^2))$, where $\mathcal{N}(\epsilon, \mathcal{F}_{N,j})$ is the covering number of $\mathcal{F}_{N,j}$ by balls with a radius ϵ .

Intuitively, let $\Theta_\delta^c = \{\|\vartheta - \vartheta_0\|_2 \geq \delta\}$, $\mathcal{F}_\epsilon^c = \{\|f - f_0\|_1 \geq \epsilon\}$, and the likelihood ratio $R_N(D, \vartheta, f) = \prod_{i=1}^N \frac{g(D_i | \vartheta, f)}{g(D_i | \vartheta_0, f_0)}$, the posterior probability of the alternative region can be decomposed as

$$\begin{aligned} \Pi(\vartheta \in \Theta_\delta^c \text{ or } f \in \mathcal{F}_\epsilon^c | D) &= \Pi^\vartheta(\vartheta \in \Theta_\delta^c | D) + \Pi(\vartheta \in \Theta_\delta \text{ and } f \in \mathcal{F}_\epsilon^c | D) \\ &= [\mathbb{P}(\vartheta \in \Theta_\delta^c, D) + \mathbb{P}(\vartheta \in \Theta_\delta, f \in \mathcal{F}_\epsilon^c, D)] / \mathbb{P}(D) \\ &= \left[\int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) + \int_{\Theta_\delta \times \mathcal{F}_\epsilon^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] / \int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f), \end{aligned}$$

and we want to show that the whole expression tends to zero as N goes to infinity. First, for the denominator, the KL property (condition 1-a) implies that the prior puts positive weight around neighborhoods of the true DGP, so the likelihood ratio integrated over the whole space is large enough. Second, the exponentially consistent sequence of tests (condition 2) takes an infimum over the alternative region $\Theta_\delta^c \times \mathcal{F}$, so it ensures that the first term in the numerator is arbitrarily small. Third, the sieve property on f (condition 3) ensures that the sieve expands to the entire alternative region 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 (Canale and De Blasi, 2017). To extend the argument to the observed g instead of the unobserved f , we need to further address convolution and common parameters. In terms of convolution, 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 , condition 1-b makes sure that the deviation generated from ϑ is small enough so that it cannot offset the difference in f . Therefore, conditions 1-b and 3 together guarantee that the data are informative enough to differentiate the true distribution from the alternatives, so the second term of the numerator can be arbitrarily small as well.

Note that the estimated individual effects h_i are not consistent because information is accumulated only along the cross-sectional dimension but not along the time dimension. Also, the result only guarantees pointwise convergence in the space of the distributions. For uniform forecasting performance in dynamic panel data models, see Liu *et al.* (2019b) which considers an empirical Bayes setup with a nonparametric kernel estimate of the marginal distribution of data.

Random Coefficients Model. In this case, f is an unconditional distribution. To ensure condition 1 in Theorem 4, we are working with space $\mathcal{F}^{\lambda*} = \left\{ f^\lambda \in \mathcal{F}^\lambda : \mathbb{E}_{f^\lambda} \|\lambda\|_2^2 \leq M^* \right\}$, for some large $M^* > 0$ (see Appendix for details).

Assumption 5. (Covariates)

1. $w_{i,0:T}$ is bounded.
2. There exists a small $m_w > 0$ such that the eigenvalues of $\sum_t w_{i,t-1} w'_{i,t-1}$ are no less than m_w almost everywhere.
3. Let $\tilde{c}_{i0} = \left(x_{i0}^P, x_{i,0:T-1}^O \right) \cdot \int \|\tilde{c}_{i0}\|_2^{2(1+\eta)} p(\tilde{c}_{i0} | w_{i,0:T}) d\tilde{c}_{i0} < \infty$.
4. $\int \left\| x_{i,t-1}^{P*} \right\|_2^{2(1+\eta)} p\left(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2} \right) dx_{i,t-1}^{P*} < \infty$.

For homoskedastic models, $\eta > 1$; for heteroskedastic models, $\eta > 0$.

The moment conditions on $x_{i,t-1}$ ensure that GMM estimates of the common parameters are asymptotically normal, so the exponentially consistent sequence of tests in Theorem 4(2) can be constructed accordingly. Homoskedastic models require higher order moments than heteroskedastic models because in the former case, σ^2 is part of the common parameters and calls for the exponentially consistent sequence of tests on it as well. These moment conditions also prevent a slight difference in β from obscuring the difference in f , and the boundedness of $w_{i,0:T}$ serves the same purpose for the heterogeneous term $\lambda'_i w_{i,t-1}$ —both are essential to Theorem 4(1-b).

Assumption 6. (Distribution of Individual Heterogeneity: Random Coefficients)

1. True distribution f_0^z :
 - (a) $f_0^z(z)$ is a continuous density.
 - (b) For some $0 < M^z < \infty$, $0 < f_0^z(z) \leq M^z$ 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. The base distribution of the DPM prior (G_0^z) follows a multivariate-normal-inverse-Wishart distribution, where the degree of freedom of the inverse Wishart component $\nu_0^z > \max \{2d_z, (2d_z + 1)(d_z - 1)\}$.

First, condition 1 ensures that the true distribution f_0 is well-behaved, and a multivariate-normal-inverse-Wishart G_0 in condition 2 guarantees that the DPM prior is general enough to contain the true distribution, so the KL property on f is established. To obtain the KL property on g (Theorem

4(1-a)), we need to account for convolution and common parameters. Intuitively, convolution with a common distribution would reduce the difference in two distributions, and the difference due to common parameters is relatively small given Assumption 5 on covariates. Second, according to Corollary 1 in Canale and De Blasi (2017), condition 2 further ensures the sieve property (Theorem 4(3)), 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 .

Theorem 7. (Posterior Consistency: Random Coefficients) *Suppose we have:*

1. *Model: Assumption 1 with (1-a) replaced by “ (λ_i, σ_i^2) are independent of c_{i0} and i.i.d. across i ”.*
 2. *Covariates: $(x_{i,0:T}, w_{i,0:T})$ satisfies Assumption 5.*
 3. *Common parameters:*
 - (a) ϑ_0 is in the interior of $\text{supp}(\Pi^\vartheta)$.
 - (b) For homoskedastic models, the domain of σ^2 is bounded by $[\underline{\sigma}^2, \bar{\sigma}^2]$ for some $\underline{\sigma}^2, \bar{\sigma}^2 > 0$.
 4. *Distributions of individual heterogeneity:*
 - (a) f_0^z and Π^z satisfy Assumption 6.
 - (b) For heteroskedastic models, $\text{supp}(f_0^{\sigma^2}) = [\underline{\sigma}^2, \bar{\sigma}^2]$ for some $\underline{\sigma}^2, \bar{\sigma}^2 > 0$.
- Then, the posterior is strongly consistent at (ϑ_0, f_0) .

Correlated Random Coefficients Model. f is now a conditional distribution, so the following discussion is based on the q_0 -induced measure. Similar to random coefficients cases, we are working with space $\mathcal{F}^{\lambda*} = \left\{ f^\lambda \in \mathcal{F}^\lambda : \int \|\lambda\|_2^2 f^\lambda(\lambda|c_0) q_0(c_0) d\lambda dc_0 \leq M^* \right\}$, for some large $M^* > 0$.

Assumption 8. (Conditioning set) *Let $\mathcal{C}$ be the support of the conditioning variables. $\mathcal{C}$ 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 among the conditioning variables, which is stronger than Assumption 5(1,3) for random coefficients models.

Assumption 9. (Distribution of Individual Heterogeneity: Correlated Random Coefficients)

1. *True distribution f_0^z :*
 - (a) $f_0^z(\cdot|\cdot)$ is jointly continuous in (z, c_0) .
 - (b) For some $0 < M^z < \infty$, $0 < f_0^z(z|c_0) \leq M^z$ for all (z, c_0) .
 - (c) $\left| \int \left[\int f_0^z(z|c_0) \log f_0^z(z|c_0) dz \right] q_0(c_0) dc_0 \right| < \infty$.
 - (d) $\int \left[\int f_0^z(z|c_0) \log \frac{f_0^z(z|c_0)}{\varphi_\delta(z|c_0)} dz \right] q_0(c_0) dc_0 < \infty$, where $\varphi_\delta(z|c_0) = \inf_{\|z'-z\|_2 < \delta} f_0^z(z'|c_0)$, for some $\delta > 0$.
 - (e) For some $\eta > 0$, $\int \left[\int \|z\|_2^{2(1+\eta)} f_0^z(z|c_0) dz \right] q_0(c_0) dc_0 < \infty$.
2. *The base distribution of the MGLR_x prior (G_0^z) is characterized by a multivariate normal distribution on $\text{vec}(\mu)$ and an inverse Wishart distribution on Ω , where the degree of freedom of the inverse Wishart component $\nu_0^z > \max\{2d_z, (2d_z + 1)(d_z - 1)\}$.*
3. *Stick-breaking process: The covariance function for Gaussian process can be specified as $V_k^z(c, \tilde{c}) = \tau^z \exp\left(-A_k^z \|c - \tilde{c}\|_2^2\right)$, where $\tau^z > 0$ is a fixed number.*

- (a) The prior for A_k^z has full support on $\mathbb{R}^+$.
- (b) There exist $\beta, \gamma > 0$ and a sequence $\delta_N = O\left(N^{-5/2}(\log N)^2\right)$ such that $\mathbb{P}(A_k^z > \delta_N) \leq \exp\left(-N^{-\beta}k^{(\beta+2)/\gamma} \log k\right)$.
- (c) For the same γ as in condition 3-b, there exists an increasing sequence $r_N \rightarrow \infty$ and $(r_N)^{d_{c_0}} = o\left(N^{1-\gamma}(\log N)^{-(d_{c_0}+1)}\right)$ 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 6. In terms of the stick-breaking process, the variability of $p_k(c_0)$ due to c_0 decreases with component index k according to condition 3-b, so the first several “sticks” would be able to capture a large fraction of the dependence of z on c_0 . Moreover, the tail of A_k^z cannot be too fat according to condition 3-c.

Theorem 10. (Posterior Consistency: Correlated Random Coefficients) *Suppose we have:*

1. *Model: Assumption 1.*
 2. *Covariates: $(x_{i,0:T}, w_{i,0:T})$ satisfy Assumptions 5(2,4) and 8.*
 3. *Common parameters:*
 - (a) ϑ_0 is in the interior of $\text{supp}(\Pi^\vartheta)$.
 - (b) For homoskedastic models, the domain of σ^2 is bounded by $[\underline{\sigma}^2, \bar{\sigma}^2]$ for some $\underline{\sigma}^2, \bar{\sigma}^2 > 0$.
 4. *Distributions of individual heterogeneity:*
 - (a) f_0^z and Π^z satisfy Assumption 9.
 - (b) For heteroskedastic models, $\text{supp}(f_0^{\sigma^2}) = [\underline{\sigma}^2, \bar{\sigma}^2]$ for some $\underline{\sigma}^2, \bar{\sigma}^2 > 0$.
- Then, the posterior is strongly consistent at (ϑ_0, f_0) .*

3.3 Density forecasts

Based on posterior consistency, we can bound the discrepancy between the proposed predictor and the oracle by estimation uncertainties in ϑ and f , and then show the asymptotical convergence of the density forecasts to the oracle forecast. Define $p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T})$, the individual-specific likelihood of $y_{i,1:T}$; and define $p(y_{i,T+1} | h_i, \vartheta, D_i)$, a component of the density forecast, which captures individual i 's uncertainty due to future shocks and is a more general version of the first term on the right hand side of (4).

Theorem 11. (Density Forecasts: General Semiparametric Model) *Suppose we have:*

1. *Posterior consistency: conditions in Theorem 4.*
2. *True DGP:*
 - (a) f_0 is bounded above by some $M > 0$.
 - (b) The second moment of h_i exists and is positive.
 - (c) If f is a conditional distribution, there exists $\underline{q} > 0$ such that $|q_0(c_0)| \geq \underline{q}$ for all $c_0 \in \mathcal{C}$.

3. *Likelihood:*

(a) $p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})$ is continuous in h_i .

(b) $0 < p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T}) \leq M(D_i)$, for some $M(D_i) > 0$.

4. There exists $\delta_1 > 0$ such that for all $\|\vartheta - \vartheta_0\|_2 < \delta_1$, for some $C_h(D_i), C_y(D_i) > 0$,

(a) $\int |p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})| dh_i \leq C_h(D_i) \|\vartheta - \vartheta_0\|_2$,

(b) $\|p(y_{i,T+1} | h_i, \vartheta, D_i) - p(y_{i,T+1} | h_i, \vartheta_0, D_i)\|_1 \leq C_y(D_i) \|\vartheta - \vartheta_0\|_2$.

Then, density forecasts converge to the oracle, i.e. given i , for all $\epsilon > 0$, as $N \rightarrow \infty$,

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

in probability with respect to the true DGP.

Corollary 12. (Density Forecasts: (Correlated) Random Coefficients) *Given conditions in Theorem 7 for random coefficients models (or conditions in Theorems 10 and 11(2-c) for correlated random coefficients models), density forecasts converge to the oracle.*

The asymptotic convergence of aggregate-level density forecasts can then be derived by summing individual-specific forecasts over different subcategories.

4 Monte Carlo Simulation

This section conducts two sets of Monte Carlo simulation experiments: the baseline setup with random effects, and the general setup with correlated random coefficients and cross-sectional heteroskedasticity. The main text focuses on density forecast results, whereas point forecast results are deferred to the Appendix.

4.1 Forecast Evaluation and Alternative Predictors

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

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

where $y_{i,T+1}$ is the realization at $T + 1$, and $\hat{p}(y_{i,T+1} | D)$ represents the predictive likelihood with respect to the estimated model conditional on the observed data D . $\exp(LPS_A - LPS_B)$ gives the odds of future realizations based on predictor A versus predictor B. I perform a test combining Amisano and Giacomini (2007) (for LPS) and Timmermann and Zhu (2019) (for panel data) to examine the significance in the LPS difference.

Different predictors can be interpreted as different priors on the distribution of λ_i . As these priors are distributions over distributions, Figure 1 plots two draws from each prior. The homogeneous prior (Homog) implies an extreme kind of pooling, which assumes that all firms share the same

Figure 1: Alternative Predictors

Notes: For easier illustration, here I consider the baseline model with univariate λ_i and homoskedasticity. The black solid line and teal dotted line represent two draws from each prior (except NP-disc, where the teal one is also solid). Homog: Because λ^* is unknown beforehand, the subgraph plots two vertical lines representing two degenerate distributions with different locations. Param: The subgraph contains two curves with different means and variances. NP-disc: See Appendix for a formal definition of DP and how it relates to DPM.

level of skill λ^* . It can be viewed as a Bayesian counterpart of the pooled OLS estimator. More rigorously, this prior is defined as $\lambda_i \sim \delta_{\lambda^*}$, where δ_{λ^*} is the Dirac delta function representing a degenerate distribution. The unknown λ^* becomes another common parameter, similar to β , so I adopt a multivariate-normal-inverse-gamma prior on $([\beta, \lambda^*]', \sigma^2)$.

The flat prior (Flat) is specified as $f(\lambda_i) \propto 1$, an uninformative prior with the posterior mode being the MLE estimate. Roughly speaking, given the common parameters, there is no pooling from the cross-section, so we learn firm i 's skill λ_i only from its own history.

The parametric prior (Param) pools the information from the cross-section via a parametric distribution, such as a Gaussian distribution with unknown mean and variance, $\lambda_i \sim N(\mu, \omega^2)$. A normal-inverse-gamma hyperprior is further imposed on (μ, ω^2) . The parametric prior can be thought of as a limit case of the DPM prior when the scale parameter $\alpha \rightarrow 0$, so there is only one component, and (μ, ω^2) 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 parametric model with the DPM model such that “the predictive (or marginal) distribution of a single observation is identical under the two models”.

The nonparametric discrete prior (NP-disc) is modeled by a DP where λ_i follows a flexible

Table 1: Simulation Setup: Baseline Model with Random Effects

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, u_{it} \sim N(0, \sigma^2)$
Common parameters	$\beta_0 = 0.8, \sigma_0^2 = \frac{1}{4}$
Initial conditions	$y_{i0} \sim N(0, 1)$
Sample size	$N = 1000, T = 6$
Random Effects:	
Degenerate	$\lambda_i = 0$
Skewed	$\lambda_i \sim \frac{1}{9}N(2, \frac{1}{2}) + \frac{8}{9}N(-\frac{1}{4}, \frac{1}{2}), \text{ so } \mathbb{V}(\lambda_i) = 1$
Bimodal	$\lambda_i \sim (0.35N(0, 1) + 0.65N(10, 1)) / \sqrt{1 + 10^2 \cdot 0.35 \cdot 0.65}, \text{ so } \mathbb{V}(\lambda_i) = 1$

nonparametric distribution, but on a discrete support. This paper focuses on continuous f , which may be more sensible for the skills 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.

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

The semiparametric 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 semiparametric ones. Therefore, it is worthwhile to implement the Monte Carlo experiments and assess which predictor produces more accurate forecasts under which circumstances.

4.2 Baseline Model with Random Effects

The specifications are summarized in Table 1. β_0 is set to 0.8, as economic data usually exhibit some degree of persistence. The initial condition y_{i0} is drawn from a standard normal distribution, which complies with the moment condition in Assumption 5. Choices of $N = 1000$ and $T = 6$ are comparable with the young firm application. There are three experiments with different true distributions of λ_i , $f_0(\cdot)$. The first experiment features a degenerate λ_i distribution, where all firms enjoy the same skill level. Note that it does not satisfy Assumption 6(1-a) requiring the true λ_i distribution to be continuous, and thus serves as a robustness check against the misspecification that the true λ_i distribution is out of the prior support. The second experiment is based on a skewed distribution, a more realistic scenario in empirical studies. The last experiment portrays a bimodal distribution with asymmetric weights on the two components. Various robustness checks are discussed in the Appendix.

I simulate 1,000 panel datasets in each setup. 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

Table 2: Density Forecast Evaluation: Baseline Model with Random Effects

	Degenerate	Skewed	Bimodal
<i>Oracle</i>	<i>-725</i>	<i>-798</i>	<i>-766</i>
Homog	-0.2***	-193***	-424***
Flat	-102***	-7***	-38***
Param	-4	-1***	-34***
NP-disc	-0.2***	-206***	-40***
NP-R	-4	-0.3	-6

Notes: The density forecasts are assessed by LPS and a test combining Amisano and Giacomini (2007) and Timmermann and Zhu (2019). For the oracle predictor, the table reports the exact values of $LPS \cdot N$ (averaged over 1,000 Monte Carlo samples). For other predictors, the table reports their differences from the oracle. The tests compare other feasible predictors with NP-R, with significance levels indicated by *: 10%, **: 5%, and ***: 1%. The entries in bold indicate the best feasible predictor in each column.

tests, the MCMC draws appear to converge to a stationary distribution (see Appendix).

Table 2 shows the forecasting comparison among alternative predictors. When the λ_i distribution is degenerate, Homog and NP-disc are the best, as expected. They are closely followed by NP-R and Param, and Flat is considerably worse. When the λ_i distribution is non-degenerate, there is a substantial gain from employing the NP-R predictor. In the bimodal case, the NP-R predictor far exceeds all alternatives. In the skewed case, 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—even the nonparametric DP prior with countably infinite support is far from enough.

To investigate why we obtain better forecasts, Figure 2 demonstrates the posterior distribution of the λ_i distribution for experiments Skewed and Bimodal. In the skewed case, the NP-R estimator better tracks the peak on the left and the tail on the right. In the bimodal case, the NP-R estimator nicely captures the M-shape. Therefore, 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 reflects the connection between distribution estimation (of individual heterogeneity) and density forecasts as in Theorem 12.

4.3 General Model

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

Figure 2: f_0 vs $\Pi(f | y_{1:N,0:T})$: Baseline Model with Random Effects

Notes: The subgraphs are constructed from the estimation results of one of the 1,000 repetitions. The black solid lines represent the true λ_i distributions, f_0 . The teal bands show the posterior distributions of f , $\Pi(f | y_{1:N,0:T})$.

because this conditional distribution exhibits rapid changes in its shape, which considerably restricts the local sample size. Their original conditional distribution is one-dimensional, and I expand it to accommodate the three-dimensional λ_i via a linear transformation. In terms of cross-sectional heteroskedasticity, I also let the σ_i^2 interact with the initial conditions, and the functional form is modified from Pelenis (2014) Case (ii). The modification guarantees that the σ_i^2 distribution is continuous with a bounded support above zero (see Theorems 10 and 12), and that the average signal-to-noise ratio is not far from 1. In addition, I consider the distribution of the innovations v_{it} to be either normal or skewed. The latter leaves the normal likelihood function misspecified. For identification, the v_{it} distributions are normalized such that $\mathbb{E}(v_{it}) = 0$ and $\mathbb{V}(v_{it}) = 1$.

The left two columns of Table 4 describe the prior setups of f^λ and f^{σ^2} . Due to cross-sectional heteroskedasticity and correlated random coefficients, the prior structures become more complicated. I further add the Homosk-NP-C predictor to examine whether it is practically relevant to model heteroskedasticity. The third column of Table 4 assesses the forecasting performance under correct specification. Heterosk-NP-C is the most accurate density predictor. Several messages can be inferred from a more detailed comparison among predictors. First, based on the comparison between Heterosk-NP-C and Homog/Homosk-NP-C, it is important to account for individual effects in both coefficients λ_i and shock size σ_i^2 . 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.

Under a misspecified v_{it} distribution, the oracle knows the true distribution of v_{it} and still serves as a legitimate benchmark for forecast evaluation, but the posterior consistency would not hold and the density forecasts would not converge to the oracle in general. Nevertheless, the proposed

Table 3: Simulation Setup: General Model

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it}, u_{it} = \sigma_i v_{it}$
Covariates	$w_{i,t-1} = [1, w_{t-1}^{(2)}, w_{i,t-1}^{(3)}]', w_{t-1}^{(2)} \sim N(0, 1) \mathbf{1}(w_{t-1}^{(2)} \leq 10),$ $w_{i,t-1}^{(3)} \sim \text{Ga}(1, 1) \mathbf{1}(w_{i,t-1}^{(3)} \leq 10)$
Common parameters	$\beta_0 = 0.8$
Initial conditions	$y_{i0} \sim U(0, 1)$
Corr. random coef.	$\lambda_i y_{i0} \sim e^{-2y_{i0}} N(y_{i0}v, 0.1^2 vv') + (1 - e^{-2y_{i0}}) N(y_{i0}^4 v, 0.2^2 vv'),$ $v = [1, 2, -1]'$
Cross-sec. heterosk.	$\sigma_i^2 y_{i0} \sim [0.454(y_{i0} + 0.5)^2 \cdot \text{IG}(51, 40) + 10^{-6}] \cdot \mathbf{1}(\sigma_i^2 \leq 10^6)$
Sample size	$N = 1000, T = 6$
Innovation distributions:	
Normal	$v_{it} \sim N(0, 1)$
Skewed	$v_{it} \sim \frac{1}{9}N(2, \frac{1}{2}) + \frac{8}{9}N(-\frac{1}{4}, \frac{1}{2})$

Notes: For the left two panels, λ_{i1} is the coefficient on $w_{i,t-1}^{(1)} = 1$ and can be interpreted as the heterogeneous intercept. In the second and the fourth distribution panels, the black solid lines, the teal dashed lines, and orange dotted lines are conditional on $y_{i0} = 0.25, 0.5$, and 0.75 , respectively. As $y_{i0} \sim U(0, 1)$, the conditional distribution equals the joint distribution for all $y_{i0} \in [0, 1]$, i.e. $f(\lambda_{i1} | y_{i0}) = f(\lambda_{i1}, y_{i0}) q_0(y_{i0}) = f(\lambda_{i1}, y_{i0})$.

semiparametric method could still be advantageous to density forecasts due to its flexibility—in the last column of Table 4, the relative ranking is the same as the correctly specified case, and NP-C is still significantly better than the alternatives.

5 Empirical Application: Young Firm Dynamics

Studies have documented that young firm performance is affected by R&D and that different firms may react differently (Akcigit and Kerr, 2018; Robb and Seamans, 2014). In this empirical application, I examine this type of firm-specific latent heterogeneity from a density forecasting perspective. I use the confidential data from the Kauffman Firm Survey (KFS), which offers a large panel of startups (4,928 firms founded in 2004, nationally representative sample), a reasonable time span (2004-2011, one baseline survey and seven follow-up annual surveys), and detailed information on

Table 4: Prior Structures and Density Forecast Evaluation: General Model

		f^λ	f^{σ^2} (or f^l)	Normal v_{it}	Skewed v_{it}
<i>Oracle</i>		<i>Known</i>	<i>Known</i>	-974	-965
Homog		$= \delta_{\lambda^*}$	$f^{\sigma^2} = \delta_{\sigma^2^*}$	-407***	-417***
Homosk	NP-C	$\sim \text{MGLR}_x$	$f^{\sigma^2} = \delta_{\sigma^2^*}$	-134***	-146***
Heterosk	Flat	$\propto 1$	$f^{\sigma^2} \propto 1$	-384***	-366***
	Param	$= \text{Normal}$	$f^{\sigma^2} = \text{IG}$	-79***	-78***
	NP-disc	$\sim \text{DP}$	$f^l \sim \text{DP}$	-79***	-78***
	NP-R	$\sim \text{DPM}$	$f^l \sim \text{DPM}$	-229***	-224***
	NP-C	$\sim \text{MGLR}_x$	$f^l \sim \text{MGLR}_x$	-70	-71

Notes: The prior structure of Heterosk-Param is detailed in the Appendix. For density forecast evaluation, see the description in Table 2. Here the tests are conducted with respect to Heterosk-NP-C.

young firms. See Robb *et al.* (2009) for further description of the survey design.

5.1 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 (2018) 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. Below, I focus on the following model specification,

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

where $\text{R\&D}_{it}$ is given by the ratio of a firm's R&D employment over its total employment. Other setups are discussed in the Appendix.

The panel used for estimation spans from 2004 ($t = 0$) to 2010 ($t = T$) with time dimension $T = 6$. The data for 2011 ($t = T + 1$) are reserved for pseudo out-of-sample forecast evaluation. The sample is constructed as follows. First, for any (i, t) , if firm i 's R&D employment is greater than its total employment, there is an incompatibility issue, and the corresponding $\text{R\&D}_{it}$ is set to NA, which only affects 0.68% of the observations. Then, I only keep firms with long enough observations for identification in unbalanced panels. This results in a cross-sectional dimension $N = 503$. The proportion of missing values is $(\# \text{missing obs}) / (NT) = 9.32\%$. Here I consider unbalanced panels with randomly omitted observations (see Appendix), which helps incorporate more individuals into estimation and elicits more information for prediction. A further extension to a panel Tobit model as in Liu *et al.* (2019a) could help accommodate firms' endogenous exit choice. The descriptive statistics for $\log \text{emp}_{it}$ and $\text{R\&D}_{it}$ are summarized in Table 5, and the corresponding densities are

Table 5: Descriptive Statistics of Observable

	10%	Mean	Med.	90%	SD	Skew.	Kurt.
log emp	0.69	1.59	1.39	2.20	1.02	0.59	3.42
R&D	0.00	0.27	0.14	0.50	0.32	1.18	3.25

plotted in Figure 12 in the Appendix. Both distributions are right skewed and may have more than one peak, so we expect that the proposed predictors with nonparametric priors could perform well in this example.

5.2 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 f^λ and a DPM prior on f^l . The conditioning variable y_{i0} is further standardized, which ensures numerical stability in practice as the conditioning variables enter exponentially into the covariance function of the Gaussian process.

The first two columns in Table 6 characterize the posterior estimates of the common parameter β . In most of the cases, the posterior means are mostly around $0.5 \sim 0.6$, which suggests that the young firm performance exhibits some degree of persistence, but not remarkably strong. 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 DGP features correlated random coefficients, the random 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.

The last column in Table 6 compares density forecasting performance. The overall best is the Heterosk-NP-C/R predictor. The main message is similar to the Monte Carlo of the general model—it is crucial to account for individual effects in both coefficients λ_i and shock size σ_i^2 through a flexible nonparametric prior that acknowledges continuity and correlated random coefficients when the underlying individual heterogeneity possesses these features. Intuitively, the odds, given by the exponential of the difference in LPS, indicate that Heterosk-NP-C/R produces density forecasts 32% (31%) more likely than Homog (Heterosk-Flat) does, on average.

Figures 3 and 13 (in the Appendix) provide 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 of how well the density forecasts coincide with the true (unobserved) conditional forecasting distributions given the current information set. Under the null hypothesis that the density forecasts coincide with the true DGP, the PITs are i.i.d. $U(0, 1)$ and

Table 6: Parameter Estimation and Density Forecast Evaluation: Young Firm Dynamics

		β		LPS*N
		Mean	SD	
<i>Heterosk</i>	<i>NP-C/R</i>	<i>0.50</i>	<i>0.02</i>	-195
Homog		0.88	0.02	-139***
Homosk	NP-C	0.48	0.02	-113***
Heterosk	Flat	0.19	0.07	-134***
	Param	0.62	0.07	-63***
	NP-disc	0.92	0.01	-88***
	NP-R	0.74	0.04	-20**
	NP-C	0.53	0.03	-6*

Notes: See the description of Table 2 for density forecast evaluation. Here Heterosk-NP-C/R is the benchmark for both normalization and significance tests. For Heterosk-NP-C/R, the table reports the exact values of $LPS \cdot N$. For other predictors, the table reports their differences from Heterosk-NP-C/R.

Figure 3: PIT

Notes: Teal lines indicate the confidence interval. See Appendix for PITs of all predictors.

the histogram is close to a flat line (Diebold *et al.*, 1998; Amisano and Geweke, 2017). We can see that, in NP-C/R, NP-C, and Flat, the histogram bars are mostly within the confidence band, while other predictors yield apparent inverse-U shapes. The reason might be that the other predictors do not take correlated random coefficients into account but instead attribute the subtlety of correlated random coefficients to the estimated variance, which leads to more diffused predictive distributions.

Figure 4 shows four types of firm-level predictive distributions: NP-C/R (a) is more concentrated, (b) is more dispersed, (c) is more skewed, or (d) exhibits extra kurtosis, compared with Homog's Gaussian predictive distributions. Figure 14 in the Appendix regroups these predictive distributions by predictors. For the Homog predictor, all predictive distributions share the same Gaussian shape paralleling with each other. On the contrary, for the NP-C/R predictor, the predictive distributions are fairly different in their shapes.

Figures 5 and 15 (in the Appendix) further aggregate the predictive distributions over sectors. It plots the predictive distributions of the log of the average employment within each sector. Com-

Figure 4: Predictive Distributions: Firm-level, 4 Types

Notes: The black solid (teal dotted) lines are the predictive distributions via the NP-C/R (Homog).

Figure 5: Predictive Distributions: Aggregated by Sectors

Notes: The black solid (teal dotted) lines are the predictive distributions via the NP-C/R (Homog). See Appendix for predictive distributions of all sectors.

paring Homog and NP-C/R across sectors, we can see 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 fact 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, 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, Homog introduces an upward bias into the forecasts, while NP-C/R reduces this bias by flexibly estimating the underlying distribution of firm-specific heterogeneity.

The latent heterogeneity structure is presented in Figure 6, which plots the joint distributions of the estimated individual effects and the conditional variable. For example, the pairwise relationship between λ_{i1} and standardized y_{i0} is nonlinear and exhibit multiple components, which reassures the utilization of nonparametric prior with correlated random coefficients. I also depict the joint

Figure 6: Joint Distributions: $\hat{\lambda}_i$ and y_{i0}

Notes: λ_{i1} is the heterogeneous intercept, and λ_{i2} is the heterogeneous coefficient on R&D.

distribution of $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ as well as the joint distribution of $\hat{\sigma}_i^2$ and standardized y_{i0} in the Appendix. There does not seem to be much correlation between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ and between $\hat{\sigma}_i^2$ and y_{i0} (the latter is in line with the forecasting performance ranking where NP-C/R provides better density forecasts than NP-C does), which, together with sanity checks on (un)conditional correlation as well as a robustness check on density forecast performance (see Appendix), partially supports the assumption that conditioning on y_{i0} , λ_i and σ_i^2 would be independent in this young firm sample.

6 Conclusion

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. The full Bayesian procedure helps capture all sources of uncertainties and, together with the flexibility in the nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients, leads to more accurate density forecasts. The proposed method is theoretically appealing as the paper proves the posterior consistency of the estimates and the convergence of the density forecasts to the oracle. The proposed method is also practically useful as demonstrated in the Monte Carlo simulations and an empirical application to young firm dynamics.

References

- AKCIGIT, U. and KERR, W. R. (2018). Growth through heterogeneous innovations. *Journal of Political Economy*, **126** (4), 1374–1443.
- AMEWOU-ATISSO, M., GHOSAL, S., GHOSH, J. K. and RAMAMOORTHY, 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.
- , BLUNDELL, R. and BONHOMME, S. (2017). Earnings and consumption dynamics: a nonlinear panel data framework. *Econometrica*, **85** (3), 693–734.
- 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.
- 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.
- 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).
- DOSS, H. and SELLKE, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *The Annals of Statistics*, pp. 1302–1305.
- 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.
- FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. 1. New York: Wiley, 3rd edn.
- 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., RAMAMOORTHY, 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 RAMAMOORTHY, 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.
- 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.
- 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. (2019a). Forecasting with a panel tobit model. *NBER Working Papers* 26569.
- , — and — (2019b). Forecasting with dynamic panel data models. *Econometrica*, forthcoming.
- 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.
- MASTEN, M. A. (2018). Random coefficients on endogenous variables in simultaneous equations models. *The Review of Economic Studies*, **85** (2), 1193–1250.
- 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.
- 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. *SSRN 1392292*.
- and SEAMANS, R. (2014). The role of R&D in entrepreneurial finance and performance. In *Finance and Strategy*, Emerald Group Publishing Limited, pp. 341–373.
- 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.
- RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian analysis*, **6** (1), 145–178.
- ROSSI, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press.
- SCUTARI, M. (2009). Learning Bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*.

- SIMS, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of econometrics*, **95** (2), 443–462.
- TIMMERMANN, A. and ZHU, Y. (2019). Comparing forecasting performance with panel data. *SSRN* 3380755.
- 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.
- WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pp. 817–838.
- 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.

Supplementary Appendix to “Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective”

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, $U(0, 1)$.

$N(\mu, \sigma^2)$ stands for a **Gaussian/normal distribution** with mean μ and variance σ^2 . Its probability distribution function (pdf) is given by $\phi(x; \mu, \sigma^2)$. 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 $\Phi(x; \mu, \sigma^2)$ 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 $\text{Ga}(a, b)$ with pdf being $f_{\text{Ga}}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$. The according **inverse gamma distribution** is given by $\text{IG}(a, b)$ with pdf being $f_{\text{IG}}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x}$. 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 $\text{IW}(\Psi, \nu)$, then its pdf is $f_{\text{IW}}(\Omega; \Psi, \nu) = \frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu d}{2}} \Gamma_d(\frac{\nu}{2})} |\Omega|^{-\frac{\nu+d+1}{2}} e^{-\frac{1}{2} \text{tr}(\Psi \Omega^{-1})}$. When Ω is a scalar, the inverse Wishart distribution is reduced to an inverse gamma distribution with $a = \nu/2$, $b = \Psi/2$.

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 all finite set of $\{c_1, c_2, \dots, c_n\}$, $[\zeta(c_1), \zeta(c_2), \dots, \zeta(c_n)]'$ has a joint Gaussian distribution with the mean vector being $[m(c_1), m(c_2), \dots, m(c_n)]'$ and the i, j -th entry of the covariance matrix being $V(c_i, c_j)$, $i, j = 1, \dots, N$.

$\mathbf{1}(\cdot)$ is an **indicator function** that equals 1 if the condition in the parenthesis is satisfied and equals 0 otherwise.

I_N is an $N \times N$ **identity matrix**.

In the **panel data** setup, for a generic variable z , which can be v , w , x , or y , z_{it} is a $d_z \times 1$ vector, and $z_{i,t_1:t_2} = (z_{it_1}, \dots, z_{it_2})$ 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 $z = [z_1, z_2, \dots, z_n]'$ is given by $\|z\|_2 = \sqrt{z_1^2 + \dots + z_n^2}$, and the L_1 -**norm** for an integrable function is given by $\|f\|_1 = \int |f(x)| dx$.

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

$\text{tr}(\cdot)$ gives the **trace** of a matrix, $\det(\cdot)$ represents the **determinant** of a matrix, $\text{vec}(\cdot)$ denotes **matrix vectorization**, and $\otimes$ is the **Kronecker product**.

B Model and Theory

B.1 Model

Short T . Which T can be considered small depends on the dimension of individual heterogeneity, 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 DGPs. 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.

Dirichlet Process (DP). The DP is another candidate as a nonparametric prior, which casts a distribution over a set of discrete distributions and constitutes a key building block of the DPM. A DP has two parameters: the base distribution G_0 characterizing the center of the DP, and the scale parameter α representing the precision (inverse-variance) of the DP. Denote

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

if for all partition $(A_1, \dots, A_K)$,

$$(G(A_1), \dots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_K)).$$

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

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

which is a multivariate generalization of the Beta distribution.

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

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

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 . Roughly speaking, the DP scale parameter α is linked to the number of unique components in the mixture density and thus

determines and reflects the flexibility of the mixture density. Let K^* denote the number of unique components. As derived in Antoniak (1974), we have

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

By definition, a draw from DP is a discrete distribution. In this sense, 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 $z_i (= \lambda_i \text{ or } l_i)$ follows a continuous parametric distribution $f(z; \theta)$ where θ are the parameters, and adopt a DP prior for the distribution of θ . 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.

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,

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

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

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

where $\tilde{\mu}_k$ is a $(d_z + d_{c_0}) \times 1$ vector, and $\tilde{\Omega}_k$ is a $(d_z + d_{c_0}) \times (d_z + d_{c_0})$ covariance matrix.

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

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

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

where $\mu_k = \left[\tilde{\mu}_{k,z} - \tilde{\Omega}_{k,zc_0} \tilde{\Omega}_{k,c_0c_0}^{-1} \tilde{\mu}_{k,c_0} \right]$, $\Omega_k = \tilde{\Omega}_{k,zz} - \tilde{\Omega}_{k,zc_0} \tilde{\Omega}_{k,c_0c_0}^{-1} \tilde{\Omega}_{k,c_0z}$. Combining all the steps

above, the conditional distribution can be approximated as

$$\begin{aligned} f(z|c_0) &\approx \sum_{k=1}^{\infty} \frac{\tilde{p}_k \phi(c_0; \tilde{\mu}_{k,c_0}, \tilde{\Omega}_{k,c_0 c_0}) \phi(z; \mu_k [1, c'_0]', \Omega_k)}{f(c_0)} \\ &= \sum_{k=1}^{\infty} p_k(c_0) \phi(z; \mu_k [1, c'_0]', \Omega_k). \end{aligned}$$

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

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 somewhat relaxed assumptions on it.

Extension: Unbalanced Panels. The discussion 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_{it} s are cross-sectionally independent, so the theoretical argument and numerical implementation are still valid in a like manner. Let $\mathcal{T}_i = \{s_{i1}, s_{i2}, \dots, s_{iT_i}\}$ denote the set of T_i periods when individual i has complete observations. That is, $(y_{it}, w_{i,t-1}, x_{i,t-1})$ are observed for all $t \in \mathcal{T}_i$. Note that:

(1) The sample is restricted to individuals with $T+1 \in \mathcal{T}_i$ (i.e. $s_{iT_i} = T+1$), so the individual forecasts could be evaluated by the pseudo-out-of-sample outcomes $y_{i,T+1}$ s. This restriction could be relaxed if one would like to estimate the model using a larger sample but only evaluate the forecasting performance on a subset of the individuals with existent $y_{i,T+1}$ s.

(2) It is also required that the conditioning variables c_{i0} exist for all individuals (in practice, 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). This assumption could also be relaxed depending on the model setup. For example, in the baseline model, it may sometimes be reasonable to assume that $(y_{i,s_{i1}-1}, \lambda_i)$ are i.i.d. across i .

(3) This structure is able to accommodate balanced panels by setting $\mathcal{T}_i = \{1, \dots, T+1\}$.

Then, we can discard the unobserved periods and redefine the conditioning set at time $t = s_{i\tau}$, $\tau = 1, \dots, T_i$, to be

$$c_{i,t-1} = (c_{i0}, x_{i,\mathcal{S}_{i\tau}-1}^P, x_{i,\mathcal{T}_i-1}^O, w_{i,\mathcal{T}_i-1}), \quad (9)$$

where $\mathcal{T}_i - 1$ indicates the set of time periods $\{s_{i1} - 1, s_{i2} - 1, \dots, s_{iT_i} - 1\}$, and $\mathcal{S}_{i\tau} - 1$ denotes the set of time periods $\{s_{i1} - 1, s_{i2} - 1, \dots, s_{i\tau} - 1\}$.

Assumption 13. (*Identification: Unbalanced Panels*) For all i ,

1. c_{i0} is observed.
2. x_{iT} and w_{iT} are observed.

3. For all i , w_{i,τ_i} has full rank d_w almost everywhere.

4. After orthogonal forward differencing, for $t = s_{i\tau}$, $\tau = 1, \dots, T_i - d_w - 1$,

$$\tilde{x}_{i,t-1} = \tilde{x}_{i,s_{i\tau}-1} = x_{i,s_{i\tau}-1} - \sum_{j=\tau+1}^{T_i-1} x_{i,s_{ij}-1} w'_{i,s_{ij}-1} \left(\sum_{j=\tau+1}^{T_i-1} w_{i,s_{ij}-1} w'_{i,s_{ij}-1} \right)^{-1} w_{i,s_{i\tau}-1}.$$

Then, the matrix $\mathbb{E} \left[\sum_{\tau=1}^{T_i-d_w-1} \tilde{x}_{i,s_{i\tau}-1} \tilde{x}'_{i,s_{i\tau}-1} \right]$ has full rank d_x .

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 1(2-b,c). They guarantee that the observed periods are 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 1 (except 2-b,c) and 13.

B.2 Identification

Conditional Independence between λ_i and σ_i^2 . Assumption 1(1-a) characterizes the correlated random coefficients model, where there can be a potential correlation between the individual heterogeneity (λ_i, σ_i^2) and the conditioning variables c_{i0} . Therefore, despite the conditional independence in Assumption 1(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.

For the random coefficients case, Assumption 1(1-a) can be altered to “ (λ_i, σ_i^2) are independent of c_{i0} and i.i.d. across i ”. Together with Assumption 1(1-d), it implies that $(\lambda_i, \sigma_i^2, c_{i0})$ are mutually independent.

In principle, we could relax the conditional independence between λ_i and σ_i^2 and still achieve identification, posterior consistency, and convergence of density forecasts under a proper set of regularity conditions. In terms of identification, one possible direction could be based on Lemma 2 in Masten (2018), but we need to at least further assume all absolute moments of λ_i and σ_i^2 are finite. In terms of implementation, we could adopt a joint MGLR_x prior on the vector of individual heterogeneity $h_i = (\lambda'_i, l_i)'$, which combines individual-specific coefficients λ_i and transformed cross-sectional heteroskedasticity $l_i = \log \frac{\bar{\sigma}^2(\sigma_i^2 - \sigma^2)}{\bar{\sigma}^2 - \sigma_i^2}$. In terms of asymptotic properties, note that Theorems 4 (posterior consistency) and 11 (convergence of density forecasts) for a general class of semiparametric models still hold. Then, the asymptotic properties could be proved in a similar manner to the heteroskedastic version of correlated random coefficients models in Theorem 10 (posterior consistency) and Corollary 12 (convergence of density forecasts). Despite the possibility of this extension, I keep the conditional independence assumption in this paper considering that Appendix E.3 yields partial evidence on the empirical relevance of this assumption.

Characteristic Function. Assumption 1(2-a) could be relaxed based on Evdokimov and White (2012).

v_{it} Distribution. Note that the normality of the shocks is a sufficient condition but not necessary. It is possible to allow some additional flexibility in v_{it} distribution. 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 and independent of (λ_i, σ_i^2) , (2) the distributions of v_{it} , $f_t^v(v_{it}|c_{i,t-1})$, 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, it seems unclear which other distribution could be a more appropriate choice *a priori*. Besides, as this paper studies panels with short time spans, time-varying shock distribution may not play a significant role.

Furthermore, it would be theoretically possible to even further extend it to the case where $f^v(v_{it}|c_{i,t-1})$ is inferred via a flexible nonparametric estimator as well (under similar identification normalizations as above). The intuition is that (λ_i, σ_i^2) varies over i whereas v_{it} varies over both i and t , so we could distinguish them with large enough cross-sectional sample size. On the other hand, empirically, it may not often be a good idea to ask too much from the finite sample, which would lead to in-sample overfitting and poor forecasts.

Example: Baseline Model. For the baseline setup in (1), we can reduce Assumption 1 and establish the identification result based on a simpler set of assumptions as follows.

Assumption 14. (*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 and independent of (y_{i0}, λ_i) .
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 and that shock is independent across firms and times, so skill and shock are intrinsically different and distinguishable. The third condition facilitates the deconvolution between the signal (skill) and the noise (shock) via Fourier 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 .

B.3 Posterior Consistency

Density Estimation. To give the intuition behind the posterior consistency argument, let us first consider a simpler scenario where we estimate the distribution of observables without deconvolution and dynamic panel data structures. 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 15. (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 $\epsilon > 0$,

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

where $d_{KL}(f_0, f) = \int f_0 \log \frac{f_0}{f}$ 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 all $\epsilon > 0$,

(a) For some $\beta > 0$, $\Pi(\mathcal{F}_N^c) = O(\exp(-\beta N))$.

(b) For some $\gamma > 0$, $\sum_j \sqrt{\mathcal{N}(\epsilon, \mathcal{F}_{N,j}) \Pi(\mathcal{F}_{N,j})} = o(\exp((1 - \gamma)N\epsilon^2))$, where $\mathcal{N}(\epsilon, \mathcal{F}_{N,j})$ is the covering number of $\mathcal{F}_{N,j}$ by balls with a radius ϵ .¹

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,

$$\Pi(U^c | x_{1:N}) = \int_{U^c} \prod_{i=1}^N \frac{f(x_i)}{f_0(x_i)} d\Pi(f) \bigg/ \int_{\mathcal{F}} \prod_{i=1}^N \frac{f(x_i)}{f_0(x_i)} d\Pi(f).$$

Intuitively speaking, for the numerator, the sieve property ensures that the sieve expands to the entire alternative region 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 around 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.

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 thus weaker assumptions on the true distribution f_0 (see Assumptions 6 and 9).

Lemma 15 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

¹As the covering number increases exponentially with the dimension of x , 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.

deconvolution problem.² Second is to incorporate an unknown shock size σ^2 in cross-sectional homoskedastic cases.³ Third is to address unknown individual-specific shock sizes σ_i^2 in cross-sectional heteroskedastic cases. Fourth is to handle strictly exogenous and predetermined variables (including lagged dependent variables) as covariates.

C Proofs

C.1 Identification

Proof. (Theorem 2)

Parts 1 and 3 for common parameters β and additive individual-heterogeneity λ_i follow earlier works such as Arellano and Bover (1995) and Arellano and Bonhomme (2012). Part 2 for cross-sectional heteroskedasticity σ_i^2 is new.

1. Identify common parameters β . First, let us perform orthogonal forward differencing of equation (2), i.e. for $t = 1, \dots, T - d_w$,

$$\tilde{y}_{it} = y_{it} - \sum_{s=t+1}^T y_{is} w'_{i,s-1} \left(\sum_{s=t+1}^T w_{i,s-1} w'_{i,s-1} \right)^{-1} w_{i,t-1}, \quad (10)$$

$$\tilde{x}_{i,t-1} = x_{i,t-1} - \sum_{s=t+1}^T x_{i,s-1} w'_{i,s-1} \left(\sum_{s=t+1}^T w_{i,s-1} w'_{i,s-1} \right)^{-1} w_{i,t-1}. \quad (11)$$

Then, β is identified given Assumption 1(2-c) and the following moment condition:

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

2. Identify the distribution of shock sizes f^{σ^2} . After orthogonal forward differencing, define

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

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

²Some 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 inferable from the data. However, here the target λ_i intertwines with u_{it} and cannot be easily inferred from the observed y_{it} .

³Note 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.

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 as⁴

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

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

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

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

Note that a chi-square distribution has non-vanishing Mellin transform, so it is legitimate to divide $M_{k^2}(\xi|c)$ on both sides

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

which recovers $M_{\sigma^2}(\xi|c)$ and hence uniquely determines 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 1(2-a), F_Λ and F_U are non-vanishing almost everywhere. Then, we obtain

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

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. Note that W is non-random conditional on c . Let $\zeta = W'\xi$ and $A_W = (W'W)^{-1}W'$, then the second derivative of $\log F_\Lambda(\zeta|c)$ is characterized by

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

⁴See the discussion on page 16 of Galambos and Simonelli (2004) for the generality of this specification.

Moreover,

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

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

Note: Once we identify $f^\lambda(\lambda_i|c_{i0})$ and $f^{\sigma^2}(\sigma_i^2|c_{i0})$, we can further recover their unconditional distributions $f^\lambda(\lambda_i)$ and $f^{\sigma^2}(\sigma_i^2)$ considering that c_{i0} is observed. $\square$

C.2 Posterior Consistency: General Semiparametric Model

Proof. (Theorem 4)

The proof builds on Canale and De Blasi (2017), which is in turn based on the early work by Barron *et al.* (1999) and Ghosal and van der Vaart (2007). Now the discussion is significantly extended to tackle convolution and common parameters. It suffices to show that: as $N \rightarrow \infty$,

- (1) for all $\delta > 0$, $\Pi^\vartheta(\vartheta \in \Theta_\delta^c | D) \rightarrow 0$,
- (2) for all $\epsilon > 0$, $\Pi^f(f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c | D) \rightarrow 0$,
- (3) for all $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that $\Pi(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N | D) \rightarrow 0$,

almost surely with respect to the true DGP (Lemmas 17, 18, and 19, respectively). We let δ depend on ϵ in point (3) because point (1) holds for all $\delta > 0$, so it holds for $\delta'(\epsilon) = \min(\delta, \delta(\epsilon))$ as well. Then, the posterior probability of the alternative region

$$\begin{aligned}& \Pi(\vartheta \in \Theta_\delta^c \text{ or } f \in \mathcal{F}_\epsilon^c | D) \\ & \leq \Pi^\vartheta(\vartheta \in \Theta_{\delta'(\epsilon)}^c | D) + \Pi^f(f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c | D) + \Pi(\vartheta \in \Theta_{\delta'(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N | D) \rightarrow 0,\end{aligned}\quad (12)$$

as $N \rightarrow \infty$, almost surely with respect to the true DGP. $\square$

Lemma 16. *Suppose condition 1-a in Theorem 4 holds, then, for all $\eta > 0$, as $N \rightarrow \infty$,*

$$\exp(N\eta) \int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow \infty,$$

almost surely with respect to the true DGP.

Proof. Similar to Barron *et al.* (1999) Lemma 4, the KL property on g (Theorem 4(1-a)) ensures that for all $\eta > 0$,

$$\mathbb{P}_0^\infty \left(\int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \leq \exp(-\eta N), \text{ infinitely often} \right) = 0,$$

where $\mathbb{P}_0^\infty$ is characterized by the true DGP when $N \rightarrow \infty$. $\square$

Lemma 17. Suppose conditions 1-a and 2 in Theorem 4 hold, then, for all $\delta > 0$, as $N \rightarrow \infty$,

$$\Pi^\vartheta (\vartheta \in \Theta_\delta^c | D) \rightarrow 0,$$

almost surely with respect to the true DGP.

Proof. Decompose the posterior probability by the sequence of exponentially consistent tests,

$$\begin{aligned} \Pi^\vartheta (\vartheta \in \Theta_\delta^c | D) &= \frac{\int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}{\int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f)} \\ &\leq \varphi_N(D) + \frac{(1 - \varphi_N(D)) \int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}{\int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}. \end{aligned} \quad (13)$$

By the Borel-Cantelli Lemma, condition 2-a in Theorem 4 implies that the first term $\varphi_N(D) \rightarrow 0$ as $N \rightarrow \infty$, almost surely with respect to the true DGP. For the numerator in the second term, note that

$$\begin{aligned} &\mathbb{E}_{\vartheta_0, f_0}^N \left[(1 - \varphi_N(D)) \int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \\ &= \int (1 - \varphi_N(D)) \left[\int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \prod_{i=1}^N g(D_i | \vartheta_0, f_0) dD \\ &= \int_{\Theta_\delta^c \times \mathcal{F}} \left[\int (1 - \varphi_N(D)) \prod_{i=1}^N g(D_i | \vartheta, f) dD \right] d\Pi(\vartheta, f) \\ &\leq \sup_{\vartheta \in \Theta^c, f \in \mathcal{F}} \mathbb{E}_{\vartheta, f} [1 - \varphi_N(D)] \\ &\leq O(\exp(-C_\varphi N)), \end{aligned}$$

where the last line follows condition 2-b in Theorem 4. Therefore, as $N \rightarrow \infty$,

$$\exp(C_\varphi N/2) (1 - \varphi_N(D)) \int_{\Theta_\delta^c \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow 0, \quad (14)$$

almost surely with respect to the true DGP. For the denominator in the second term, condition 1-a in Theorem 4 ensures Lemma 16. If we let $\eta = C_\varphi/4$, then as $N \rightarrow \infty$,

$$\exp(C_\varphi N/4) \int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow \infty, \quad (15)$$

almost surely with respect to the true DGP. Combining (13), (14), and (15), we prove the lemma. $\square$

Lemma 18. Suppose conditions 1-a and 3-a in Theorem 4 hold, then, for all $\epsilon > 0$, as $N \rightarrow \infty$,

$$\Pi^f \left(f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c \middle| D \right) \rightarrow 0,$$

almost surely with respect to the true DGP.

Proof. Decompose the posterior probability as follows,

$$\Pi^f \left(f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c \middle| D \right) = \frac{\int_{\Theta \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}{\int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}. \quad (16)$$

For the numerator,

$$\begin{aligned} & \mathbb{E}_{\vartheta_0, f_0}^N \left[\int_{\Theta \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \\ & \leq \mathbb{E}_{\vartheta_0, f_0}^N \left[\int_{\Theta \times \mathcal{F}_N^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \\ & = \int \left[\int_{\Theta \times \mathcal{F}_N^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \prod_{i=1}^N g(D_i | \vartheta_0, f_0) dD \\ & = \int_{\Theta \times \mathcal{F}_N^c} \left[\int \prod_{i=1}^N g(D_i | \vartheta, f) dD \right] d\Pi(\vartheta, f) \\ & = \Pi(\mathcal{F}_N^c) \\ & = O(\exp(-\beta N)), \end{aligned}$$

where the last line follows condition 3-a in Theorem 4. Therefore, as $N \rightarrow \infty$,

$$\exp(\beta N/2) (1 - \varphi_N(D)) \int_{\Theta \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow 0, \quad (17)$$

almost surely with respect to the true DGP. For the denominator, condition 1-a in Theorem 4 ensures Lemma 16. If we let $\eta = \beta/4$, then as $N \rightarrow \infty$,

$$\exp(\beta N/4) \int_{\Theta \times \mathcal{F}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow \infty, \quad (18)$$

almost surely with respect to the true DGP. Combining (16), (17), and (18), we prove the lemma. $\square$

Lemma 19. Suppose conditions 1 and 3-b in Theorem 4 hold, then, for all $\epsilon > 0$, there exists a

$\delta(\epsilon) > 0$,⁵ such that as $N \rightarrow \infty$,

$$\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \mid D \right) \rightarrow 0,$$

in probability with respect to the true DGP.

Proof. For all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_0$ and $f_1, f_2 \in \mathcal{F}$,

$$\|g(D_i | \vartheta_1, f_1) - g(D_i | \vartheta_2, f_2)\|_1 \leq \|g(D_i | \vartheta_1, f_1) - g(D_i | \vartheta_2, f_1)\|_1 + \|g(D_i | \vartheta_2, f_1) - g(D_i | \vartheta_2, f_2)\|_1.$$

The first term is bounded by condition 1-b in Theorem 4,

$$\|g(D_i | \vartheta_1, f_1) - g(D_i | \vartheta_2, f_1)\|_1 \leq C_g \|\vartheta_1 - \vartheta_2\|_2,$$

for some $C_g > 0$. The second term is bounded by $\|f - f_0\|_1$ based on the general “convolution” form of the individual likelihood in (8). If f is an unconditional distribution,

$$\begin{aligned} \|g(D_i | \vartheta_2, f_1) - g(D_i | \vartheta_2, f_2)\|_1 &= \int \left| \int p(D_i | \vartheta_2, h_i) f_1(h_i) dh_i - \int p(D_i | \vartheta_2, h_i) f_2(h_i) dh_i \right| dD_i \\ &\leq \int \left[\int p(D_i | \vartheta_2, h_i) dD_i \right] |f_1(h_i) - f_2(h_i)| dh_i \\ &= \|f_1 - f_2\|_1. \end{aligned}$$

If f is a conditional distribution,

$$\begin{aligned} &\|g(D_i | \vartheta_2, f_1) - g(D_i | \vartheta_2, f_2)\|_1 \\ &= \int \left| \int p(D_i \setminus c_{i0} | \vartheta_2, h_i) f_1(h_i | c_{i0}) q_0(c_{i0}) dh_i - \int p(D_i \setminus c_{i0} | \vartheta_2, h_i) f_2(h_i | c_{i0}) q_0(c_{i0}) dh_i \right| dD_i \\ &\leq \int \left[\int \left[\int p(D_i \setminus c_{i0} | \vartheta_2, h_i) d(D_i \setminus c_{i0}) \right] |f_1(h_i | c_{i0}) - f_2(h_i | c_{i0})| dh_i \right] q_0(c_{i0}) dc_{i0} \\ &= \|f_1 - f_2\|_1. \end{aligned}$$

Intuitively, the distance between g s is comparable to the distance between f s when the distance between ϑ s are small enough.

Define $\epsilon' = \epsilon/9$ and $\delta(\epsilon) = \min(\epsilon/(20C_g), \delta_0)$, then for all $\|\vartheta - \vartheta_0\|_2 < \delta(\epsilon)$, $f \in \mathcal{F}_\epsilon^c$,

$$\begin{aligned} \|g(D_i | \vartheta, f) - g(D_i | \vartheta_0, f_0)\|_1 &\geq \|f - f_0\|_1 - C_g \|\vartheta_1 - \vartheta_2\|_2 \\ &> \epsilon - \frac{1}{20}\epsilon > 8\epsilon'. \end{aligned} \tag{19}$$

⁵We let δ depend on ϵ here because Lemma 17 holds for all $\delta > 0$.

Moreover, for all $\|\vartheta_1 - \vartheta_0\|_2 < \delta(\epsilon)$ and $\|\vartheta_2 - \vartheta_0\|_2 < \delta(\epsilon)$, we have $\|\vartheta_1 - \vartheta_2\|_2 < 2\delta(\epsilon)$. Then, for all $\|f_1 - f_2\|_1 < \epsilon'/10$,

$$\begin{aligned} \|g(D_i|\vartheta_1, f_1) - g(D_i|\vartheta_2, f_2)\|_1 &\leq \|f_1 - f_2\|_1 + C_g \|\vartheta_1 - \vartheta_2\|_2 \\ &< \frac{1}{10}\epsilon' + \frac{1}{10}\epsilon = \epsilon'. \end{aligned}$$

Let $\mathcal{G}$ be the space induced by $f \in \mathcal{F}$ and $\|\vartheta - \vartheta_0\| < \delta(\epsilon)$ according to the likelihood function in (8), then for all $G \in \mathcal{G}$ induced by $F \in \mathcal{F}$, the covering number

$$\mathcal{N}(\epsilon', G) \leq \mathcal{N}(\epsilon'/10, F). \quad (20)$$

Denote $g_0(D_i) = g(D_i|\vartheta_0, f_0)$ and $g(D_i) = g(D_i|\vartheta, f)$. Based on Ghosal and van der Vaart (2007) Corollary 1, for all set G with $\inf_{g \in G} \|g - g_0\|_1 \geq 8\epsilon'$,⁶ for all $\gamma_1, \gamma_2 > 0$, there exists a test $\tilde{\varphi}_N(D)$ such that

$$\mathbb{E}_{g_0}^N \tilde{\varphi}_N(D) \leq \sqrt{\frac{\gamma_2}{\gamma_1}} \mathcal{N}(\epsilon', G) \exp(-N\epsilon'^2) \text{ and } \sup_{g \in G} \mathbb{E}_g^N [1 - \tilde{\varphi}_N(D)] \leq \sqrt{\frac{\gamma_1}{\gamma_2}} \exp(-N\epsilon'^2). \quad (21)$$

Let $\mathcal{G}_\epsilon^c$ be the induced set by $\mathcal{F}_\epsilon^c$, and $\mathcal{G}_{N,j}$ be the induced set by $\mathcal{F}_{N,j}$ in Theorem 4 condition 3-b. Then, there exists a $\tilde{\varphi}_{N,j}(D)$ for each $\mathcal{G}_\epsilon^c \cap \mathcal{G}_{N,j}$.

Define

$$H_N = \left\{ D : \int R_N(D, \vartheta, f) d\Pi(\vartheta, f) \geq \exp(-\gamma_0 N \epsilon_1^2) \right\}, \quad (22)$$

where $\gamma_0 \leq (99 + \gamma)/100$ for the γ in Theorem 4 condition 3-b. Lemma 16 implies that as $N \rightarrow \infty$, $\mathbb{P}_0^N(H_N) \rightarrow 1$, where $\mathbb{P}_0^N$ is characterized by the true DGP with the sample size being N . Hence

$$\begin{aligned} &\mathbb{E}_{g_0}^N \left[\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N \mid D \right) \right] \\ &= \mathbb{E}_{g_0}^N \left[\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_N \mid D \right) \mathbf{1}(H_N) \right] + o(1) \\ &= \sum_j \mathbb{E}_{g_0}^N \left[\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_{N,j} \mid D \right) \mathbf{1}(H_N) \right] + o(1). \end{aligned} \quad (23)$$

For each j , given (19), we have $\inf_{g \in \mathcal{G}_\epsilon^c \cap \mathcal{G}_{N,j}} \|g - g_0\|_1 \geq 8\epsilon'$, so we can decompose the posterior probability by the above sequence of tests,

$$\begin{aligned} &\mathbb{E}_{g_0}^N \left[\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \cap \mathcal{F}_{N,j} \mid D \right) \mathbf{1}(H_N) \right] \\ &\leq \mathbb{E}_{g_0}^N \tilde{\varphi}_{N,j}(D) + \mathbb{E}_{g_0}^N \left[(1 - \tilde{\varphi}_{N,j}(D)) \int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_{N,j}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \exp(\gamma_0 N \epsilon_1^2). \end{aligned} \quad (24)$$

⁶The original Ghosal and van der Vaart (2007) Corollary 1 considers the Hellinger distance, which is defined as $d_H(g, g_0) = \sqrt{\int (\sqrt{g} - \sqrt{g_0})^2}$. Note that $d_H^2(g, g_0) \leq \|g - g_0\|_1 \leq 2d_H(g, g_0)$, so $\inf_{g \in \mathcal{Q}} d_H(g, g_0) \geq 4\epsilon'$.

Let $\gamma_{1,j} = \mathcal{N}(\epsilon', \mathcal{G}_{N,j})$ and $\gamma_{2,j} = \Pi^f(\mathcal{F}_{N,j})$. For the first term,

$$\begin{aligned} & \mathbb{E}_{g_0}^N \tilde{\varphi}_{N,j}(D) \\ & \leq \sqrt{\frac{\gamma_{2,j}}{\gamma_{1,j}}} \mathcal{N}(\epsilon', \mathcal{G}_{\epsilon}^c \cap \mathcal{G}_{N,j}) \exp(-N\epsilon'^2) \\ & \leq \sqrt{\mathcal{N}(\epsilon', \mathcal{G}_{N,j}) \Pi^f(\mathcal{F}_{N,j})} \exp(-N\epsilon'^2), \end{aligned} \tag{25}$$

where the second line is given by the test in (21). For the second term, note that

$$\begin{aligned} & \mathbb{E}_{g_0}^N \left[(1 - \tilde{\varphi}_{N,j}(D)) \int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_{\epsilon}^c \cap \mathcal{F}_{N,j}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \\ & = \int (1 - \varphi_N(D)) \left[\int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_{\epsilon}^c \cap \mathcal{F}_{N,j}} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \right] \prod_{i=1}^N g(D_i | \vartheta_0, f_0) dD \\ & = \int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_{\epsilon}^c \cap \mathcal{F}_{N,j}} \left[\int (1 - \varphi_N(D)) \prod_{i=1}^N g(D_i | \vartheta, f) dD \right] d\Pi(\vartheta, f) \\ & \leq \sup_{g \in \mathcal{G}_{\epsilon}^c \cap \mathcal{G}_{N,j}} \mathbb{E}_g^N [1 - \tilde{\varphi}_N(D)] \cdot \Pi(\Theta_{\delta(\epsilon)}, \mathcal{F}_{\epsilon}^c \cap \mathcal{F}_{N,j}) \\ & \leq \sup_{g \in \mathcal{G}_{\epsilon}^c \cap \mathcal{G}_{N,j}} \mathbb{E}_g^N [1 - \tilde{\varphi}_N(D)] \cdot \Pi^f(\mathcal{F}_{N,j}) \\ & \leq \sqrt{\frac{\gamma_{1,j}}{\gamma_{2,j}}} \exp(-N\epsilon'^2) \cdot \Pi^f(\mathcal{F}_{N,j}) \\ & \leq \sqrt{\mathcal{N}(\epsilon', \mathcal{G}_{N,j}) \Pi^f(\mathcal{F}_{N,j})} \exp(-N\epsilon'^2), \end{aligned} \tag{26}$$

where the second to last line is given by the test in (21). Combining (24), (25), and (26), as $N \rightarrow \infty$,

$$\begin{aligned} & \sum_j \mathbb{E}_{g_0}^N \left[\Pi(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_{\epsilon}^c \cap \mathcal{F}_{N,j} | D) \mathbf{1}(H_N) \right] \\ & \leq \sum_j \sqrt{\mathcal{N}(\epsilon', \mathcal{G}_{N,j}) \Pi^f(\mathcal{F}_{N,j})} \exp(-N\epsilon'^2(1 - \gamma_0)) \\ & \leq \sum_j \sqrt{\mathcal{N}(\epsilon'/10, \mathcal{F}_{N,j}) \Pi^f(\mathcal{F}_{N,j})} \exp(-N\epsilon'^2(1 - \gamma_0)) \\ & = o(\exp((1 - \gamma)N\epsilon'^2/100) \exp(-N\epsilon'^2(1 - \gamma_0))) \\ & = o(\exp(-N\epsilon'^2(1 - \gamma_0 - (1 - \gamma)/100))) \\ & \rightarrow 0. \end{aligned}$$

The third line converting the covering number from the space of g to the space of f using (20), the fourth line follows the summability condition of covering numbers as in Theorem 4 condition 3-b,

and the last line is given by $\gamma_0 \leq (99 + \gamma)/100$. Then, according to (23), as $N \rightarrow \infty$,

$$\mathbb{E}_{g_0}^N \left[\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D \right) \right] \rightarrow 0.$$

Further applying Markov inequality, we obtain that as $N \rightarrow \infty$,

$$\Pi \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D \right) \rightarrow 0,$$

almost surely with respect to the true DGP. □

C.3 Posterior Consistency: (Correlated) Random Coefficients Model

C.3.1 Random Coefficients: Cross-sectional Homoskedasticity

The space for common parameters $\vartheta = (\beta, \sigma^2)$ is

$$\Theta = \mathbb{R}^{d_x} \times [\underline{\sigma}^2, \bar{\sigma}^2].$$

And the individual-specific likelihood function is characterized as

$$g(D_i | \vartheta, f) = \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2) f(\lambda_i) d\lambda_i.$$

To ensure condition 1 in Theorem 4, we work with space

$$\mathcal{F}^* = \left\{ f \in \mathcal{F} : \mathbb{E}_f \|\lambda\|_2^2 \leq M^* \right\}$$

for some large $M^* > 0$. Given Assumption 6(1-e), the second moment of f_0 exists, so f_0 satisfies this condition when M^* is large enough. Moreover, based on Doss and Sellke (1982) and Egorov's Theorem, we can establish that for all $\tau \in (0, 1)$, there exist $M^* > 0$ such that

$$\Pi^f(\mathcal{F}^*) > 1 - \tau.$$

Let Π^* be the corresponding probability measure on $\mathcal{F}^*$, i.e. for any event A ,

$$\Pi^*(A) = \Pi(A | \mathcal{F}^*).$$

We want to show that as $N \rightarrow \infty$,

$$\Pi^*(\vartheta \in \Theta_\delta^c \text{ or } f \in \mathcal{F}_\epsilon^c | D) \rightarrow 0,$$

in probability with respect to the true DGP. Then, equations (13), (16), and (24) in Appendix C.2 become

$$\begin{aligned}
\Pi^{\vartheta*}(\vartheta \in \Theta_\delta^c | D) &\leq \varphi_N(D) + \frac{(1 - \varphi_N(D)) \int_{\Theta_\delta^c \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}{\int_{\Theta \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}, \\
\Pi^{f*}\left(f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N^c \middle| D\right) &= \frac{\int_{\Theta \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_N^c \cap \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}{\int_{\Theta \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f)}, \\
\mathbb{P}_0^N \left[\Pi^* \left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_{N,j} \middle| D \right) \mathbf{1}(H_N^*) \right] \\
&\leq \mathbb{P}_0^N \tilde{\varphi}_{N,j}(D) + \mathbb{P}_0^N \left[(1 - \tilde{\varphi}_{N,j}(D)) \int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_\epsilon^c \cap \mathcal{F}_{N,j} \cap \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \mathbf{1}(H_N^*) \right], \\
\text{where } H_N^* &= \left\{ D : \int_{\Theta \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \geq \exp(-\gamma_0 N \epsilon_1^2) \right\},
\end{aligned}$$

respectively. For stand-alone terms and terms in the numerators, the set union with $\mathcal{F}^*$ renders the integrations even smaller, so it suffices to prove the original conditions without $\mathcal{F}^*$ in (14), (17), (25), and (26). For terms in the denominators, Remark 20 shows that if for all $\epsilon > 0$, $\Pi^f(f \in \mathcal{F} : d_{KL}(f_0, f) < \epsilon) > 0$, then

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

and then condition 1-a in Theorem 4 is established in the following proof.

Proof. (Theorem 7, Homoskedastic Version)

1. Condition 1-a in Theorem 4. Based on Lemma 1 in Canale and De Blasi (2017), Assumption 6 ensures that the KL property holds for f (the distribution of λ), i.e. for all $\epsilon > 0$,

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

Then, Remark 20 shows that

$$\Pi^f(f \in \mathcal{F}^* : d_{KL}(f_0, f) < \epsilon) > 0. \quad (27)$$

Now, we need to establish an altered KL property specified on g (the distribution of observables). The KL divergence of $g(D_i | \vartheta, f)$ with respect to $g(D_i | \vartheta_0, f_0)$ can be decomposed as

$$\begin{aligned}
0 &\leq \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta, f)} dD_i \\
&= \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i + \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i.
\end{aligned} \quad (28)$$

First term: Crossing out common factors in the numerator and denominator, we have

$$\int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i = \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} dD_i.$$

We can apply the convolution property of the KL divergence in Lemma 26(1) to the integral over λ_i

$$\begin{aligned} & \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} \\ & \leq \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) \log \frac{f_0(\lambda_i)}{f(\lambda_i)} d\lambda_i. \end{aligned}$$

Then, further integrating the above expression over D_i , we have

$$\begin{aligned} 0 & \leq \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i \\ & = \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} dD_i \\ & \leq \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \left[\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) \log \frac{f_0(\lambda_i)}{f(\lambda_i)} d\lambda_i \right] dD_i \\ & = \int \left[\int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) dD_i \right] f_0(\lambda_i) \log \frac{f_0(\lambda_i)}{f(\lambda_i)} d\lambda_i \\ & = d_{KL}(f_0, f). \end{aligned}$$

According to the KL property on f in (27), define

$$S_\epsilon^f = \left\{ f \in \mathcal{F}^* : d_{KL}(f_0, f) < \frac{\epsilon}{3} \right\},$$

then $\Pi^f(S_\epsilon^f) > 0$, and for all $f \in S_\epsilon^f$, the first term

$$0 \leq \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i < \frac{\epsilon}{3}. \quad (29)$$

Second term: Given the bounds in (28) and (29), we have

$$\int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i > -\frac{\epsilon}{3}.$$

Then, we only need to find an upper bound of the second term.

$$\begin{aligned}
-\frac{\epsilon}{3} &< \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i \\
&= \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2) f(\lambda_i) d\lambda_i} dD_i \\
&= \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} \\
&\quad \cdot \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2) f(\lambda_i) d\lambda_i} dD_i \\
&\leq \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} \\
&\quad \cdot \left[\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2)} d\lambda_i \right] dD_i \\
&= \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i \\
&\quad \cdot \left[\int \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i)}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f(\lambda_i) d\lambda_i} \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2)} d\lambda_i \right] dD_i \\
&= \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i \\
&\quad \cdot \left[\int \frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i) d\lambda_i} \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2)} d\lambda_i \right] dD_i,
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
m_i(\beta_0) &= \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \sum_t w_{i,t-1} (y_{it} - \beta'_0 x_{i,t-1}), \\
\Sigma_i(\sigma_0^2) &= \sigma_0^2 \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1}.
\end{aligned}$$

The second line in (30) crosses out common factors in the numerator and denominator. The third line rearranges the expression so that we can apply the convolution property of the KL divergence in Lemma 26(1) in the fourth line. The fifth line rearranges the expression so that we can cross out common factors in the numerator and denominator in the last line. Note that the log of the ratio

of normal distributions has an analytical form,

$$\begin{aligned}
& \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma^2)} \\
&= \frac{T}{2} (\log \sigma^2 - \log \sigma_0^2) + \frac{1}{2} \sum_t (y_{it} - \beta' x_{i,t-1} - \lambda'_i w_{i,t-1})^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2} \right) \\
&\quad + \sum_t \frac{(y_{it} - \beta' x_{i,t-1} - \lambda'_i w_{i,t-1})^2 - (y_{it} - \beta'_0 x_{i,t-1} - \lambda'_i w_{i,t-1})^2}{2\sigma_0^2} \\
&= \frac{T}{2} (\log \sigma^2 - \log \sigma_0^2) + \frac{1}{2} \sum_t (y_{it} - \beta' x_{i,t-1} - \lambda'_i w_{i,t-1})^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2} \right) \\
&\quad + \sum_t \frac{(\beta' x_{i,t-1})^2 - (\beta'_0 x_{i,t-1})^2 - 2(y_{it} - \lambda'_i w_{i,t-1})(\beta - \beta_0)' x_{i,t-1}}{2\sigma_0^2}. \tag{31}
\end{aligned}$$

Define

$$S_\epsilon^{\sigma^2} = \left\{ \sigma^2 \in \sigma_0^2 \left[1, \exp \left(\frac{2\epsilon}{3T} \right) \right] \right\},$$

then $\Pi^{\sigma^2}(S_\epsilon^{\sigma^2}) > 0$, and for all $\sigma^2 \in S_\epsilon^{\sigma^2}$, the sum of the first two terms is less than $\epsilon/3$. Note that $S_\epsilon^{\sigma^2}$ is asymmetric with respect to σ_0^2 because we only need to find an upper bound of $\int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i$. For the last term, if $\|\beta - \beta_0\|_2 \leq \delta_\beta$ for some $\delta_\beta > 0$,

$$\begin{aligned}
\left| (\beta' x_{i,t-1})^2 - (\beta'_0 x_{i,t-1})^2 \right| &\leq 2 \|\beta_0\|_2 \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2 + \|\beta - \beta_0\|_2^2 \|x_{i,t-1}\|_2^2 \\
&\leq (2 \|\beta_0\|_2 + \delta_\beta) \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2 \\
&= C_{11} \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2, \tag{32}
\end{aligned}$$

where C s are constants (same below). At the same time,

$$\begin{aligned}
\left| 2(y_{it} - \lambda'_i w_{i,t-1})(\beta - \beta_0)' x_{i,t-1} \right| &\leq \|\beta - \beta_0\|_2 \cdot 2 \|x_{i,t-1}\|_2 (|y_{it}| + \|\lambda_i\|_2 \|w_{i,t-1}\|_2) \\
&\leq \|\beta - \beta_0\|_2 \left(2 \|x_{i,t-1}\|_2^2 + y_{it}^2 + \|\lambda_i\|_2^2 \|w_{i,t-1}\|_2^2 \right) \\
&\leq C_{12} \|\beta - \beta_0\|_2 \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 + \|\lambda_i\|_2^2 \right). \tag{33}
\end{aligned}$$

The last line follows that $w_{i,0:T-1}$ is bounded due to Assumption 5(1). Given that T is finite, combining (32) and (33), the last term in (31) is bounded by

$$C_2 \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + \|\lambda_i\|_2^2 \right]. \tag{34}$$

Note that for all $f \in \mathcal{F}^*$, the second moment of f is bounded by M^* , we can treat f as a “prior”, ϕ as a Gaussian “likelihood”, and $\frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i) d\lambda_i}$ as the “posterior”, then the second moment

with respect to the “posterior”

$$\begin{aligned} & \int \frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i) d\lambda_i} \|\lambda_i\|_2^2 d\lambda_i \\ & \leq C_3 \left(\|m_i(\beta_0)\|_2^2 + \text{tr}(\Sigma_i(\sigma_0^2)) + M^* \right) \leq \tilde{C}_3 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + 1 \right]. \end{aligned} \quad (35)$$

Assumption 5(2) ensures that $\text{tr}(\Sigma_i(\sigma_0^2))$ is bounded above. Plugging (35) back to (34), we see that the expression in the brackets in the last line of (30) is bounded by

$$C_4 \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + 1 \right]. \quad (36)$$

Assumptions 6(1-e) and 5(1,3,4) ensure that $\mathbb{E} y_{it}^2$ and $\mathbb{E} \|x_{i,t-1}\|_2^2$ exist, so the rest of the integration in (30) is bounded by

$$C_5 \|\beta - \beta_0\|_2. \quad (37)$$

Define

$$S_\epsilon^\beta = \left\{ \|\beta - \beta_0\|_2 < \min \left(\frac{\epsilon}{3C_5}, \delta_\beta \right) \right\},$$

then $\Pi^\beta(S_\epsilon^\beta) > 0$, and for all $\beta \in S_\epsilon^\beta$, the integral associated with the last term in (31) is less than $\epsilon/3$, i.e.

$$\begin{aligned} & \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_0^2) f_0(\lambda_i) d\lambda_i \\ & \cdot \left[\int \frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i) d\lambda_i} \sum_t \frac{(\beta' x_{i,t-1})^2 - (\beta'_0 x_{i,t-1})^2 - 2(y_{it} - \lambda'_i w_{i,t-1})(\beta - \beta_0)' x_{i,t-1}}{2\sigma_0^2} d\lambda_i \right] dD_i \\ & \leq \epsilon/3. \end{aligned}$$

Therefore, for all $(\beta, \sigma^2, f) \in S_\epsilon^\beta \times S_\epsilon^{\sigma^2} \times S_\epsilon^f$, $d_{KL}(g(D_i | \vartheta_0, f_0), g(D_i | \vartheta, f)) < \epsilon$. Considering that $\Pi((\beta, \sigma^2, f) \in S_\epsilon^\beta \times S_\epsilon^{\sigma^2} \times S_\epsilon^f) > 0$ and that $S_\epsilon^f \subseteq \mathcal{F}^*$, we prove that for all $\epsilon > 0$,

$$\Pi(\vartheta \in \Theta, f \in \mathcal{F}^* : d_{KL}(g(D_i | \vartheta_0, f_0), g(D_i | \vartheta, f)) < \epsilon) > 0.$$

Then, similar to Lemma 16, we obtain that for all $\eta > 0$, as $N \rightarrow \infty$,

$$\exp(N\eta) \int_{\Theta \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow \infty,$$

almost surely with respect to the true DGP.

2. Condition 1-b in Theorem 4. For all $\vartheta_1, \vartheta_2 \in \Theta$ and $f \in \mathcal{F}$,

$$\begin{aligned}
& \|g(D_i | \vartheta_1, f) - g(D_i | \vartheta_2, f)\|_1 \\
& \leq \sum_{\tau} \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \left[\prod_{t=1}^{\tau-1} \phi(y_{it}; \beta'_1 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_1^2) \prod_{t=\tau+1}^T \phi(y_{it}; \beta'_2 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_2^2) \right. \\
& \quad \cdot \left. \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_1^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_2^2) \right| \right] \cdot f(\lambda_i) d\lambda_i dD_i \\
& = \sum_{\tau} \int \prod_{t=2}^{\tau-1} p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \left[\prod_{t=1}^{\tau-1} \phi(y_{it}; \beta'_1 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_1^2) \right. \\
& \quad \cdot \left. \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_1^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_2^2) \right| \right] \cdot f(\lambda_i) dy_{i,1:\tau} dx_{i,1:\tau-1}^{P*} dc_{i0} d\lambda_i.
\end{aligned}$$

The last line is given by integrating out y_{it} and $x_{i,t-1}^{P*}$ iteratively for $t = T, T-1, \dots, \tau+1$. According to Lemma 28 on L_1 -distance between normal distributions,

$$\begin{aligned}
& \int \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_1^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_2^2) \right| dy_{i\tau} \\
& \leq \sqrt{\frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 + \sigma_2^{-2} [(\beta_1 - \beta_2)' x_{i,\tau-1}]^2} \\
& \leq \sqrt{\frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1} + \sqrt{\sigma_2^{-2} [(\beta_1 - \beta_2)' x_{i,\tau-1}]^2}.
\end{aligned}$$

The last line follows the facts that $\log x \leq x - 1$ for all $x > -1$ and that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$. For the first term, given condition 3-b in Theorem 7, as long as $|\sigma_1^2 - \sigma_2^2| < \underline{\sigma}^2$, we have $\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_2^2} < \frac{\sigma^2}{\sigma_2^2} \leq 1$, so

$$\begin{aligned}
\sqrt{\frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1} & \leq \sqrt{\sum_{k=1}^{\infty} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 \right)^{2k} \left(\frac{1}{2k} - \frac{1}{2k+1} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 \right) \right)} \\
& \leq \sqrt{\frac{1}{6} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 \right)^2} \leq \frac{|\sigma_1^2 - \sigma_2^2|}{\sqrt{6}\underline{\sigma}^2}.
\end{aligned}$$

For the second term,

$$\sqrt{\sigma_2^{-2} [(\beta_1 - \beta_2)' x_{i,\tau-1}]^2} \leq \frac{1}{\sqrt{\underline{\sigma}^2}} \|\beta_1 - \beta_2\|_2 \|x_{i,\tau-1}\|_2.$$

$\mathbb{E}_f \|x_{i,\tau-1}\|_2$ exists based on Assumption 5(1,3,4) and the fact that for all $f \in \mathcal{F}^*$, the first moment with respect to f is bounded by $\sqrt{M^*}$. Therefore, we can set $\delta_0 = \underline{\sigma}^2$, then for all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_0$,

there exists $C_g > 0$ not depending on f such that

$$\|g(D_i|\vartheta_1, f) - g(D_i|\vartheta_2, f)\|_1 \leq C_g \|\vartheta_1 - \vartheta_2\|_2.$$

3. Condition 2 in Theorem 4. After orthogonal forward differencing in (10) and (11), we can estimate

$$\begin{aligned}\hat{\beta}_{GMM} &= \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right)^{-1} \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{y}_{it} \right), \\ \hat{\sigma}_{GMM}^2 &= \frac{1}{N(T-d_w)} \left(\sum_{i,t} \tilde{y}_{it}^2 - \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{y}_{it} \right)' \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right)^{-1} \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{y}_{it} \right) \right),\end{aligned}$$

given Assumption 1(2-c), i.e. $\mathbb{E}[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}]$ has full rank. Suppose the alternative region $\Theta^c = \{(\beta, \sigma^2) : \|\beta - \beta_0\|_2 > \Delta \text{ or } |\sigma^2 - \sigma_0^2| > \Delta'\}$. Define test

$$\varphi_N(D) = \mathbf{1} \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \text{ or } |\hat{\sigma}_{GMM}^2 - \sigma_0^2| > \frac{\Delta'}{2} \right).$$

Under the null hypothesis,

$$\begin{aligned}\hat{\beta}_{GMM} &\xrightarrow{d} N \left(\beta_0, \frac{\sigma_0^2}{N} \mathbb{E} \left[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right]^{-1} \right), \\ \hat{\sigma}_{GMM}^2 &\xrightarrow{d} N \left(\sigma_0^2, \frac{2\sigma_0^2}{N(T-d_w) - d_x} \right).\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) &= \mathbb{P}_0^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \text{ or } |\hat{\sigma}_{GMM}^2 - \sigma_0^2| > \frac{\Delta'}{2} \right) \\ &\leq \mathbb{P}_0^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \right) + \mathbb{P}_0^N \left(|\hat{\sigma}_{GMM}^2 - \sigma_0^2| > \frac{\Delta'}{2} \right) \\ &\leq \sum_{j=1}^{d_x} \mathbb{P}_0^N \left(\left| \hat{\beta}_{GMM,j} - \beta_{0,j} \right| > \frac{\Delta}{2\sqrt{d_x}} \right) + \mathbb{P}_0^N \left(|\hat{\sigma}_{GMM}^2 - \sigma_0^2| > \frac{\Delta'}{2} \right) \\ &\leq \frac{2d_x \phi \left(\frac{\Delta}{2\sqrt{d_x}} \middle/ \sqrt{\frac{\sigma_0^2}{N} \Lambda_{\min,xx}^{-1}} \right)}{\frac{\Delta}{2\sqrt{d_x}}} + \frac{2\phi \left(\frac{\Delta'}{2} \middle/ \sqrt{\frac{2\sigma_0^2}{N(T-d_w) - d_x}} \right)}{\frac{\Delta'}{2}} \\ &= \frac{4d_x^{3/2}}{\Delta} \phi \left(\frac{\Delta}{2} \sqrt{\frac{\Lambda_{\min,xx} N}{d_x \sigma_0^2}} \right) + \frac{4}{\Delta'} \phi \left(\frac{\Delta'}{2} \sqrt{\frac{N(T-d_w) - d_x}{2\sigma_0^2}} \right),\end{aligned} \tag{38}$$

where $\Lambda_{\min,xx}$ is the smallest eigenvalue of $\mathbb{E}[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}]$. The third line is given by the fact

that $\|\hat{\beta}_{GMM} - \beta_0\|_2 > \frac{\Delta}{2}$ implies that $|\hat{\beta}_{GMM,j} - \beta_{0,j}| > \frac{\Delta}{2\sqrt{d_x}}$ for at least one $j = 1, \dots, d_x$. The fourth line follows the bound of the tail of a standard normal distribution in Lemma 29. Under the alternative hypothesis,

$$\begin{aligned}\hat{\beta}_{GMM} &\xrightarrow{d} N\left(\beta, \frac{\sigma^2}{N} \mathbb{E}\left[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}\right]^{-1}\right), \\ \hat{\sigma}_{GMM}^2 &\xrightarrow{d} N\left(\sigma^2, \frac{2\sigma^2}{N(T - d_w) - d_x}\right).\end{aligned}$$

The moment conditions in Assumptions 5 ensures the existence of these asymptotic distributions. Then,

$$\begin{aligned}\mathbb{E}_{\vartheta,f}[1 - \varphi_N(D)] &= \mathbb{P}_{\vartheta,f}^N\left(\|\hat{\beta}_{GMM} - \beta_0\|_2 \leq \frac{\Delta}{2} \text{ and } |\hat{\sigma}_{GMM}^2 - \sigma_0^2| \leq \frac{\Delta'}{2}\right) \\ &\leq \mathbb{P}_{\vartheta,f}^N\left(\|\hat{\beta}_{GMM} - \beta\|_2 > \frac{\Delta}{2} \text{ and } |\hat{\sigma}_{GMM}^2 - \sigma^2| > \frac{\Delta'}{2}\right) \\ &\leq \mathbb{P}_{\vartheta,f}^N\left(\|\hat{\beta}_{GMM} - \beta\|_2 > \frac{\Delta}{2} \text{ or } |\hat{\sigma}_{GMM}^2 - \sigma^2| > \frac{\Delta'}{2}\right) \\ &\leq \frac{4d_x^{3/2}}{\Delta} \phi\left(\frac{\Delta}{2} \sqrt{\frac{\Lambda_{\min,xx} N}{d_x \sigma^2}}\right) + \frac{4}{\Delta'} \phi\left(\frac{\Delta'}{2} \sqrt{\frac{N(T - d_w) - d_x}{2\sigma^2}}\right).\end{aligned}\tag{39}$$

The second line is given by the triangular inequality. The last line follows the same argument as the calculation under the null hypothesis. As both σ_0^2 and σ^2 are bounded above by $\bar{\sigma}^2$ (condition 3 in Theorem 7), we can combine (38) and (39) and set

$$C_\varphi = \min\left\{\frac{\Delta^2 \Lambda_{\min,xx}}{8d_x \bar{\sigma}^2}, \frac{\Delta'^2 (T - d_w)}{16\bar{\sigma}^2}\right\},$$

which leads to

$$\mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) = O(e^{-C_\varphi N}), \text{ and } \sup_{\vartheta \in \Theta^c, f \in \mathcal{F}} \mathbb{E}_{\vartheta, f}[1 - \varphi_N(D)] = O(e^{-C_\varphi N}).$$

4. Condition 3 in Theorem 4. According to Corollary 1 in Canale and De Blasi (2017), Assumption 6(2) ensures that for some $c_1, c_2, c_3 > 0$, $r > (d_z - 1)/2$, and $\kappa > d_z(d_z - 1)$, for

sufficiently large $x > 0$,

$$\begin{aligned} G_0^z(\|\mu\|_2 > x) &= O\left(x^{-2(r+1)}\right), \\ G_0^z(\Lambda_1 > x) &= O\left(\exp(-c_1 x^{c_2})\right), \\ G_0^z\left(\Lambda_{d_z} < \frac{1}{x}\right) &= O\left(x^{-c_3}\right), \\ G_0^z\left(\frac{\Lambda_1}{\Lambda_{d_z}} > x\right) &= O\left(x^{-\kappa}\right), \end{aligned}$$

where Λ_1 and Λ_{d_z} are the largest and smallest eigenvalues of Ω^{-1} , respectively. Then, we can establish the sieve property on f based on Theorem 2 in Canale and De Blasi (2017). $\square$

Remark 20. Here I demonstrate that if for all $\epsilon > 0$, $\Pi^f(f \in \mathcal{F} : d_{KL}(f_0, f) < \epsilon) > 0$, then

$$\Pi^f(f \in \mathcal{F}^* : d_{KL}(f_0, f) < \epsilon) > 0. \quad (40)$$

Let $\mathcal{F}_{KL,\epsilon} = \{f \in \mathcal{F} : d_{KL}(f_0, f) < \epsilon\}$. First, based on Doss and Sellke (1982) and Egorov's Theorem, we can establish that for all $\tau \in (0, 1)$, there exists a $M^* > 0$ with $\mathcal{F}^* = \left\{f \in \mathcal{F} : \mathbb{E}_f \|\lambda\|_2^2 \leq M^*\right\}$ such that

$$\Pi^f(\mathcal{F}^*) > 1 - \tau.$$

Then, there exists $\epsilon^* > 0$ such that $\Pi^f(\mathcal{F}_{KL,\epsilon^*}) > \tau$, so we have

$$\Pi^f(\mathcal{F}^* \cap \mathcal{F}_{KL,\epsilon^*}) > 0. \quad (41)$$

(1) If $\epsilon \geq \epsilon^*$, the above expression implies $\Pi^f(\mathcal{F}^* \cap \mathcal{F}_{KL,\epsilon}) > 0$, which is equivalent to (40).

(2) If $\epsilon < \epsilon^*$, let $w = \epsilon/\epsilon^*$, then for all $f^* \in \mathcal{F}^* \cap \mathcal{F}_{KL,\epsilon^*}$, we can construct $f = wf^* + (1-w)f_0$. Thus,

$$\begin{aligned} d_{KL}(f_0, f) &= \int f_0 \log \frac{f_0}{f} \\ &\leq w \int f_0 \log \frac{f_0}{f^*} + (1-w) \int f_0 \log \frac{f_0}{f_0} \\ &< w\epsilon^* = \epsilon, \end{aligned} \quad (42)$$

where the second line is given by the convexity of $(-\log x)$. At the same time, when M^* is sufficiently large, $\mathbb{E}_{f_0} \|\lambda\|_2^2 \leq M^*$, then,

$$\int \|\lambda_i\|_2^2 f(\lambda_i) d\lambda_i = w \int \|\lambda_i\|_2^2 f^*(\lambda_i) d\lambda_i + (1-w) \int \|\lambda_i\|_2^2 f_0(\lambda_i) d\lambda_i \leq M^*. \quad (43)$$

Combining (42) and (43), we obtain $f \in \mathcal{F}^* \cap \mathcal{F}_{KL,\epsilon}$. Also note that if f_1^* and f_2^* are different, their

corresponding f_1 and f_2 are different as well. Therefore, considering (41) and the fact that Π has full support,⁷ we have

$$\Pi^f(\mathcal{F}^* \cap \mathcal{F}_{KL,\epsilon}) > 0.$$

C.3.2 Random Coefficients: Cross-sectional Heteroskedasticity

Similar to the homoskedastic case, we work with space $\mathcal{F}^{\lambda*} = \left\{ f^\lambda \in \mathcal{F}^\lambda : \mathbb{E}_{f^\lambda} \|\lambda\|_2^2 \leq M^* \right\}$ for some large $M^* > 0$. The space for common parameters $\vartheta = \beta$ is $\Theta = \mathbb{R}^{d_x}$. The individual-specific likelihood function is characterized as

$$g(D_i | \vartheta, f) = \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2.$$

Proof. (Theorem 7, Heteroskedastic Version)

1. Condition 1-a in Theorem 4. In the prior, λ and σ^2 are independent, so Lemma 26(2) implies that

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

and Lemma 26(3) implies that

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

Based on Lemma 1 in Canale and De Blasi (2017), Assumption 6 ensures that the KL property holds for f^λ and f^l . Then, Remark 20 shows that the f^λ part can be restricted to $\mathcal{F}^{\lambda*}$, so

$$\Pi^f \left(f \in \mathcal{F}^{\lambda*} \times \mathcal{F}^{\sigma^2} : d_{KL}(f_0, f) < \epsilon \right) > 0. \quad (44)$$

Now, we need to establish an altered KL property specified on g (the distribution of observables). The decomposition (28) is still valid as in the homoskedastic case.

First term: Crossing out common factors in the numerator and denominator, we have

$$\begin{aligned} & \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i \\ &= \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} dD_i. \end{aligned}$$

We can apply the general “convolution” property of the KL divergence in Remark 27 to the integral

⁷More specifically, for all f with $\text{supp}(f) \in \text{supp}(G_0)$, $f \in \text{supp}(\Pi)$ (see Theorem 3.2.4 in Ghosh and Ramamoorthi (2003)). Especially, if G_0 has full support on Θ , then Π has full support on $\mathcal{F}$. Here, G_0 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 (the spectral norm is induced by the L_2 -norm on vectors, $\|\Omega\|_2 = \max_{x \neq 0} \frac{\|\Omega x\|_2}{\|x\|_2}$).

over (λ_i, σ_i^2)

$$\begin{aligned}
& \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2 \\
& \cdot \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \\
& \leq \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) \log \frac{f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2)}{f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2)} d\lambda_i d\sigma_i^2.
\end{aligned}$$

Then, further integrating the above expression over D_i , we have

$$\begin{aligned}
0 & \leq \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i \\
& = \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} dD_i \\
& \leq \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \\
& \quad \cdot \left[\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) \log \frac{f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2)}{f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2)} d\lambda_i d\sigma_i^2 \right] dD_i \\
& = \int \left[\int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) dD_i \right] \\
& \quad \cdot f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) \log \frac{f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2)}{f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2)} d\lambda_i d\sigma_i^2 \\
& = d_{KL}(f_0, f).
\end{aligned}$$

According to the KL property on f in (44), define

$$S_\epsilon^f = \left\{ f \in \mathcal{F}^{\lambda*} \times \mathcal{F}^{\sigma^2} : d_{KL}(f_0, f) < \frac{\epsilon}{2} \right\},$$

then $\Pi^f(S_\epsilon^f) > 0$, and for all $f \in S_\epsilon^f$, the first term

$$0 \leq \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f_0)}{g(D_i | \vartheta_0, f)} dD_i < \frac{\epsilon}{2}. \quad (45)$$

Second term: Given the bounds in (28) and (45), we have

$$\int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i > -\frac{\epsilon}{2}.$$

Then, we only need to find an upper bound of the second term.

$$\begin{aligned}
-\frac{\epsilon}{2} &< \int g(D_i | \vartheta_0, f_0) \log \frac{g(D_i | \vartheta_0, f)}{g(D_i | \vartheta, f)} dD_i \\
&= \int g(D_i | \vartheta_0, f_0) \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} dD_i \\
&= \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \\
&\quad \cdot \left[\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2 \right. \\
&\quad \cdot \log \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \left. \right] dD_i \\
&\leq \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \frac{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \\
&\quad \cdot \left[\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)} d\lambda_i d\sigma_i^2 \right] dD_i \\
&= \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f_0^\lambda(\lambda_i) f_0^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2 \\
&\quad \cdot \left[\int \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2)}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \right. \\
&\quad \cdot \log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)} d\lambda_i d\sigma_i^2 \left. \right] dD_i.
\end{aligned} \tag{46}$$

The second line in (46) crosses out common factors in the numerator and denominator. The third line rearranges the expression so that we can apply the general “convolution” property of the KL divergence in Remark 27 in the fourth line. The last line rearranges the expression. Again, the log of the ratio of normal distributions has an analytical form,

$$\begin{aligned}
&\log \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)}{\prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2)} \\
&= \sum_t \frac{(y_{it} - \beta' x_{i,t-1} - \lambda'_i w_{i,t-1})^2 - (y_{it} - \beta'_0 x_{i,t-1} - \lambda'_i w_{i,t-1})^2}{2\sigma_i^2} \\
&= \sum_t \frac{(\beta' x_{i,t-1})^2 - (\beta'_0 x_{i,t-1})^2 - 2(y_{it} - \lambda'_i w_{i,t-1})(\beta - \beta_0)' x_{i,t-1}}{2\sigma_i^2} \\
&\leq C_2 \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + \|\lambda_i\|_2^2 \right],
\end{aligned} \tag{47}$$

where C s are constants (same below). The last line is similar to the homoskedastic case in Appendix C.3.1, given that T is finite, $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$, $w_{i,0:T-1}$ is bounded due to Assumption 5(1), and $\|\beta - \beta_0\|_2 \leq \delta_\beta$ for some $\delta_\beta > 0$. Note that for all $f^\lambda \in \mathcal{F}^{\lambda*}$, the second moment of f^λ is bounded by M^* . Let

$$\begin{aligned} m_i(\beta_0) &= \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \sum_t w_{i,t-1} (y_{it} - \beta'_0 x_{i,t-1}), \\ \Sigma_i(\sigma_i^2) &= \sigma_i^2 \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1}. \end{aligned}$$

We can treat f^λ as a “prior”, $\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2))$ as a Gaussian “likelihood” (conditional on σ_i^2), and $\frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) f(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) f(\lambda_i) d\lambda_i}$ as the “conditional posterior”. Then, the second moment with respect to the “conditional posterior”

$$\begin{aligned} \mathbb{E} \left[\|\lambda_i\|_2^2 \middle| \sigma_i^2, D_i \right] &= \int \frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) f^\lambda(\lambda_i)}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) f^\lambda(\lambda_i) d\lambda_i} \|\lambda_i\|_2^2 d\lambda_i \\ &\leq C_3 \left(\|m_i(\beta_0)\|_2^2 + \text{tr}(\Sigma_i(\sigma_i^2)) + M^* \right) \\ &\leq C_3 \left(\|m_i(\beta_0)\|_2^2 + \text{tr}(\Sigma_i(\bar{\sigma}^2)) + M^* \right) \\ &\leq \tilde{C}_3 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + 1 \right]. \end{aligned}$$

The third line is given by $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$. Assumption 5(2) ensures that $\text{tr}(\Sigma_i(\bar{\sigma}^2))$ is bounded above. Further integrating out $\sigma_i^2 | D_i$, we have

$$\begin{aligned} &\int \frac{\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2)}{\int \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2} \|\lambda_i\|_2^2 d\lambda_i d\sigma_i^2 \\ &= \mathbb{E} \left[\|\lambda_i\|_2^2 \middle| D_i \right] = \mathbb{E} \left[\mathbb{E} \left[\|\lambda_i\|_2^2 \middle| \sigma_i^2, D_i \right] \right] \\ &\leq \tilde{C}_3 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + 1 \right]. \end{aligned} \tag{48}$$

Plugging (48) back to (47), we see that the expression in the brackets in the last line of (46) is bounded by

$$C_4 \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + 1 \right]. \tag{49}$$

Assumptions 6(1-e) and 5(1,3,4) as well as $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$ ensure that $\mathbb{E} y_{it}^2$ and $\mathbb{E} \|x_{i,t-1}\|_2^2$ exist, so

the rest of the integration in (46) is bounded by

$$C_5 \|\beta - \beta_0\|_2.$$

Define

$$S_\epsilon^\beta = \left\{ \|\beta - \beta_0\|_2 < \min \left(\frac{\epsilon}{2C_5}, \delta_\beta \right) \right\},$$

then $\Pi^\beta(S_\epsilon^\beta) > 0$, and for all $\beta \in S_\epsilon^\beta$, the integral in (46) is less than $\epsilon/2$.

Therefore, for all $(\beta, f) \in S_\epsilon^\beta \times S_\epsilon^f$, $d_{KL}(g(D_i|\vartheta_0, f_0), g(D_i|\vartheta, f)) < \epsilon$. Considering that $\Pi((\beta, f) \in S_\epsilon^\beta \times S_\epsilon^f) > 0$ and that $S_\epsilon^f \subseteq \mathcal{F}^{\lambda*} \times \mathcal{F}^{\sigma^2}$, we prove that for all $\epsilon > 0$,

$$\Pi(\vartheta \in \Theta, f \in \mathcal{F}^* : d_{KL}(g(D_i|\vartheta_0, f_0), g(D_i|\vartheta, f)) < \epsilon) > 0.$$

Then, similar to Lemma 16, we obtain that for all $\eta > 0$, as $N \rightarrow \infty$,

$$\exp(N\eta) \int_{\Theta \times \mathcal{F}^*} R_N(D, \vartheta, f) d\Pi(\vartheta, f) \rightarrow \infty,$$

almost surely with respect to the true DGP.

2. Condition 1-b in Theorem 4. For all $\vartheta_1, \vartheta_2 \in \Theta$ and $f \in \mathcal{F}$,

$$\begin{aligned} & \|g(D_i|\vartheta_1, f) - g(D_i|\vartheta_2, f)\|_1 \\ & \leq \sum_\tau \int \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \left[\prod_{t=1}^{\tau-1} \phi(y_{it}; \beta'_1 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) \prod_{t=\tau+1}^T \phi(y_{it}; \beta'_2 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) \right. \\ & \quad \cdot \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) \right| \cdot f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) d\lambda_i d\sigma_i^2 dD_i \\ & = \sum_\tau \int \prod_{t=2}^{\tau-1} p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) p(c_{i0}) \left[\prod_{t=1}^{\tau-1} \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) \right. \\ & \quad \cdot \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) \right| \cdot f^\lambda(\lambda_i) f^{\sigma^2}(\sigma_i^2) \\ & \quad \cdot dy_{i,1:\tau} dx_{i,1:\tau-1}^{P*} dc_{i0} d\lambda_i d\sigma_i^2. \end{aligned}$$

The last line is given by integrating out y_{it} and $x_{i,t-1}^{P*}$ iteratively for $t = T, T-1, \dots, \tau+1$. According to Lemma 28 on L_1 -distance between normal distributions,

$$\begin{aligned} & \int \left| \phi(y_{i\tau}; \beta'_1 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) - \phi(y_{i\tau}; \beta'_2 x_{i,\tau-1} + \lambda'_i w_{i,\tau-1}, \sigma_i^2) \right| dy_{i\tau} \\ & \leq \sqrt{\sigma_i^{-2} [(\beta_1 - \beta_2)' x_{i,\tau-1}]^2} \\ & \leq C_1 \|\beta_1 - \beta_2\|_2 \|x_{i,\tau-1}\|_2, \end{aligned}$$

where C_1 is a constant. The last line is given by $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$. $\mathbb{E}_f \|x_{i,t-1}\|_2$ exists based on Assumption 5(1,3,4), $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$, and the fact that for all $f^\lambda \in \mathcal{F}^{\lambda*}$, the first moment with respect to f is bounded by $\sqrt{M^*}$. Therefore, we can set δ_0 to be any positive number, then for all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_0$, there exists $C_g > 0$ not depending on f such that

$$\|g(D_i | \vartheta_1, f) - g(D_i | \vartheta_2, f)\|_1 \leq C_g \|\vartheta_1 - \vartheta_2\|_2.$$

3. Condition 2 in Theorem 4. After orthogonal forward differencing in (10) and (11), we can estimate

$$\hat{\beta}_{GMM} = \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right)^{-1} \left(\sum_{i,t} \tilde{x}_{i,t-1} \tilde{y}_{it} \right),$$

given Assumption 1(2-c), i.e. $\mathbb{E}[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}]$ has full rank. Suppose the alternative region $\Theta^c = \{\beta : \|\beta - \beta_0\|_2 > \Delta\}$. Define test

$$\varphi_N(D) = \mathbf{1} \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \right).$$

As in White (1980), under the null hypothesis,

$$\begin{aligned} \hat{\beta}_{GMM} &\xrightarrow{d} N \left(\beta_0, \frac{1}{N} V_\beta \right), \\ V_\beta &= \mathbb{E} \left[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right]^{-1} \mathbb{E} \left[\sum_t \tilde{u}_{it}^2 \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right] \mathbb{E} \left[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1} \right]^{-1}. \end{aligned}$$

The moment conditions in Assumptions 5 ensures the existence of these asymptotic distributions.⁸ Then,

$$\begin{aligned} \mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) &= \mathbb{P}_0^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \right) \\ &\leq \sum_{j=1}^{d_x} \mathbb{P}_0^N \left(\left| \hat{\beta}_{GMM,j} - \beta_{0,j} \right| > \frac{\Delta}{2\sqrt{d_x}} \right) \\ &\leq \frac{4d_x^{3/2}}{\Delta} \phi \left(\frac{\Delta}{2} \sqrt{\frac{N}{d_x \Lambda_{\max, \beta}}} \right), \end{aligned} \tag{50}$$

where $\Lambda_{\max, \beta}$ is the largest eigenvalue of V_β . The argument is similar to the homoskedastic case in

⁸ $\mathbb{E}[\sum_t \tilde{u}_{it}^2 \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}]$ exists since $\tilde{u}_{it}^2 = \sigma_i^2 \tilde{v}_{it}$, $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$, and $\tilde{v}_{it}$ and $\tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}$ are independent.

Appendix C.3.1. Under the alternative hypothesis,

$$\hat{\beta}_{GMM} \xrightarrow{d} N\left(\beta, \frac{1}{N} V_\beta\right).$$

Then,

$$\begin{aligned} \mathbb{E}_{\vartheta, f} [1 - \varphi_N(D)] &= \mathbb{P}_{\vartheta, f}^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 \leq \frac{\Delta}{2} \right) \\ &\leq \mathbb{P}_{\vartheta, f}^N \left(\left\| \hat{\beta}_{GMM} - \beta \right\|_2 > \frac{\Delta}{2} \right) \\ &\leq \frac{4d_x^{3/2}}{\Delta} \phi \left(\frac{\Delta}{2} \sqrt{\frac{N}{d_x \Lambda_{\max, \beta}}} \right). \end{aligned} \tag{51}$$

The second line is given by the triangular inequality. The last line follows the same argument as the calculation under the null hypothesis. We can combine (50) and (51) and set

$$C_\varphi = \frac{\Delta^2}{8d_x \Lambda_{\max, \beta}},$$

which leads to

$$\mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) = O(e^{-C_\varphi N}), \text{ and } \sup_{\vartheta \in \Theta^c, f \in \mathcal{F}} \mathbb{E}_{\vartheta, f} [1 - \varphi_N(D)] = O(e^{-C_\varphi N}).$$

4. Condition 3 in Theorem 4. Same as the homoskedastic case in Appendix C.3.1. □

C.3.3 Correlated Random Coefficients: Cross-sectional Homoskedasticity

The proofs of correlated random coefficients models build on Pati *et al.* (2013)'s study on univariate conditional density estimation and introduce two major extensions: multivariate conditional density estimation based on location-scale mixture, and deconvolution and dynamic panel data structures.

Similar to the random coefficients homoskedastic case, we work with space

$$\mathcal{F}^* = \left\{ f \in \mathcal{F} : \int \|\lambda\|_2^2 f(\lambda|c_0) q_0(c_0) d\lambda dc_0 \leq M^* \right\}$$

for some large $M^* > 0$. The space for common parameters $\vartheta = (\beta, \sigma^2)$ is $\Theta = \mathbb{R}^{d_x} \times [\underline{\sigma}^2, \bar{\sigma}^2]$. The individual-specific likelihood function is characterized as

$$g(D_i | \vartheta, f) = \prod_t p(x_{i,t-1}^{P^*} | y_{i,t-1}, c_{i,0:t-2}) \int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2) f(\lambda_i | c_{i0}) q_{i0}(c_{i0}) d\lambda_i.$$

For conditional distributions, the induced q_0 -integrated L_1 -distance is defined as

$$\|f - f_0\|_1 \equiv \int \left[\int |f(\lambda|c_0) - f_0(\lambda|c_0)| d\lambda \right] q_0(c_0) dc_0,$$

and the induced q_0 -integrated KL divergence is

$$d_{KL}(f_0, f) \equiv \int \left[\int f_0(\lambda|c_0) \log \frac{f_0(\lambda|c_0)}{f(\lambda|c_0)} d\lambda \right] q_0(c_0) dc_0.$$

Proof. (Theorem 10, Homoskedastic Version)

1. Condition 1-a in Theorem 4. Assumption 9(1,2,3-a) ensures the induced q_0 -integrated KL property on f , i.e. for all $\epsilon > 0$,

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

Pati *et al.* (2013) Theorem 5.3 proved it for univariate λ . Here, for multivariate λ , we work with the spectral norm for the covariance matrices Ω and consider $\|\Omega\|_2 \in [\underline{\omega}, \bar{\omega}]$ as the approximating compact set in the proof of Lemma 5.5, Theorem 5.6, and Corollary 5.7 in Pati *et al.* (2013). The rest of the proof of part 1 parallels the random coefficients homoskedastic case in Appendix C.3.1, except for changing $f(\lambda_i)$ and $p(c_{i0})$ to $f(\lambda_i|c_{i0})$ and $q_{i0}(c_{i0})$, respectively, and modifying (35) and (36) for the second term: let

$$M^*(c_0) = \int \|\lambda\|_2^2 f^\lambda(\lambda|c_0) d\lambda,$$

then, as we work with $\mathcal{F}^*$,

$$\int M^*(c_0) q_0(c_0) dc_0 \leq M^*. \quad (52)$$

Now (35) becomes

$$\begin{aligned} & \int \frac{\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i|c_{i0})}{\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) f(\lambda_i|c_{i0}) d\lambda_i} \|\lambda_i\|_2^2 d\lambda_i \\ & \leq C_3 \left(\|m_i(\beta_0)\|_2^2 + \text{tr}(\Sigma_i(\sigma_0^2)) + M^*(c_{i0}) \right) \leq \tilde{C}_3 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + M^*(c_{i0}) + 1 \right], \end{aligned}$$

and (36) turns to be

$$C_4 \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 \right) + M^*(c_{i0}) + 1 \right].$$

Consider (52), once we integrate out c_{i0} , there still exists a constant C_5 as in (37).

2. Condition 1-b in Theorem 4. Similar to the random coefficients homoskedastic case in

Appendix C.3.1, except changing $f(\lambda_i)$ and $p(c_{i0})$ to $f(\lambda_i | c_{i0})$ and $q_{i0}(c_{i0})$, respectively.

3. Condition 2 in Theorem 4. Same as the random coefficients homoskedastic case in Appendix C.3.1.

4. Condition 3 in Theorem 4. Assumption 9(2,3-b,3-c) addresses the sieve property. Now the covering number is based on the induced q_0 -integrated L_1 -distance. Condition 2 resembles the random coefficients homoskedastic case in Appendix C.3.1 while expands component means to include coefficients on c_{i0} . Comparing to Theorem 5.10 in Pati *et al.* (2013), condition 2 here imposes weaker tail conditions on G_0 and hence is able to accommodate multivariate-normal-inverse-Wishart components. Conditions 3-b,c on stick breaking process directly follow Remark 5.12 and Lemma 5.15 in Pati *et al.* (2013). $\square$

C.3.4 Correlated Random Coefficients: Cross-sectional Heteroskedasticity

Similar to the random coefficients heteroskedastic case, we work with space

$$\mathcal{F}^{\lambda*} = \left\{ f^\lambda \in \mathcal{F}^\lambda : \int \|\lambda\|_2^2 f^\lambda(\lambda | c_0) q_0(c_0) d\lambda dc_0 \leq M^* \right\}$$

for some large $M^* > 0$. The space for common parameters $\vartheta = \beta$ is $\Theta = \mathbb{R}^{d_x}$. The individual-specific likelihood function is characterized as

$$\begin{aligned} g(D_i | \vartheta, f) \\ = \prod_t p(x_{i,t-1}^{P*} | y_{i,t-1}, c_{i,0:t-2}) \int \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_i^2) f^\lambda(\lambda_i | c_{i0}) f^{\sigma^2}(\sigma_i^2 | c_{i0}) q_{i0}(c_{i0}) d\lambda_i d\sigma_i^2. \end{aligned}$$

Proof. (Theorem 10, Heteroskedastic Version)

1. Condition 1-a in Theorem 4. Similar to the random coefficients heteroskedastic case in Appendix C.3.2, except changing $f^\lambda(\lambda_i)$, $f^{\sigma^2}(\sigma_i^2)$, and $p(c_{i0})$ to $f^\lambda(\lambda_i | c_{i0})$, $f^{\sigma^2}(\sigma_i^2 | c_{i0})$, and $q_{i0}(c_{i0})$, respectively and treating the conditional densities as in the correlated random coefficients homoskedastic case in Appendix C.3.3.

2. Condition 1-b in Theorem 4. Similar to the random coefficients heteroskedastic case in Appendix C.3.2, except changing $f^\lambda(\lambda_i)$, $f^{\sigma^2}(\sigma_i^2)$, and $p(c_{i0})$ to $f^\lambda(\lambda_i | c_{i0})$, $f^{\sigma^2}(\sigma_i^2 | c_{i0})$, and $q_{i0}(c_{i0})$, respectively.

3. Condition 2 in Theorem 4. Same as the random coefficients heteroskedastic case in Appendix C.3.2.

4. Condition 3 in Theorem 4. Same as the correlated random coefficients homoskedastic case in Appendix C.3.3. $\square$

C.4 Density Forecasts: General Semiparametric Model

Proof. (Theorem 11) Following the definitions in Sections 2.2 and 3.3,

$$\begin{aligned}
& \int \left| f_{i,T+1}^{cond}(y|\vartheta, f, D_i) - f_{i,T+1}^{oracle}(y|D_i) \right| dy \\
&= \int \left| \int p(y|h_i, \vartheta, D_i) p(h_i|\vartheta, f, D_i) dh_i - \int p(y|h_i, \vartheta_0, D_i) p(h_i|\vartheta_0, f_0, D_i) dh_i \right| dy \\
&= \int \left| \frac{\int p(y|h_i, \vartheta, D_i) p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) f(h_i|c_{i0}) dh_i}{\int p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) f(h_i|c_{i0}) dh_i} \right. \\
&\quad \left. - \frac{\int p(y|h_i, \vartheta_0, D_i) p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i}{\int p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i} \right| dy. \tag{53}
\end{aligned}$$

The last line follows Bayes' theorem. Here I combine the cases where f could be an unconditional distribution or a conditional distribution. In the former case, $f(h_i|c_{i0}) = f(h_i)$. Set

$$\begin{aligned}
A_i &= \int p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) f(h_i|c_{i0}) dh_i, \\
B_i(y) &= \int p(y|h_i, \vartheta, D_i) p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) f(h_i|c_{i0}) dh_i,
\end{aligned}$$

with A_{i0} and $B_{i0}(y)$ being the counterparts for the oracle predictor. Then,

$$\begin{aligned}
& \int \left| f_{i,T+1}^{cond}(y|\vartheta, f, D_i) - f_{i,T+1}^{oracle}(y|D_i) \right| dy \\
&= \int \left| \frac{B_i(y)}{A_i} - \frac{B_{i0}(y)}{A_{i0}} \right| dy \\
&\leq \frac{\int B_{i0}(y) dy \cdot |A_i - A_{i0}|}{A_{i0}A_i} + \frac{\int |B_i(y) - B_{i0}(y)| dy}{A_i},
\end{aligned}$$

and it is sufficient to establish the following four statements (Lemmas 21, 22, 23, and 25). Note that both A_i and $B_i(y)$ are non-negative by definition, so we can get rid of $|\cdot|$ for these terms. $\square$

Lemma 21. *Suppose conditions 2-a, 2-c, 3-b, and 4-a in Theorem 11 hold, then for all $\epsilon > 0$, as $N \rightarrow \infty$,*

$$\mathbb{P}(|A_i - A_{i0}| < \epsilon | D) \rightarrow 1,$$

in probability with respect to the true DGP.

Proof. Note that

$$\begin{aligned}
& |A_i - A_{i0}| \\
&\leq \int p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i \\
&\quad + \int |p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i.
\end{aligned}$$

For the first term, for $\|\vartheta - \vartheta_0\|_2 < \delta_1$,

$$\begin{aligned}
& \int p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i) - f_0(h_i)| dh_i \\
& \leq \int (p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T}) + C_h(D_i) \|\vartheta - \vartheta_0\|_2) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i \\
& \leq (M(D_i) + C_h(D_i) \delta_1) \int |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i.
\end{aligned} \tag{54}$$

The second/third line is based on Theorem 11(4-a/3-b). If f is an unconditional distribution,

$$\int |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i = \int |f(h_i) - f_0(h_i)| dh_i = \|f - f_0\|_1.$$

If f is a conditional distribution, the posterior consistency is based on the q_0 -integrated L_1 -distance. Further considering Theorem 11(2-c), we have

$$\begin{aligned}
\int |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i & \leq \frac{1}{\underline{q}} \int |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| q_0(c_{i0}) dh_i dc_{i0} \\
& \leq \frac{1}{\underline{q}} \|f - f_0\|_1.
\end{aligned}$$

Then, considering the posterior consistency of f , as $N \rightarrow \infty$, the first term

$$\mathbb{P} \left(\int p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i) - f_0(h_i)| dh_i < \frac{\epsilon}{2} \middle| D \right) \rightarrow 1,$$

in probability with respect to the true DGP. For the second term,

$$\begin{aligned}
& \int |p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i \\
& \leq M \int |p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})| dh_i \\
& \leq M C_h(D_i) \|\vartheta - \vartheta_0\|_2,
\end{aligned} \tag{55}$$

where the second/third line follows condition 2-a/4-a in Theorem 11. Since ϑ enjoys posterior consistency, as $N \rightarrow \infty$, the second term

$$\mathbb{P} \left(\int |p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i < \frac{\epsilon}{2} \middle| D \right) \rightarrow 1,$$

in probability with respect to the true DGP. $\square$

Lemma 22. Suppose conditions 2-a, 2-c, 3-b, 4-a, and 4-b in Theorem 11 hold, then for all $\epsilon > 0$,

as $N \rightarrow \infty$,

$$\mathbb{P} \left(\int |B_i(y) - B_{i0}(y)| dy < \epsilon \middle| D \right) \rightarrow 1,$$

in probability with respect to the true DGP.

Proof. Note that

$$\begin{aligned} & \int |B_i(y) - B_{i0}(y)| dy \\ & \leq \int p(y|h_i, \vartheta, D_i) p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i dy \\ & \quad + \int p(y|h_i, \vartheta, D_i) |p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i dy \\ & \quad + \int |p(y|h_i, \vartheta, D_i) - p(y|h_i, \vartheta_0, D_i)| p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i dy. \end{aligned}$$

Similar to the proof of Lemma 21, the first term can be reduced to (54),

$$\begin{aligned} & \int p(y|h_i, \vartheta, D_i) p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i dy \\ & = \int \left[\int p(y|h_i, \vartheta, D_i) dy \right] p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i \\ & = \int p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) |f(h_i|c_{i0}) - f_0(h_i|c_{i0})| dh_i, \end{aligned}$$

the second term can be reduced to (55),

$$\begin{aligned} & \int p(y|h_i, \vartheta, D_i) |p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i dy \\ & = \int \left[\int p(y|h_i, \vartheta, D_i) dy \right] |p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i \\ & = \int |p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})| f_0(h_i|c_{i0}) dh_i. \end{aligned}$$

In addition, conditions 3-b and 4-b in Theorem 11 bound the third term,

$$\begin{aligned} & \int |p(y|h_i, \vartheta, D_i) - p(y|h_i, \vartheta_0, D_i)| p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i dy \\ & \leq M(D_i) \int \left[\int |p(y|h_i, \vartheta, D_i) - p(y|h_i, \vartheta_0, D_i)| dy \right] f_0(h_i|c_{i0}) dh_i \\ & \leq M(D_i) C_y(D_i) \|\vartheta - \vartheta_0\|_2. \end{aligned}$$

Together with the posterior consistency of ϑ , as $N \rightarrow \infty$, the third term

$$\mathbb{P} \left(\int |p(y|h_i, \vartheta, D_i) - p(y|h_i, \vartheta_0, D_i)| p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i dy < \frac{\epsilon}{3} \middle| D \right) \rightarrow 1,$$

in probability with respect to the true DGP. $\square$

Lemma 23. *Suppose conditions 2-b, 2-c, 3-a, and 3-b in Theorem 11 hold, then there exists $\underline{A}_i > 0$ such that*

$$A_{i0} > \underline{A}_i.$$

Proof. Let μ_0 and V_0 be the (conditional) mean and variance of h_i based on the true distribution f_0 . Theorem 11(2-b) ensures the existence of unconditional moments. If f_0 is a conditional distribution, given Theorem 11(2-c),

$$\begin{aligned} \int \|h_i\|_2^2 f_0(h_i|c_{i0}) dh_i &\leq \frac{1}{\underline{q}} \int \|h_i\|_2^2 f_0(h_i|c_{i0}) q_0(c_{i0}) dh_i dc_{i0} \\ &\leq \frac{1}{\underline{q}} \mathbb{E}_{f_0, q_0} [\|h_i\|_2^2], \end{aligned}$$

so conditional moments exist as well. Following Chebyshev's inequality, let d_h be the dimension of h_i , we have

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

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

$$\begin{aligned} A_{i0} &= \int p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i \\ &\geq \int_{h_i \in K} p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i \\ &\geq \left(1 - \frac{d_h}{k^2} \right) \min_{h_i \in K} p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) \stackrel{\text{def}}{=} \underline{A}_i. \end{aligned}$$

Based on the extreme value theorem, the minimum exists considering the continuity condition in Theorem 11(3-a); and the minimum is positive given Theorem 11(3-b). Intuitively, since the domains of h_i in $p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T})$ and $f_0(h_i|c_{i0})$ overlap, the integral is bounded below by some positive number. $\square$

Remark 24. Moreover, together with Lemma 21, let $\epsilon = \underline{A}_i/2$, then as $N \rightarrow \infty$,

$$\mathbb{P}(A_i > \underline{A}_i/2 | D) \geq \mathbb{P}(A_{i0} - \underline{A}_i/2 > \underline{A}_i/2 | D) \rightarrow 1,$$

in probability with respect to the true DGP.

Lemma 25. *Suppose condition 3-b in Theorem 11 holds, then*

$$\int B_{i0}(y) dy < \bar{B}_i.$$

Proof. We have

$$\begin{aligned} & \int B_{i0}(y) dy \\ &= \int p(y|h_i, \vartheta_0, D_i) p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i dy \\ &= \int \left[\int p(y|h_i, \vartheta, D_i) dy \right] p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i \\ &= \int p(y_{i,1:T}|h_i, \vartheta_0, D_i \setminus y_{i,1:T}) f_0(h_i|c_{i0}) dh_i \\ &\leq M(D_i) \int f_0(h_i|c_{i0}) dh_i \\ &= M(D_i) \stackrel{\text{def}}{=} \bar{B}_i. \end{aligned}$$

The second to last line follows condition 3-b in Theorem 11. □

C.5 Density Forecasts: (Correlated) Random Coefficients Model

Proof. **(Theorem 12)**

1. Condition 2 in Theorem 11. Conditions 2-a,b in Theorem 11 are given by Theorem 7(1-b,e) (for random coefficients cases) and Theorem 10(1-b,e) (for correlated random coefficients cases), together with $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$ for heteroskedastic models (for a change of variables between l and σ^2). Conditions 2-c in Theorem 11 is assumed in Corollary 12 for correlated random coefficients cases.

2. Condition 3 in Theorem 11.

(a) Homoskedastic models. In homoskedastic models,

$$\begin{aligned} p(y_{i,1:T}|h_i, \vartheta, D_i \setminus y_{i,1:T}) &= \prod_t \phi(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2) \\ &= C_i(\beta, \sigma^2) \phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma^2)), \end{aligned}$$

where

$$\begin{aligned}
m_i(\beta) &= \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \sum_t w_{i,t-1} (y_{it} - \beta' x_{i,t-1}), \\
\Sigma_i(\sigma^2) &= \sigma^2 \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1}, \\
C_i(\beta, \sigma^2) &= \frac{1}{\sqrt{(2\pi)^{T-d_w} \left| \sum_t w_{i,t-1} w'_{i,t-1} \right|}} (\sigma^2)^{-\frac{T-d_w}{2}} \exp\left(-\frac{b_i(\beta)}{2\sigma^2}\right), \\
b_i(\beta) &= \sum_t (y_{it} - \beta' x_{i,t-1})^2 \\
&\quad - \left(\sum_t w_{i,t-1} (y_{it} - \beta' x_{i,t-1}) \right)' \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \left(\sum_t w_{i,t-1} (y_{it} - \beta' x_{i,t-1}) \right) \\
&= (y_{i,1:T} - \beta' x_{i,0:T-1}) M_{w,i} (y_{i,1:T} - \beta' x_{i,0:T-1})', \\
M_{w,i} &= I_{d_w} - w'_{i,0:T-1} (w_{i,0:T-1} w'_{i,0:T-1})^{-1} w_{i,0:T-1}.
\end{aligned} \tag{56}$$

$y_{i,1:T}$, $x_{i,0:T-1}$, and $w_{i,0:T-1}$ are $1 \times T$, $d_x \times T$, and $d_w \times T$ matrices, respectively. $M_{w,i}$ is a projection matrix projecting to the null space of $w_{i,0:T-1}$. As $C_i(\beta, \sigma^2)$ can be cancelled in the numerator and denominator of (53), we can reduce $p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T})$ to

$$\tilde{p}(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) = \phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma^2)).$$

Given the boundedness and rank condition on $w_{i,t-1}$, condition 3-a in Theorem 11 is satisfied. For condition 3-b in Theorem 11,

$$\phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)) \leq \sqrt{(2\pi\sigma_0^2)^{-d_w} \left| \sum_t w_{i,t-1} w'_{i,t-1} \right|} \stackrel{\text{def}}{=} M(D_i).$$

(b) Heteroskedastic models. In heteroskedastic models,

$$\begin{aligned}
p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T}) &= \prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) \\
&= C_i(\beta, \sigma_i^2) \phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma_i^2)),
\end{aligned}$$

with $C_i(\cdot, \cdot)$, $m_i(\cdot)$, and $\Sigma_i(\cdot)$ defined in (56). Again, given $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$ and the boundedness and rank condition on $w_{i,t-1}$, condition 3-a in Theorem 11 is satisfied. For condition 3-b in Theorem

11, as $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$,

$$\prod_t \phi(y_{it}; \beta'_0 x_{i,t-1} + \lambda'_i w_{i,t-1}, \sigma_i^2) \leq (2\pi\sigma_i^2)^{-T/2} \leq (2\pi\underline{\sigma}^2)^{-T/2} \stackrel{\text{def}}{=} M(D_i).$$

3. Condition 4 in Theorem 11.

(a) Homosekdstastic models. Condition 4 in Theorem 11 is established via Lemma 28 on L_1 -distance between normal distributions. For **condition 4-a**, similar to the argument on page A-22, as long as $|\sigma^2 - \sigma_0^2| < \sigma_0^2$,

$$\begin{aligned} & \int |\phi(\lambda_i; m(\beta), \Sigma(\sigma^2)) - \phi(\lambda_i; m(\beta_0), \Sigma(\sigma_0^2))| d\lambda_i \\ & \leq \sqrt{d_w \left(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln \frac{\sigma^2}{\sigma_0^2} \right) + \sigma_0^{-2} (\beta - \beta_0)' M_{xw,i} (\beta - \beta_0)} \\ & \leq \sqrt{\frac{d_w}{6} \frac{|\sigma^2 - \sigma_0^2|}{\sigma_0^2}} + \sqrt{\frac{\Lambda_{\max,i}}{\sigma_0^2}} \|\beta - \beta_0\|_2 \\ & \leq C_h(D_i) (\|\vartheta - \vartheta_0\|_2), \end{aligned}$$

where $M_{xw,i} = \sum_t x_{i,t-1} w'_{i,t-1} \left(\sum_t w_{i,t-1} w'_{i,t-1} \right)^{-1} \sum_t w_{i,t-1} x'_{i,t-1}$, and $\Lambda_{\max,i}$ is the largest eigenvalue of $M_{xw,i}$. The term in **condition 4-b** is

$$p(y|h_i, \vartheta_0, D_i) = \phi(y; \beta' x_{iT} + \lambda'_i w_{iT}, \sigma^2)$$

in homoskedastic models. Similarly, using Lemma 28 to bound the L_1 -distance between normal distributions, as long as $|\sigma^2 - \sigma_0^2| < \sigma_0^2$,

$$\begin{aligned} & \int |\phi(y; \beta' x_{iT} + \lambda'_i w_{iT}, \sigma^2) - \phi(y; \beta'_0 x_{iT} + \lambda'_i w_{iT}, \sigma_0^2)| dy \\ & \leq \sqrt{\frac{\sigma^2}{\sigma_0^2} - 1 - \ln \frac{\sigma^2}{\sigma_0^2} + \sigma_0^{-2} (\beta - \beta_0)' x_{iT} x'_{iT} (\beta - \beta_0)} \\ & \leq \frac{|\sigma^2 - \sigma_0^2|}{\sqrt{6}\sigma_0^2} + \frac{1}{\sqrt{\sigma_0^2}} \|\beta - \beta_0\|_2 \|x_{iT}\|_2 \\ & \leq C_y(D_i) (\|\vartheta - \vartheta_0\|_2). \end{aligned}$$

Therefore, we can set $\delta_1 = \sigma_0^2$ and obtain condition 4 in Theorem 11.

(b) **Heterosekastic models.** For **condition 4-a**,

$$\begin{aligned} & \int |p(y_{i,1:T} | h_i, \vartheta, D_i \setminus y_{i,1:T}) - p(y_{i,1:T} | h_i, \vartheta_0, D_i \setminus y_{i,1:T})| dh_i \\ & \leq \int C_i(\beta, \sigma_i^2) |\phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma_i^2)) - \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2))| d\lambda_i d\sigma_i^2 \\ & \quad + \int |C_i(\beta, \sigma_i^2) - C_i(\beta_0, \sigma_i^2)| \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) d\lambda_i d\sigma_i^2 \end{aligned}$$

For the first term,

$$\begin{aligned} & \int C_i(\beta, \sigma_i^2) |\phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma_i^2)) - \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2))| d\lambda_i d\sigma_i^2 \\ & = \int C_i(\beta, \sigma_i^2) \left[\int |\phi(\lambda_i; m_i(\beta), \Sigma_i(\sigma_i^2)) - \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2))| d\lambda_i \right] d\sigma_i^2 \\ & \leq \int C_i(\beta, \sigma_i^2) \sqrt{\sigma_i^{-2} (\beta - \beta_0)' M_{xw,i} (\beta - \beta_0)} d\sigma_i^2 \\ & \leq \tilde{C}_i(\underline{\sigma}^2)^{-\frac{T-d_w+1}{2}} \sqrt{(\beta - \beta_0)' M_{xw,i} (\beta - \beta_0)} \int d\sigma_i^2 \\ & \leq \tilde{C}_i(\underline{\sigma}^2)^{-\frac{T-d_w+1}{2}} (\bar{\sigma}^2 - \underline{\sigma}^2) \sqrt{\Lambda_{\max,i}} \|\beta - \beta_0\|_2, \end{aligned}$$

with $C_i(\cdot, \cdot)$, $m_i(\cdot)$, and $\Sigma_i(\cdot)$ defined in (56), $M_{xw,i}$ and $\Lambda_{\max,i}$ defined in part 3(a), and

$$\tilde{C}_i = \left((2\pi)^{T-d_w} \left| \sum_t w_{i,t-1} w'_{i,t-1} \right| \right)^{-1/2}.$$

The third line is based on Lemma 28 on the L_1 -distance between normal distributions. The fourth and last lines incorporate $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$. For the second term,

$$\begin{aligned} & \int |C_i(\beta, \sigma_i^2) - C_i(\beta_0, \sigma_i^2)| \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) d\lambda_i d\sigma_i^2 \\ & \leq \int |C_i(\beta, \sigma_i^2) - C_i(\beta_0, \sigma_i^2)| \left[\int \phi(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_i^2)) d\lambda_i \right] d\sigma_i^2 \\ & = \int |C_i(\beta, \sigma_i^2) - C_i(\beta_0, \sigma_i^2)| d\sigma_i^2 \\ & \leq \tilde{C}_i \int (\sigma_i^2)^{-\frac{T-d_w}{2}} \frac{|b_i(\beta) - b_i(\beta_0)|}{2\sigma_i^2} d\sigma_i^2 \\ & \leq \frac{1}{2} \tilde{C}_i(\underline{\sigma}^2)^{-\frac{T-d_w+2}{2}} |b_i(\beta) - b_i(\beta_0)| \int d\sigma_i^2 \\ & = \frac{1}{2} \tilde{C}_i(\underline{\sigma}^2)^{-\frac{T-d_w+2}{2}} (\bar{\sigma}^2 - \underline{\sigma}^2) |b_i(\beta) - b_i(\beta_0)|. \end{aligned}$$

The fourth line follows the fact that for all $x_1, x_2 \geq 0$, $|e^{-x_1} - e^{-x_2}| \leq |x_1 - x_2|$. The fifth and last lines incorporate $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$. Note that as $M_{w,i}$ is a projection matrix, the largest eigenvalue of

$M_{w,i}$ is 1. If $\|\beta - \beta_0\|_2 \leq \delta_\beta$ for some $\delta_\beta > 0$,

$$\begin{aligned} |b_i(\beta) - b_i(\beta_0)| &\leq 2\|\beta - \beta_0\|_2 \|x_{i,0:T-1}\|_2 (\|y_{i,1:T}\|_2 + \|\beta_0\|_2 \|x_{i,0:T-1}\|_2) \\ &\quad + \|\beta - \beta_0\|_2^2 \|x_{i,0:T-1}\|_2^2 \\ &\leq \|\beta - \beta_0\|_2 \left[2\|y_{i,1:T}\|_2 \|x_{i,0:T-1}\|_2 + (2\|\beta_0\|_2 + \delta_\beta) \|x_{i,0:T-1}\|_2^2 \right] \end{aligned}$$

The term in **condition 4-b** is

$$p(y|h_i, \vartheta_0, D_i) = \phi(y; \beta'x_{iT} + \lambda'_i w_{iT}, \sigma_i^2)$$

in heteroskedastic models. Similarly, using Lemma 28 to bound the L_1 -distance between normal distributions,

$$\begin{aligned} &\int |\phi(y; \beta'x_{iT} + \lambda'_i w_{iT}, \sigma_i^2) - \phi(y; \beta'_0 x_{iT} + \lambda'_i w_{iT}, \sigma_i^2)| dy \\ &\leq \sqrt{\sigma_i^{-2} (\beta - \beta_0)' x_{iT} x'_{iT} (\beta - \beta_0)} \\ &\leq \sqrt{\underline{\sigma}^{-2} (\beta - \beta_0)' x_{iT} x'_{iT} (\beta - \beta_0)} \\ &\leq \frac{1}{\sqrt{\underline{\sigma}^2}} \|\beta - \beta_0\|_2 \|x_{iT}\|_2 \\ &= C_y(D_i) \|\beta - \beta_0\|_2. \end{aligned}$$

Therefore, we can set $\delta_1 = \delta_\beta$ and obtain condition 4 in Theorem 11. $\square$

C.6 Useful Lemmas

Lemma 26. (Properties of KL Divergence)

1. (Convolution) The KL divergence is non-increasing after convolution. If f_0 , f , and p are distributions, let $g_0(y) = \int p(y-x) f_0(x) dx$ be the convolution of f_0 and p , and similarly $g(y) = \int p(y-x) f(x) dx$, then,

$$d_{KL}(g_0, g) \leq d_{KL}(f_0, f).$$

2. (Independence) The KL divergence is additive for independent distributions. If f_0^x and f_0^y are independent distributions with joint distribution $f_0(x, y) = f_0^x(x) f_0^y(y)$, and similarly $f(x, y) = f^x(x) f^y(y)$, then,

$$d_{KL}(f_0, f) = d_{KL}(f_0^x f_0^y, f^x f^y) = d_{KL}(f_0^x, f^x) + d_{KL}(f_0^y, f^y).$$

3. (Variable Transformation) The KL divergence is invariant under variable transformation. If

$y = h(x)$, where h is a continuous function,

$$d_{KL}(f_0^y, f^y) = d_{KL}(f_0^x, f^x).$$

Proof. Property 1 (Convolution): Define $h(x) = x \log x$, then $h(x)$ is a concave function. Note that

$$\begin{aligned} g_0(y) \log \frac{g_0(y)}{g(y)} &= g(y) h\left(\frac{g_0(y)}{g(y)}\right) \\ &= g(y) h\left(\int \frac{p(y-x)f(x)}{\int p(y-x)f(x)dx} \cdot \frac{f_0(x)}{f(x)} dx\right) \\ &\leq g(y) \int \frac{p(y-x)f(x)}{\int p(y-x)f(x)dx} \cdot h\left(\frac{f_0(x)}{f(x)}\right) dx \\ &= \int p(y-x)f(x) h\left(\frac{f_0(x)}{f(x)}\right) dx \\ &= \int p(y-x)f_0(x) \log \frac{f_0(x)}{f(x)} dx, \end{aligned} \tag{57}$$

where the inequality is given by Jensen's inequality. Then, further integrating the above expression over y , we have

$$\begin{aligned} d_{KL}(g_0, g) &= \int g_0(y) \log \frac{g_0(y)}{g(y)} dy \\ &\leq \int p(y-x)f_0(x) \log \frac{f_0(x)}{f(x)} dx dy \\ &= \int f_0(x) \log \frac{f_0(x)}{f(x)} dx \\ &= d_{KL}(f_0, f), \end{aligned}$$

where the inequality follow the above derivation (57).

Properties 2 (Independence) and 3 (Variable Transformation) can be directly derived from the definition of the KL divergence. $\square$

Remark 27. We can extend Property 1 to a more general “convolution” form. Let $u = h(x, y)$, then $g_0(y) = \int \frac{\partial h(x, y)}{\partial y} p(h(x, y)) f_0(x) dx$ and $g(y) = \int \frac{\partial h(x, y)}{\partial y} p(h(x, y)) f(x) dx$. We can obtain $d_{KL}(g_0, g) \leq d_{KL}(f_0, f)$ in a similar fashion. Especially, in Property 1 above, $u = y - x$.

Lemma 28. (L_1 -Distance between Normal Distributions) Suppose we have two multivariate normal distributions $\phi(x; \mu_1, \Sigma_1)$ and $\phi(x; \mu_2, \Sigma_2)$, where x is a $d_x \times 1$ vector, then

$$\|\phi(x; \mu_1, \Sigma_1) - \phi(x; \mu_2, \Sigma_2)\|_1 \leq \sqrt{\text{tr}(\Sigma_2^{-1}\Sigma_1) + \log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d_x + (\mu_2 - \mu_1)' \Sigma_2^{-1} (\mu_2 - \mu_1)}.$$

Proof. We can first bound the L_1 -distance by the KL divergence using Pinsker’s inequality

$$\|\phi(x; \mu_1, \Sigma_1) - \phi(x; \mu_2, \Sigma_2)\|_1 \leq \sqrt{2d_{KL}(\phi(x; \mu_1, \Sigma_1), \phi(x; \mu_2, \Sigma_2))},$$

and then plug in the formula of the KL divergence between multivariate normals. $\square$

Lemma 29. (Tail of Normal Distribution) *If x follows a standard normal distribution, $x \sim N(0, 1)$, then for $x^* > 0$,*

$$\mathbb{P}(x > x^*) \leq \frac{\phi(x^*)}{x^*}.$$

Proof. See Feller (1968). $\square$

D Algorithms

D.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 36). The estimates and forecasts of 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 .

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

Note that due to the truncation approximation, the probability for component K is different from its infinite mixture counterpart in (6). I denote the above truncated stick-breaking process as $p_k \sim \text{TSB}(1, \alpha, K)$, 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 31(2). For individual het-

erogeneity $z = \lambda, l$, let γ_i^z be individual i 's component affiliation, which can take values $\{1, \dots, K^z\}$, 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 parameters is given by

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

where $z = \lambda, l$, $k = 1, \dots, K^z$, $i = 1, \dots, N$, and $t = 1, \dots, T$. 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) .⁹

The proposed 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 30. (Random Coefficients with Cross-sectional Heteroskedasticity)¹⁰ *For each iteration $s = 1, \dots, n_{sim}$,*

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

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

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

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

⁹The 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 D.3 for details of the baseline model with random effects.

¹⁰Below, 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 D.5.

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

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

2. *Component parameters:* For $z = \lambda, l$, for $k = 1, \dots, K^z$, 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)

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

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

(b) For $i = 1, \dots, N$, draw $l_i^{(s)}$ via the random-walk Metropolis-Hastings approach,

$$p\left(l_i^{(s)}\left|\mu_{\gamma_i^{(s)}}^{l(s)}, \Omega_{\gamma_i^{(s)}}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i\right.\right) \\ \propto \phi\left(l_i^{(s)}; \mu_{\gamma_i^{(s)}}^{l(s)}, \Omega_{\gamma_i^{(s)}}^{l(s)}\right) \prod_{t=1}^T \phi\left(y_{it}; \lambda_i^{(s)'} w_{i,t-1} + \beta^{(s-1)'} x_{i,t-1}, \sigma^2\left(l_i^{(s)}\right)\right),$$

where $\sigma^2(l) = \frac{\bar{\sigma}^2 - \sigma^2}{1 + \bar{\sigma}^2 \exp(-l)} + \underline{\sigma}^2$. Then, calculate $(\sigma_i^2)^{(s)}$ based on $\sigma^2(l)$.

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

Remark 31. (1) 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 33 for details.

(2) 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, $(\beta^{(s)}, \sigma^{2(s)})$ are drawn from a linear regression model with “unknown” variance, $p\left(\beta^{(s)}, \sigma^{2(s)} \mid \left\{\lambda_i^{(s)}\right\}, D\right)$.

D.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 expression (3) in the main text for balanced panels or expression (9) in Appendix B.1 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 $A_k = C_k B_k$. According to the expressions in Assumption 9(3), we can denote $\eta = \beta/3$ and let B_k^η follow the standard exponential distribution, i.e. $p(B_k^\eta) = \exp(-B_k^\eta)$. We can also let $C_k = C_* k^{-(3\eta+2)/(\gamma\eta)} (\log k)^{-1/\eta}$ for large ks , where C_* is a constant, $\gamma \in (0, 1)$, and $\eta(1 - \gamma) > d_{c_0}$. This prior structure satisfies Pati *et al.* (2013) Remark 5.12 that ensures strong consistency.¹¹ Furthermore, it is helpful to introduce a set of auxiliary stochastic functions $\xi_k(c_{i0})$, $k = 1, 2, \dots$, 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 (7) can be recovered by marginalizing over $\{\xi_k(c_{i0})\}$. Finally, I combine 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 $\boldsymbol{\zeta}_k = [\zeta_k(c_{10}), \zeta_k(c_{20}), \dots, \zeta_k(c_{N0})]'$ and $\boldsymbol{\xi}_k = [\xi_k(c_{10}), \xi_k(c_{20}), \dots, \xi_k(c_{N0})]'$, as well as an $N \times N$ matrix $\mathbf{V}_k \equiv \tilde{V}(A_k)$ with the i, j -th element being $(\mathbf{V}_k)_{ij} = \exp\left(-A_k \|c_{i0} - c_{j0}\|_2^2\right)$. The next algorithm extends Algorithm 30 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 30.

¹¹In practice, to ensure that $V_k(c, \tilde{c})$ would not decay too fast to an identity matrix as k increases, we can set η to be very large, and γ to be smaller than but very close to 1. Then, C_k would be close to $C_* k^{-3}$ essentially. I choose C_* to be 5 in Monte Carlo simulations and the empirical application, but the results are robust across a range of C_* , e.g. from 1 to 10.

Algorithm 32. (Correlated Random Coefficients with Cross-sectional Heteroskedasticity)¹² For each iteration $s = 1, \dots, n_{sim}$,

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

(a) For $k = 1, \dots, K^z - 1$, draw $A_k^{z(s)}$ via the random-walk Metropolis-Hastings approach,¹³

$$p\left(A_k^{z(s)} \mid \zeta_k^{z(s-1)}, \{c_{i0}\}\right) \propto \left(A_k^{z(s)}\right)^{\eta-1} \exp\left(-\left(\frac{A_k^{z(s)}}{C_k}\right)^\eta\right) \cdot \phi\left(\zeta_k^{z(s-1)}; 0, \tilde{V}\left(A_k^{z(s)}\right)\right).$$

Then, calculate $\mathbf{V}_k^{z(s)} = \tilde{V}\left(A_k^{z(s)}\right)$.

(b) For $k = 1, \dots, K^z - 1$, and $i = 1, \dots, N$, draw $\xi_k^{z(s)}(c_{i0})$ from a truncated normal distribution $p\left(\xi_k^{z(s)}(c_{i0}) \mid \zeta_k^{z(s-1)}(c_{i0}), \gamma_i^{z(s-1)}\right)$.¹⁴

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

(c) For $k = 1, \dots, K^z - 1$, $\zeta_k^{z(s)}$ from a multivariate normal distribution $p\left(\zeta_k^{z(s)} \mid \mathbf{V}_k^{z(s)}, \xi_k^{z(s)}\right)$:

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

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

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

2. *Component parameters:* For $z = \lambda, l$, for $k = 1, \dots, K^z$,

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

(b) Draw $\Omega_k^{z(s)}$ from an inverse Wishart distribution (or an inverse gamma distribution if z is a scalar) $p\left(\Omega_k^{z(s)} \mid \mu_k^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_k^{z(s-1)}}\right)$.

3. *Component memberships:* For $z = \lambda, l$, for $i = 1, \dots, N$, draw $\gamma_i^{z(s)}$ from a multinomial distribu-

¹²See Remark 31(2) for the adaption to cross-sectional homoskedastic models.

¹³The first term comes from the change of variables from B_k^η to A_k .

¹⁴ $\mathbf{1}(\cdot)$ is an indicator function.

tion $p\left(\left\{\gamma_i^{z(s)}\right\}\left|\left\{p_k^{z(s)}, \mu_k^{z(s)}, \Omega_k^{z(s)}\right\}, z_i^{(s-1)}, c_{i0}\right)\right):$

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

4. *Individual-specific parameters:*

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

(b) For $i = 1, \dots, N$, draw $l_i^{(s)}$ via the random-walk Metropolis-Hastings approach $p\left(l_i^{(s)} \left| \mu_{\gamma_i^{(s)}}^{l(s)}, \Omega_{\gamma_i^{(s)}}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i\right.\right)$, then calculate $(\sigma_i^2)^{(s)}$ based on $\sigma^2(l)$.

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

D.3 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 normal-inverse-gamma form,

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

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.

$$a_0^{\sigma^2} = 2, \tag{58}$$

$$b_0^{\sigma^2} = \hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right) \cdot (a_0^{\sigma^2} - 1) = \hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right), \tag{59}$$

$$m_0^\beta = 0.5, \tag{60}$$

$$\psi_0^\beta = \frac{1}{b_0^{\sigma^2} / (a_0^{\sigma^2} - 1)} = \frac{1}{\hat{E}^i\left(\widehat{Var}_i^t(y_{it})\right)}. \tag{61}$$

In (59) here and (62) below, $\hat{E}_i^t$ and $\widehat{Var}_i^t$ stand for the sample mean and variance for firm i over $t = 1, \dots, T$, and $\hat{E}^i$ and $\widehat{Var}^i$ are the sample mean and variance over the whole cross-section $i = 1, \dots, N$. Equation (59) ensures that on average the prior and the data have a similar scale. Equation (60) 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 (58) characterizes a rather less informative prior on σ^2 with infinite variance, and (61) assumes that the prior variance of β is equal to 1 on average.

The hyperpriors for the DPM prior are specified as:

$$G_0(\mu_k, \omega_k^2) = N(m_0^\lambda, \psi_0^\lambda \omega_k^2) \text{IG}(a_0^\lambda, b_0^\lambda),$$

$$\alpha \sim \text{Ga}(a_0^\alpha, b_0^\alpha).$$

Similarly, the hyperparameters are chosen to be:

$$a_0^\lambda = 2, b_0^\lambda = \widehat{Var}^i(\hat{E}_i^t(y_{it})) \cdot (a_0^\lambda - 1) = \widehat{Var}^i(\hat{E}_i^t(y_{it})), \quad (62)$$

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

$$a_0^\alpha = 2, b_0^\alpha = 2. \quad (63)$$

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 (63) 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:

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

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 the results are not very sensitive to hyperparameters as long as the implied priors are flexible enough to cover the range of observables.

D.4 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 33. (Adaptive RWMH) *Let us consider a generic variable θ . For each iteration $s = 1, \dots, n_{sim}$,*

1. *Draw candidate $\tilde{\theta}$ from the random-walk proposal density $\tilde{\theta} \sim N(\theta^{(s-1)}, \zeta^{(s)}\Sigma)$.*
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,

$$\log \zeta^{(s+1)} = \rho \left(\log \zeta^{(s)} + s^{-c} \left(a.r.(\tilde{\theta}|\theta^{(s-1)}) - a.r.^* \right) \right),$$

where $0.5 < c \leq 1$, $a.r.^*$ is the target acceptance rate, and

$$\rho(x) = \min(|x|, \bar{x}) \cdot \text{sgn}(x),$$

where $\bar{x} > 0$ is a very large number.

Remark 34. (1) In step 1, since the algorithms in this paper only consider RWMH on conditionally independent scalar variables, Σ is simply taken to be 1.

(2) In step 4, I choose $c = 0.55$, $a.r.^* = 30\%$ in the numerical exercises, following Griffin (2016).

D.5 Details on Posterior Samplers

D.5.1 Step 2: Component Parameters

Random Coefficients Model For $z = \lambda, l$ and $k = 1, \dots, K^z$, 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) $p \left(\mu_k^{z(s)}, \Omega_k^{z(s)} \middle| \left\{ z_i^{(s-1)} \right\}_{i \in J_k^{z(s-1)}} \right)$:

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

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 $k = 1, \dots, K^z$,

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

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

(b) Draw $\Omega_k^{z(s)}$ from an inverse Wishart distribution (or an inverse gamma distribution if z is a scalar) $p \left(\Omega_k^{z(s)} \left| \mu_k^{z(s)}, \left\{ z_i^{(s-1)}, c_{i0} \right\}_{i \in J_k^{z(s-1)}} \right. \right)$:

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

D.5.2 Step 4: Individual-specific Parameters

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

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

where the conditional “prior” mean is characterized by

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

D.5.3 Step 5: Common parameters

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

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

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

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

Remark 35. For unbalanced panels, the summations and products in steps 4 and 5 (Subsections D.5.2 and D.5.3) are instead over $t \in s_{i,1:T_i-1}$, the observed periods of individual i used for estimation.

D.6 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, 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 , $i = 1, \dots, N$, are i.i.d. standard uniform random variables, i.e. $u_i^z \sim U(0, 1)$. Then, the mixture of components in (5) can be rewritten as

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

where $z = \lambda, l$. By marginalizing over u_i^z , we can recover (5). Accordingly, we can define the number of active components as

$$K^{z,A} = \max_{1 \leq i \leq N} \gamma_i^z,$$

and the number of potential components (including active components) as

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

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 36. (Slice-Retrospective: Random Coefficients with Cross-sectional Heteroskedasticity)

For each iteration $s = 1, \dots, n_{sim}$, steps 1-3 in Algorithm 30 are modified as follows. For $z = \lambda, l$,

1. Active components:

(a) Number of active components:

$$K^{z,A} = \max_{1 \leq i \leq N} \gamma_i^{z(s-1)}.$$

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

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

(c) Component parameters: for $k = 1, \dots, K^{z,A}$, 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 30 step 2.

(d) Label switching: jointly update $\{p_k^{z(s)}, \theta_k^{z(s)}, \gamma_i^{z(s)}\}_{k=1}^{K^{z,A}}$ based on $\{p_k^{z*}, \theta_k^{z*}, \gamma_i^{z(s-1)}\}_{k=1}^{K^{z,A}}$ 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 $i = 1, \dots, N$, draw $u_i^{z(s)}$ from a uniform distribution $p\left(u_i^{z(s)} \mid \left\{p_k^{z(s)}\right\}, \gamma_i^{z*}\right)$:

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

3. DP scale parameter:

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

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

- (b) Draw $\alpha^{z(s)}$ from a mixture of two gamma distributions $p\left(\alpha^{z(s)} \mid \xi^{z(s)}, K^{z,A}, N\right)$:

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

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

4. Potential components:

- (a) Component probabilities: start with $K^{z*} = K^{z,A}$,

- i. if $\left(1 - \sum_{j=1}^{K^{z*}} p_j^{z(s)}\right) < \min_{1 \leq i \leq N} u_i^{z(s)}$, set $K^{z,P} = K^{z*}$ and stop;

- ii. otherwise, let $K^{z*} = K^{z*} + 1$, draw $\zeta_{K^{z*}}^z \sim \text{Beta}\left(1, \alpha^{z(s)}\right)$, update $p_{K^{z*}}^{z(s)} = \zeta_{K^{z*}}^z \prod_{j < K^{z*}} (1 - \zeta_j^z)$, and go to step (a-i).

- (b) Component parameters: for $k = K^{z,A} + 1, \dots, K^{z,P}$, draw $\theta_k^{z(s)}$ from the DP base distribution G_0^z .

5. Component memberships: For $i = 1, \dots, N$, draw $\gamma_i^{z(s)}$ from a multinomial distribution $p\left(\left\{\gamma_i^{z(s)}\right\} \mid \left\{p_k^{z(s)}, \mu_k^{z(s)}, \Omega_k^{z(s)}\right\}, u_i^{z(s)}, z_i^{(s-1)}\right)$:

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

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

The remaining part of the algorithm resembles steps 4 and 5 in Algorithm 30.

Remark 37. Note that:

- (1) 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^z s being integrated out. Thus, extra caution is needed in dealing with the order of the steps.

(2) 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.

(3) Step 3 for DP scale parameter α^z follows Escobar and West (1995). It is different from step 1-a in Algorithm 30 due to the unrestricted number of components in the current sampler.

(4) 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.

(5) 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.

D.7 Parametric Specification of Heteroskedasticity

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

$$\sigma_i^2 \sim \text{IG}(a, b).$$

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:

$$\begin{aligned} b &\sim \text{Ga}(a_0^b, b_0^b), \\ p(a|b, a_0^a, b_0^a, c_0^a) &\propto \frac{(a_0^a)^{-1-a} (b)^{ac_0^a}}{\Gamma(a)^{b_0^a}}. \end{aligned} \tag{64}$$

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 38. (Parametric Specification: Cross-sectional Heteroskedasticity) *For each iteration $s = 1, \dots, n_{sim}$,*

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

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

which is characterized by the same kernel form as expression (64) with

$$\log(a_1^a) = \log(a_0^a) + \sum_{i=1}^N \log(\sigma_i^{2(s-1)}),$$

$$b_1^a = b_0^a + N,$$

$$c_1^a = c_0^a + N.$$

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

$$b^{(s)} \sim Ga\left(a_1^b, b_1^b\right),$$

$$a_1^b = a_0^b + Na^{(s)},$$

$$b_1^b = b_0^b + \sum_{i=1}^N \left(\sigma_i^{2(s-1)}\right)^{-1}.$$

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

$$\sigma_i^{2(s)} \sim IG(a_i, b_i),$$

$$a_i = a^{(s)} + T/2,$$

$$b_i = b^{(s)} + \frac{1}{2} \sum_{t=1}^T \left(y_{it} - \beta^{(s-1)'} x_{i,t-1} - \lambda_i^{(s)'} w_{i,t-1}\right)^2.$$

E Monte Carlo Simulation and Empirical Application

E.1 Point Forecasts

Point Forecast Evaluation. 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

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

Table 7: Point Forecast Evaluation: Baseline Model with Random Effects

	Degenerate	Skewed	Bimodal
<i>Oracle</i>	<i>0.250</i>	<i>0.289</i>	<i>0.270</i>
Homog	0.039***	0.092***	0.340***
Flat	0.099***	0.004***	0.021***
Param	0.041	0.001***	0.019***
NP-disc	0.039***	0.091***	0.019***
NP-R	0.041	0.0001	0.003

Notes: The point forecasts are assessed by MSE and the Timmermann and Zhu (2019) test. For the oracle predictor, the table reports the exact values of MSE (averaged over 1,000 Monte Carlo samples). For other predictors, the table reports their differences from the oracle. The tests compare other feasible predictors with NP-R, with significance levels indicated by *: 10%, **: 5%, and ***: 1%. The entries in bold indicate the best feasible predictor in each column.

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 Timmermann and Zhu (2019) test, which extends the Diebold and Mariano (1995) test to panel data setups, is further implemented to assess whether the difference in the MSE is significant.

Baseline Model with Random Effects. For each experiment, point forecasts and density forecasts share comparable rankings (Table 7).

General Model. Considering point forecasts, Heterosk-Param and Heterosk-NP-disc constitute the first tier, Heterosk-NP-R can be viewed as the second tier, Heterosk-NP-C and Homosk-NP-C are the third tier, and Homog and Heterosk-Flat are markedly inferior (Table 8). It is not very 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.

Empirical Application. Most of the estimators are comparable according to MSE, with only Flat performing significantly poorly (Table 9). Intuitively, shrinkage in general leads to better forecasting performance, especially for point forecasts, but the Flat prior does not introduce any shrinkage to individual effects (λ_i, σ_i^2) . Conditional on the common parameter β , the Flat estimator of (λ_i, σ_i^2) 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).

Table 8: Point Forecast Evaluation: General Model

		Normal v_{it}	Skewed v_{it}
<i>Oracle</i>		<i>0.492</i>	<i>0.486</i>
Homog		0.444***	0.451***
Homosk	NP-C	0.076***	0.084**
Heterosk	Flat	0.580***	0.596***
	Param	0.043***	0.052***
	NP-disc	0.045***	0.053***
	NP-R	0.059***	0.066***
	NP-C	0.079	0.082

Notes: For point forecast evaluation, see the description in Table 7. Here the tests are conducted with respect to Heterosk-NP-C.

Table 9: Point Forecast Evaluation: Young Firm Dynamics

		MSE
<i>Heterosk</i>	<i>NP-C/R</i>	<i>0.197</i>
Homog		0.015
Homosk	NP-C	0.005
Heterosk	Flat	0.292**
	Param	-0.0001
	NP-disc	0.009
	NP-R	0.001
	NP-C	-0.002**

Notes: See the description of Table 7 for point forecast evaluation. Here Heterosk-NP-C/R is the benchmark for both normalization and significance tests. For Heterosk-NP-C/R, the table reports the exact values of MSE. For other predictors, the table reports their differences from Heterosk-NP-C/R.

Figure 7: Convergence Diagnostics: β

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

E.2 Baseline Model with Random Effects

MCMC convergence. Both the Brook-Draper diagnostic and the Raftery-Lewis diagnostic yield desirable MCMC accuracy. Figures 7 to 10 show trace plots, prior/posterior distributions, rolling means, and autocorrelation graphs of β , σ^2 , α , and λ_i ($i = 1$).

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

Figure 8: Convergence Diagnostics: σ^2

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

Figure 9: Convergence Diagnostics: α

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

Figure 10: Convergence Diagnostics: λ_i ($i = 1$)

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

Figure 11: f_0 vs $\Pi(f|y_{1:N,0:T})$: Baseline Model with Bimodal Random Effects, $N = 10^5$

Notes: The black solid lines represent the true λ_i distributions, f_0 . The teal bands show the posterior distribution of f , $\Pi(f|y_{1:N,0:T})$.

ratio), and different underlying λ_i distributions (including standard normal and fat tail distributions). In 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 NP-R predictor is more advantageous 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 teal bands in Figure 2 get closer to the true f_0 and eventually completely overlaps it (see Figure 11), 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 DGPs. They provide some improvement in the corresponding situations, but call for extra computation efforts.

E.3 Empirical Application

E.3.1 Additional Figures and Tables.

Below are additional Figures and Tables that complement the main results in the text.

Figure 12: Distributions of Observables

Figure 13: PIT: All Predictors

Notes: Teal lines indicate the confidence interval.

Figure 14: Predictive Distributions: Firm-level, 4 Types (Regrouped)

Notes: Predictive distributions are regrouped according to predictors. The blue solid / orange dotted / yellow dashed / purple dash-dot lines are the predictive distributions of typical firms a/b/c/d in Figure 4 in the main text, respectively.

Table 10: 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)

Figure 15: Predictive Distributions: Aggregated by Sectors, All Sectors

Notes: Subgraph titles are two-digit NAICS codes. Only sectors with more than 10 firms are shown. The black solid (teal dotted) lines are the predictive distributions via the NP-C/R (Homog).

E.3.2 Discussions

Other Setups. (1) Choices of variables: In the main text, y_{it} is chosen to be the log of employment. 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. $R\&D_{it}$ 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.

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.

(2) 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:¹⁵

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

(b) $w_{i,t-1} = [1, \text{rec}_{t-1}]'$. 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.

Results show that for common parameter β , the posterior means are around $0.4 \sim 0.6$ in most cases. For point forecasts, most of the estimators are comparable according to MSE, with only Flat performing poorly in all three setups. For density forecasts, the overall best across all three setups is the Heterosk-NP-C/R predictor in the R&D setup. Comparing across setups, the one with recession produces the worst density forecasts (and worst point forecasts as well), so the recession dummy with heterogeneous effects does not contribute much to forecasting and may even incur overfitting.

β Estimates in the Literature. Compared to the literature, the closest setup 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 (their Table 2) which is close to Homog, and 0.228 with firm fixed effects estimated via OLS (their Table 4) which is to Flat.

Conditional Independence between λ_i and σ_i^2 . First, Figure 16 shows the joint distribution of $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ as well as the joint distribution of $\hat{\sigma}_i^2$ and standardized y_{i0} , the conditioning variable. There does not seem to be much correlation between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ and between $\hat{\sigma}_i^2$ and y_{i0} .

Second, the correlation matrix together with p -values in parentheses (Table 11) delivers a similar message that the unconditional correlations between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ and between $\hat{\sigma}_i^2$ and y_{i0} are roughly insignificant.

Third, to assess conditional correlation, I considered a regression

$$\hat{\sigma}_i^2 = b_0 + b_1 \hat{\lambda}_{i1} + b_2 \hat{\lambda}_{i2} + b_3 y_{i0} + \epsilon_i,$$

where the joint significance of (b_1, b_2) could give us an idea regarding the conditional correlation

¹⁵I do not jointly incorporate recession and R&D because such specification largely restricts the cross-sectional sample size due to the rank requirement.

Figure 16: Joint Distributions: $\hat{\lambda}_i$, $\hat{\sigma}_i^2$, and y_{i0}

Notes: λ_{i1} is the heterogeneous intercept, and λ_{i2} is the heterogeneous coefficient on R&D.

between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ conditioning on y_{i0} . The estimated $\hat{b}_1 = -0.02$ with the 95% interval being $[-0.09, 0.05]$, and $\hat{b}_2 = 0.06$ with the 95% interval being $[-0.07, 0.18]$. Both intervals contain 0. The p -value of the F -test on (b_1, b_2) is 0.92, which is not significant either.

Fourth, to examine conditional independence beyond correlation, I conducted various pairwise conditional independence tests via the R package “bnlearn” (Scutari, 2009). It cannot reject the null hypothesis that $(\hat{\lambda}_{i,\text{level}}, \hat{\sigma}_i^2) \Big| (\hat{\lambda}_{i,\text{RD}}, y_{i0})$, $(\hat{\lambda}_{i,\text{RD}}, \hat{\sigma}_i^2) \Big| (\hat{\lambda}_{i,\text{level}}, y_{i0})$, $(\hat{\lambda}_{i,\text{level}}, \hat{\sigma}_i^2) \Big| y_{i0}$, and $(\hat{\lambda}_{i,\text{RD}}, \hat{\sigma}_i^2) \Big| y_{i0}$ are pairwise conditional independent (the corresponding p -values are all larger than 0.3). Note that all exercises here are in a “sanity check” manner, and an asymptotic theory of tests is beyond the scope of this paper.

Last but not least, I have also explored the alternative predictor with a joint MGLR_x prior on $h_i = (\lambda'_i, l'_i)'$ mentioned in Appendix B.2. However, the density forecasts significantly deteriorate in both the Monte Carlo simulation and the empirical application (see the first row versus the second row in Table 12). One possible explanation could be that if the true DGP exhibits conditional independence between λ_i and σ_i^2 ,¹⁶ then although in principle the alternative predictor could approximate the conditional independence structure asymptotically, it could generate overfitting problems and cause inferior out-of-sample density forecasts in finite samples.

Combining the unconditional and conditional evidence based on posterior means of individual heterogeneity as well as the robustness check on density forecast performance, one would be partially confident about the conditional independence assumption in this young firm sample.

Heterogeneous AR(1) Coefficients. Heterogeneous AR(1) coefficients could be interesting in empirical studies (e.g., Arellano *et al.* (2017) analyzed earnings and consumption dynamics in a nonlinear panel setup). As a robustness check, I have experimented with a version of heterogeneous

¹⁶It is the case in the Monte Carlo simulation and could be the case in the empirical application (see the sanity checks above).

Table 11: Unconditional Correlations: $\hat{\lambda}_i$, $\hat{\sigma}_i^2$, and y_{i0}

	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\sigma}_i^2$	y_{i0}
$\hat{\lambda}_{i1}$	1.00 (1.00)	0.33 (0.00)	-0.08 (0.06)	0.70 (0.00)
$\hat{\lambda}_{i2}$	0.33 (0.00)	1.00 (1.00)	0.02 (0.62)	0.10 (0.02)
$\hat{\sigma}_i^2$	-0.08 (0.06)	0.02 (0.62)	1.00 (1.00)	-0.10 (0.02)
y_{i0}	0.70 (0.00)	0.10 (0.02)	-0.10 (0.02)	1.00 (1.00)

Notes: λ_{i1} is the heterogeneous intercept, and λ_{i2} is the heterogeneous coefficient on R&D. p -values are in parentheses. The entries in bold are significant at the 5% level.

Table 12: Density Forecast Evaluation: Robustness Checks

	Empirical	MonteCarlo
NP-C/R	-195	-1193
cond. correlated (λ_i, σ_i^2)	-506***	-1240***
Heterogeneous β_i	-371**	-1274***

Notes: NP-C/R is the best density predictor in Table 6, which features homogeneous β and conditional independence between λ_i and σ_i^2 . The tests are conducted with respect to NP-C/R, with significance levels indicated by *: 10%, **: 5%, and ***: 1%. The Monte Carlo part is based on one of the 100 repetitions.

persistence β_i constructed from the best density predictor, NP-C/R, in the empirical application. Unfortunately, its density forecast is significantly worse than the forecast from the specification with homogeneous β (see entries (1,1) versus (3,1) in Table 12).

To investigate why this is the case, I turned to the Monte Carlo simulation. In the general model (with normal v_{it}) where the true DGP features homogeneous β , if we fit a model with heterogeneous persistence, the range of the posterior mean $\hat{\beta}_i$ is fairly dispersed, most of $\hat{\beta}_i$ s appear to be smaller than the true value (this usually happens in time series settings where both the persistence and the initial condition are positive and the sample size is relatively small), and density forecast performance deteriorates in a similar manner (see entries (1,2) versus (3,2) in Table 12). Therefore, one possible explanation could be that in a relatively small sample, heterogeneous β_i may tend to fit noise and thus hamper out-of-sample density forecast performance.