Density Forecasts in Panel Data Models:

A Semiparametric Bayesian Perspective*

Laura Liu†

September 26, 2021

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 as well as 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 combining information from the whole panel. Theoretically, I prove that in cross-sectional homoskedastic cases, 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, as well as my committee members, Xu Cheng and Francis J. DiTraglia, for much help and guidance throughout this project. I also thank Todd Clark and Christian Hansen (Co-Editors), an anonymous Associate Editor, and two anonymous referees for their constructive comments and suggestions. I further benefited from many helpful discussions with Stéphane Bonhomme, Evan Chan, Benjamin Connault, Hyungsik R. Moon, Alexandre Poirier, and seminar participants at University of Pennsylvania, FRB Philadelphia, Federal Reserve Board, University of Virginia, Microsoft, UC Berkeley, UC San Diego (Rady), Boston University, 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, Tinbergen Institute, University of Toronto, as well as conference participants at the Midwest Econometrics Group, NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics, Bayesian Nonparametrics Conference, Microeconometrics Class of 2017 Conference, Interactions Workshop, First Italian Workshop of Econometrics and Empirical Economics, International Association for Applied Econometrics Annual Conference, and North American Winter Meeting of the Econometric Society. I would also like to acknowledge the Kauffman Foundation and the NORC Data Enclave for providing research 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. It can also 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.*, 2020). This paper constructs individual-specific density forecasts using a dynamic linear panel data model with common and heterogeneous coefficients as well as cross-sectional heteroskedasticity.

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

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

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

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 facilitate good forecasts of $y_{i,T+1}$. 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 combining information from the whole panel. 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. This calls for a flexible modeling of f, and 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 treat it as a prior distribution and combine it with firm-specific data to obtain the firm-specific posterior via Bayes' theorem. In a special case where the common parameters are $\left(\beta, \sigma^2\right) = (0, 1)$, the firm-specific posterior is

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

This firm-specific posterior helps better infer the firm-specific unobserved skill λ_i and better forecast the firm-specific future performance, thanks to the estimated underlying distribution f that integrates the information from the whole panel in an efficient and flexible way. This is only an intuitive explanation of why the skill distribution f is crucial. In the actual implementation, the correlated random effect distribution f, common parameters $\left(\beta, \sigma^2\right)$, and firm-specific skill λ_i are all inferred simultaneously.

It is natural to construct density forecasts based on the firm-specific posterior. In general, forecasting can be done in a point, interval, or density manner, with density forecasts giving the richest insight into 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 $u_{i,T+1}$, unobserved individual heterogeneity λ_i, and estimation uncertainty of common parameters $\left(\beta, \sigma^2\right)$ and of skill distribution f. Moreover, once density forecasts are obtained, one can easily recover point and interval forecasts.

The contributions of this paper are threefold. First, I establish the theoretical properties of the proposed predictor when the cross-sectional dimension N tends to infinity. To begin, I provide conditions for identifying the common parameters and the distribution of the individual heterogeneity in both cross-sectional homoskedastic and heteroskedastic models. Then, I prove that the proposed estimator achieves posterior consistency in cross-sectional homoskedastic cases. Compared with previous literature on posterior consistency in density estimation problems, 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) strictly exogenous and predetermined variables (including lagged dependent variables) as covariates, and (4) 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 alternative predictors with various parametric priors on f, evaluated by the log predictive score. An application to young firm dynamics also shows that the proposed predictor provides more accurate density predictions. The better forecasting performance is largely due to three key features (in order of importance): the nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients. The estimated model also helps shed light on the latent heterogeneity structure of firm-specific coefficients and cross-sectional heteroskedasticity, as well as whether and how the unobserved heterogeneity depends on the initial condition of the firms.

Moreover, the proposed method is applicable beyond forecasting. Here estimating heterogeneous parameters is important because we want to generate good individual-specific 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.* (2020) and Gu and Koenker (2017a,b). Liu *et al.* (2020) 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.* (2020) 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. adopting a prior on the λ_i distribution and updating the prior belief by the observed data), whereas they employ an empirical Bayes approach (i.e. choosing 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 approach. In addition, this paper features correlated random coefficients allowing the cross-sectional heterogeneity to interact with the initial conditions, whereas Gu and Koenker (2017b) focus 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 on 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 Fisher and Jensen (2021) 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. Fisher and Jensen (2021) assume random effects instead of correlated random effects. Burda and Harding (2013) and Rossi (2014) use a panel probit model and a panel 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. 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 posterior consistency for a multivariate conditional density estimator featuring infinite location-scale mixtures with a probit stick-breaking process.

To account for deconvolution, I construct an inversion inequality that links the convergence of the distribution of observables to the convergence of the distribution of the unobserved individual heterogeneity. The latter is in the Wasserstein metric, which is useful in handling deconvolution problems as found in the recent literature. For example, Nguyen (2013) considers the unobserved distribution on a discrete support, and Su *et al.* (2020) flexibly model a symmetric unimodal unobserved distribution using a mixture of symmetric uniforms where the bounds are drawn from a Dirichlet process location-mixture of Gammas. Their setups, however, differ from the current framework, which calls for a new inversion inequality developed in this paper. Then, I further take into account the dynamic panel data structure, as well as obtain the convergence of the proposed predictor to the oracle predictor.

Last but not least, the empirical application in this paper also relates to the young firm dynamics literature. Akcigit and Kerr (2018) document that R&D intensive firms grow faster, especially 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

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

where $i = 1, \cdots, N$, and $t = 1, \cdots, T + h$. 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 0 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, e.g. 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 $x_{i,t-1} = \left[x_{i,t-1}^{O\prime}, x_{i,t-1}^{P\prime}\right]'$, 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 $x_{i,t-1} = \left[x_{i,t-1}^{O\prime}, x_{i,t-1}^{P*\prime}, y_{i,t-1}\right]'$ with $\beta = \left[\beta^{O\prime}, \beta^{P*\prime}, \rho\right]'$. 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 $c_{i,t-1} = \left(x_{i,0:t-1}^{P}, x_{i,0:T}^{O}, w_{i,0:T}\right)$. We further define $D = \left(\{D_i\}_{i=1}^N \right)$, 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, $\vartheta = \beta,\ h_i = \left(\lambda_i, \sigma_i^2\right)$ 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 types of assumptions on this distribution. One is the random coefficients model, where the individual effects h_i are independent of the conditioning variables $c_{i0} = \left(x_{i0}^P, x_{i,0:T}^O, w_{i,0:T} \right)$. The other is the correlated random coefficients model, where h_i and c_{i0} could be correlated. 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 more feasible to only take into account a subset of c_{i0} or a function of c_{i0} that is relevant for the specific study.

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,

$$f_{i,T+1}^{cond} \left(y \,|\vartheta, f, D_i \right) = \int \underbrace{p\left(y |\, h_i, \vartheta, w_{iT}, x_{iT} \right)}_{\text{future shock}} \cdot \underbrace{p\left(h_i \,|\vartheta, f, D_i \right)}_{\text{individual heterogeneity}} \; dh_i, \tag{3}$$

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 $p\left(y |\, h_i, \vartheta, w_{iT}, x_{iT} \right)$ captures individual i's uncertainty due to the future shock $u_{i,T+1}$. The second term

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

is the individual-specific posterior. It characterizes individual i's uncertainty due to unobserved individual heterogeneity that arises from insufficient time-series information to infer individual h_i. The common distribution f helps regulate 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 (3),

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

In practice, (ϑ, f) are unknown and need to be estimated, thus introducing another source of uncertainty. For the common parameters ϑ, I adopt a conjugate prior (e.g. mulitvariate 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),

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

The conditional predictor reflects uncertainties due to future shocks and unobserved individual heterogeneity, whereas the posterior of (ϑ, f) captures estimation uncertainty. Note that the inference of (ϑ, f) combines information from the whole panel. Once conditioned on (ϑ, f), we have that individuals' outcomes are independent across i and that 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 formulate the nonparametric Bayesian prior using mixture models, 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 $l_i = \log \frac{\bar{\sigma}^2\left(\sigma_i^2 - \underline{\sigma}^2\right)}{\bar{\sigma}^2 - \sigma_i^2}$, where $\underline{\sigma}^2$ ($\bar{\sigma}^2$) is some small (large) positive number. This transformation ensures that the support of f_{σ^2} is bounded by $\left[\underline{\sigma}^2, \bar{\sigma}^2\right]$ for numerical stability, whereas the 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 conditioning on c_{i0}, so their mixture structures can be modeled separately. For a concise exposition, I define a generic variable z that can represent either λ or l, and include z in the subscript as an indicator. When there is no confusion, z and i in the subscript are suppressed.

2.3.1 Random Coefficients Model

In the random coefficients model, the individual heterogeneity $z_i\,(=\lambda_i$ or $l_i)$ is assumed to be independent of the conditioning variables c_{i0}, so the inference of the f part can be considered as an unconditional density estimation problem, and then the DPM prior is a typical choice in the nonparametric Bayesian literature. 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,

$$z_i \sim \sum_{k=1}^{\infty} p_k N\left(\mu_k, \Omega_k\right). \tag{4}$$

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. This is why the DPM prior is flexible enough to approximate a wide range of continous 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.

$$(\mu_k, \Omega_k) \sim G_0, \text{ and } p_k \sim \zeta_k \prod_{j<k} (1-\zeta_j), \text{ where } \zeta_k \sim \mathrm{Beta}\left(1, \alpha\right). \tag{5}$$

The scale parameter α controls the number of unique components in the mixture density and thus determines the flexibility of the mixture density. One advantage 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 $\alpha \sim \mathrm{Ga}\left(a_{\alpha,0}, b_{\alpha,0}\right)$, 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 $z_i\,(=\lambda_i$ 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 ($\mathrm{MGLR_x}$) prior, a generalization of the Gaussian-mixture prior for conditional density estimation, and extend it to the multivariate setup. Conditioning on c_{i0},

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

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 matricvariate-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

$$p_k\left(c_{i0}\right) = \Phi\left(\zeta_k\left(c_{i0}\right)\right) \prod_{j<k}\left(1 - \Phi\left(\zeta_j\left(c_{i0}\right)\right)\right), \tag{6}$$

where stochastic function ζ_k is drawn from Gaussian process $\zeta_k \sim GP\left(0, V_k\right)$ for $k = 1, 2, \cdots$. 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 further dependency of component means and covariances on c_{i0} beyond the MGLR_x specification (see details in the Appendix).

3 Theoretical Properties

In general, it is desirable to ensure that the prior belief does not dominate the posterior inference asymptotically. For Bayesians with 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 can be viewed 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 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 result.

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 coefficients model with cross-sectional heteroskedasticity, where random coefficients and cross-sectional homoskedasticity can be viewed as special cases and will be discussed in Remark 3.

Assumption 1. (Identification: General Model)

1. *Model setup: Consider the panel data model in (2),*
   - (a) $\left(c_{i0}, \lambda_i, \sigma_i^2\right)$ *are i.i.d. across* i.
   - (b) *For all* t, *conditional on* $(y_{it}, c_{i,t-1})$, x_{it}^{P*} *is independent of* $\left(\lambda_i, \sigma_i^2\right)$.
   - (c) $\left(x_{i,0:T}^{O}, w_{i,0:T}\right)$ *are independent of* $\left(\lambda_i, \sigma_i^2\right)$.
   - (d) *Conditioning on* c_{i0}, λ_i *and* σ_i^2 *are independent of each other.*
   - (e) *Let* $u_{it} = \sigma_i v_{it}$. $v_{it} \sim N(0,1)$ *is i.i.d. across* i *and* t *and independent of* $\left(c_{i,t-1}, \lambda_i, \sigma_i^2\right)$.
2. *Identification:*
   - (a) *The characteristic functions of* $\lambda_i|c_{i0}$ *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) *Let* $\tilde{x}_{i,t-1} = x_{i,t-1} - \sum_{s=t+1}^{T} x_{i,s-1} w_{i,s-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}'\right)^{-1} w_{i,t-1}$ *given by orthogonal forward differencing. Then, the matrix* $\mathbb{E}\left[\sum_{t=1}^{T-d_w} \tilde{x}_{i,t-1}\tilde{x}_{i,t-1}'\right]$ *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 standard proof based on characteristic functions 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 obtain the identification of f_{σ^2}.

Remark 3. (1) For random coefficients models, Assumption 1(1-a) is replaced by "$\left(\lambda_i, \sigma_i^2\right)$ are independent of c_{i0} and i.i.d. across i."

(2) Under cross-sectional homoskedasticity, we can delete Assumption 1(1-d), get rid of σ_i^2 in conditions 1-a,b,c and $\sigma_i^2|c_{i0}$ in condition 2-a, and replace condition 1-e by "u_{it} is i.i.d. across i and t and independent of $(c_{i,t-1}, \lambda_i)$."

3.2 Posterior Consistency

Most of the 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 semiparametric setup, and then I specify and verify these conditions in cases of (correlated) random coefficients models.

General Semiparametric Model. Let Θ be the space of the common parameters ϑ, $\mathcal{F}$ be a set of the underlying distributions f with finite second moments, $\Pi(\cdot,\cdot)$ 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

$$g\left(D_i|\,\vartheta, f\right) = \begin{cases} \int p\left(D_i|\,\vartheta, h_i\right) f\left(h_i\right) dh_i, & \text{if } f \text{ is an unconditional dist.,} \\ \int p\left(\,D_i\backslash\, c_{i0}|\,\vartheta, h_i\right) f\left(\,h_i|\,c_{i0}\right) q_0\left(c_{i0}\right) dh_i, & \text{if } f \text{ is a conditional dist.,} \end{cases} \tag{7}$$

where $D_i\backslash\, c_{i0}$ 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 Wasserstein metric on f. Let $\Gamma(f_1, f_2)$ be the collection of all joint measures with marginals f_1 and f_2. We define the second Wasserstein distance, $W_2(f_1, f_2) = \left(\inf_{\gamma\in\Gamma(f_1,f_2)} \int \|h_1 - h_2\|_2^2\, d\gamma(h_1, h_2)\right)^{1/2}$. Note that convergence in the W_2 metric is equivalent to weak convergence plus convergence of the second moment (Santambrogio, 2015).

When f is a conditional distribution, it is helpful to link the properties of the conditional density to the corresponding joint density $f(h, c_0) = f(h|c_0)\, q_0(c_0)$ without explicitly modeling q_0, which circumvents the difficulty associated with an uncountable set of conditional densities (Pati *et al.*, 2013). Note that q_0 is only for theoretical derivation, and there is no need to estimate it in practice.

Theorem 4. (Posterior Consistency: General Semiparametric Model) *Suppose we have:*

1. Individual-specific likelihood:

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

$$\Pi\left((\vartheta, f):\ D_{KL}\left(g\left(D_i|\,\vartheta_0, f_0\right) \,\|\, g\left(D_i|\,\vartheta, f\right)\right) < \epsilon\right) > 0.$$

(b) There exists $\delta_\vartheta > 0$ such that for all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_\vartheta$ 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.

(c) There exists an increasing function $\mathfrak{C}: \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ *with* $\lim_{x\to 0} \mathfrak{C}(x) = 0$ *such that for all* $f \in \mathcal{F}$, $W_2(f, f_0) \leq \mathfrak{C}(\|g(D_i|\vartheta_0, f) - g(D_i|\vartheta_0, f_0)\|_1)$.

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*

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

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* $\beta > 0$, $\Pi_f(\mathcal{F}_N^c) = O(\exp(-\beta N))$.
   - *(b) For some* $\gamma > 0$, $\sum_j \sqrt{\mathcal{N}(\epsilon, \mathcal{F}_{N,j}) \Pi_f(\mathcal{F}_{N,j})} = o\left(\exp\left((1-\gamma) N\epsilon^2\right)\right)$, *where* $\mathcal{N}(\epsilon, \mathcal{F}_{N,j})$ *is the covering number of* $\mathcal{F}_{N,j}$ *by balls with radius* ϵ *in the* L_1*-norm.*

Then, the posterior achieves consistency at (ϑ_0, f_0), *i.e. for all* $\epsilon, \delta > 0$, *as* $N \to \infty$,

$$\Pi\left((\vartheta, f): \|\vartheta - \vartheta_0\|_2 < \delta,\ W_2(f, f_0) < \epsilon \middle| D\right) \to 1,$$

in probability with respect to the true DGP.

Intuitively, let $\Theta_\delta^c = \{\|\vartheta - \vartheta_0\|_2 \geq \delta\}$, $\mathcal{F}_\epsilon^c = \{W_2(f, f_0) \geq \epsilon\}$, 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

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

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 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).

When g is observed instead of f, we need to further address convolution and common parameters. In terms of convolution, it preserves the L_1-norm as well as the number of balls that cover the sieve. Moreover, the inversion inequality in condition 1-c helps identify the underlying f based on the observed g. I use the Wasserstein metric on f because there are technical difficulties in establishing a similar inversion inequality in the L_1-norm, whereas recent literature found that the Wasserstein metric circumvents the issue (Nguyen, 2013; Su *et al.*, 2020). We can extend both condition 1-c and the posterior consistency result to the W_p metric with $p \geq 1$. 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,c 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.* (2020), 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. Here I focus on the cross-sectional homoskedastic case due to the difficulty in constructing a suitable mollifier in the cross-sectional heteroskedastic setup, which is left for future research. Then, the space for common parameters $\vartheta = \left(\beta, \sigma^2\right)$ is $\Theta = \mathbb{R}^{d_x} \times \left[\underline{\sigma}^2,\ \bar{\sigma}^2\right]$. Let $\mathbb{E}_f\left[\mathfrak{g}(\lambda)\right] = \int \mathfrak{g}(\lambda) f(\lambda)\, d\lambda$ for a generic function $\mathfrak{g}(\lambda)$. To ensure condition 1 in Theorem 4, we consider space $\mathcal{F} = \left\{f : \mathbb{E}_f \left\|\lambda\right\|_2^{2(1+\eta)} \leq M\right\}$, for some large $M > 0$, and η is defined in Assumption 6(1-e) below.

Assumption 5. (Covariates)

1. $w_{i,0:T-1}$ *is bounded.*
2. *The eigenvalues of* $\sum_t w_{i,t-1} w_{i,t-1}'$ *are no less than some small* $m_w > 0$.
3. $x_{i,0:T-1}^{O}$, $x_{i,0:T-1}^{P*}$, *and* y_{i0} *have finite* $4\left(1+\eta'\right)$*-th moments with* $\eta' > 0$.

The conditions on $w_{i,0:T-1}$ help obtain an upper bound on the W_2-distance between f and its convolution with a mollifier and hence ensure Theorem 4(1-c). Both conditions can be relaxed to "almost everywhere" with slight adjustments in the proofs. The moment conditions on $x_{i,t-1}$ ensure that the GMM estimates of the common parameters are asymptotically normal, so the exponentially consistent sequence of tests in Theorem 4(2) can be constructed accordingly. All three conditions also prevent a slight difference in β from obscuring the difference in f, and are essential to Theorem 4(1-a,b).

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

1. *True distribution* f_0*:*
   (a) $f_0(\lambda)$ *is a continuous density.*
   (b) *For some* $M_\lambda > 0$, $0 < f_0(\lambda) \leq M_\lambda$ *for all* λ.
   (c) $\mathbb{E}_{f_0}\left[\log f_0(\lambda)\right] < \infty$.
   (d) $\mathbb{E}_{f_0}\left[\log \frac{f_0(\lambda)}{\varphi_\delta(\lambda)}\right] < \infty$, *where* $\varphi_\delta(\lambda) = \inf_{\left\|\lambda' - \lambda\right\|_2 < \delta} f_0(\lambda')$, *for some* $\delta > 0$.

(e) *For some* $\eta > 0$, $\mathbb{E}_{f_0}\left[\int \|\lambda\|_2^{2(1+\eta)}\right] < \infty$.

2. *The base distribution of the DPM prior* (G_0) *follows a multivariate-normal-inverse-Wishart distribution, where the degree of freedom of the inverse Wishart component* $\nu_0 > \max\left(2d_w, (2d_w+1)(d_w-1)\right)$.

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. Second, according to Corollary 1 in Canale and De Blasi (2017), condition 2 further ensures the sieve property (Theorem 4(3)), where $2d_w$ controls the tail behavior of component mean μ and $(2d_w+1)(d_w-1)$ regulates the eigenvalue structure of component variance Ω.

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

1. *Model: Remark 3 for random coefficients models with cross-sectional homoskedasticity.*
2. *Covariates:* $(x_{i,0:T}, w_{i,0:T})$ *satisfies Assumption 5.*
3. *Common parameters:*
   (a) ϑ_0 *is in the interior of supp* (Π_ϑ).
   (b) *The domain of* σ^2 *is bounded by* $\left[\underline{\sigma}^2,\ \bar{\sigma}^2\right]$ *for some* $\underline{\sigma}^2, \bar{\sigma}^2 > 0$.
4. *Distributions of individual heterogeneity:* f_0 *and* Π_f *satisfy Assumption 6.*

Then, the posterior achieves consistency 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. Let $\mathcal{C}$ be the support of the conditioning variables, and $\mathcal{F}^*$ be a subset of conditional distributions such that mapping $c_0 \mapsto f(\cdot|c_0)$ is a continous function from $\mathcal{C}$ to the space of Lebesgue integrable functions on $\mathbb{R}^{d_w}$. Similar to the above discussion on random coefficients models, I focus on the cross-sectional homoskedastic case and consider space $\mathcal{F} = \left\{f : \left\{\mathbb{E}_{f,q_0}\|\lambda\|_2^{2(1+\eta)} \le M\right\} \cap \mathcal{F}^*\right\}$, where $\mathbb{E}_{f,q_0}\left[\mathfrak{g}(\lambda, c_0)\right] = \int \mathfrak{g}(\lambda, c_0) f(\lambda|c_0) q_0(c_0)\, d\lambda dc_0$ for a generic function $\mathfrak{g}(\lambda, c_0)$. M is some large positive constant, and η is defined in Assumption 9(1-e) below.

Assumption 8. (Conditioning set) $\mathcal{C}$ *is compact, and* $q_0(c_0) > 0$ *for all* $c_0 \in \mathcal{C}$.

The compactness ensures uniform convergence on $\mathcal{C}$ in the proof of the KL property. It is stronger than the $\mathcal{C}$ part in Assumption 5(1,3) for random coefficients models.

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

1. *True distribution* f_0:
   (a) $f_0(\cdot|\cdot)$ *is jointly continuous in* (λ, c_0).
   (b) *For some* $M_\lambda > 0$, $0 < f_0(\lambda|c_0) \le M_\lambda$ *for all* (λ, c_0).
   (c) $\mathbb{E}_{f_0,q_0}\left[\log f_0(\lambda|c_0)\right] < \infty$.
   (d) $\mathbb{E}_{f_0,q_0}\left[\log \frac{f_0(\lambda|c_0)}{\varphi_\delta(\lambda|c_0)}\right] < \infty$, *where* $\varphi_\delta(\lambda|c_0) = \inf_{\|\lambda'-\lambda\|_2<\delta} f_0(\lambda'|c_0)$, *for some* $\delta > 0$.

(e) *For some* $\eta > 0$, $\mathbb{E}_{f_0,q_0}\left[\int \|\lambda\|_2^{2(1+\eta)}\right] < \infty$.

2. *The base distribution of the* $MGLR_x$ *prior* (G_0) *is characterized by* a *multivariate normal distribution on* $vec\,(\mu)$ *and an inverse Wishart distribution on* Ω, *where the degree of freedom of the inverse Wishart component* $\nu_0 > \max\left(2d_w, (2d_w+1)(d_w-1)\right)$.
3. *Stick-breaking process: The covariance function for Gaussian process can be specified as* $V_k(c,\tilde{c}) = \tau \exp\left(-A_k \|c-\tilde{c}\|_2^2\right)$, *where* $\tau > 0$ *is a fixed number.*
   (a) *The prior for* A_k *has full support on* $\mathbb{R}^+$.
   (b) *There exist* β, $\gamma > 0$ *and a sequence* $\delta_N = O\left(N^{-5/2}(\log N)^2\right)$ *such that* $\mathbb{P}(A_k > \delta_N) \leq \exp\left(-N^{-\beta} k^{(\beta+2)/\gamma} \log k\right)$.
   (c) *For the same* γ *as in condition 3-b, there exists an increasing sequence* $r_N \to \infty$ *and* $(r_N)^{d_{c0}} = o\left(N^{1-\gamma}(\log N)^{-\left(d_{c_0}+1\right)}\right)$ *such that* $\mathbb{P}(A_k > r_N) \leq \exp(-N)$.

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 and G_0 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 λ on c_0. Moreover, the tail of A_k cannot be too fat according to condition 3-c.

Theorem 10. (Posterior Consistency: Correlated Random Coefficients) *Suppose we have:*
1. *Model: Remark 3(2) for cross-sectional homoskedastic models.*
2. *Covariates:* $(x_{i,0:T}, w_{i,0:T})$ *satisfy Assumptions 5(2,3) and 8.*
3. *Common parameters: Theorem 7(3).*
4. *Distributions of individual heterogeneity:* f_0 *and* Π_f *satisfy Assumption 9.*

Then, the posterior achieves consistency 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 asymptotic convergence of the density forecasts to the oracle forecast. Theorem 21 in the Appendix established the convergence result in the general semiparametric setup, and the following theorem focuses on the (correlated) random coefficients models considered in the paper.

Theorem 11. (Density Forecasts: (Correlated) Random Coefficients with Cross-sectional Homoskedasticity) *Given conditions in Theorem 7 for random coefficients models (or conditions in Theorems 10 and continuity of* $q_0(c_0)$ *for correlated random coefficients models), density forecasts converge to* *the oracle for all i with $\mathbb{E}_{f_0}\left[\|\lambda\|_2^2 \middle| c_{i0}\right] < \infty$, i.e. given i, for all $\epsilon > 0$, as $N \to \infty$,*

$$\mathbb{P}\left(W_2\left(f_{i,T+1}^{cond}, f_{i,T+1}^{oracle}\right) < \epsilon \middle| D\right) \to 1,$$

in probability with respect to the true DGP.

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

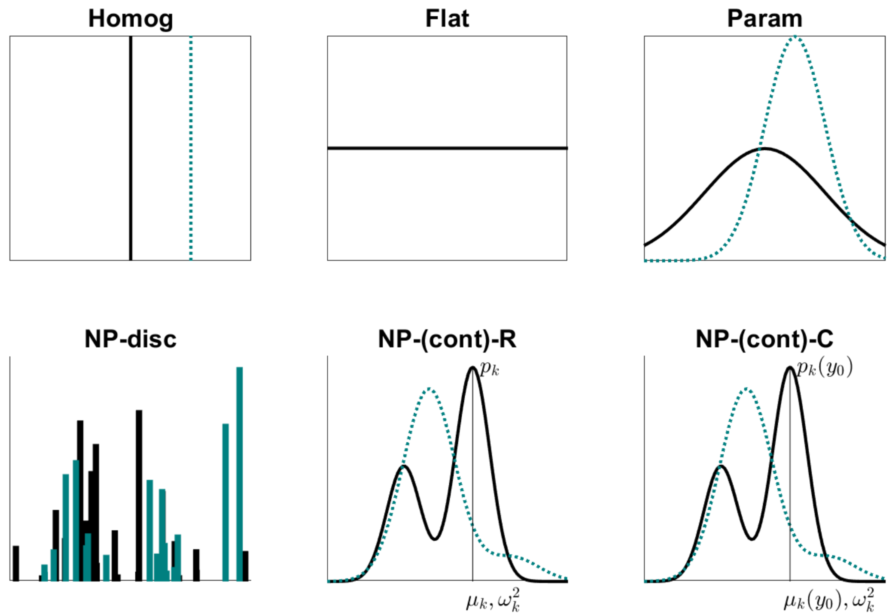
The accuracy of the density forecasts is measured by the log predictive score (LPS) as suggested in Geweke and Amisano (2010), $LPS = \frac{1}{N}\sum_i \log \hat{p}\left(y_{i,T+1}|D\right)$, where $y_{i,T+1}$ is the realization at $T+1$, and $\hat{p}\left(y_{i,T+1}|D\right)$ represents the predictive likelihood with respect to the estimated model conditional on the observed data D. $\exp\left(LPS_A - LPS_B\right)$ gives the odds of future realizations based on predictor A versus predictor B. I performed a test combining Amisano and Giacomini (2007) (for the LPS) and Qu *et al.* (2020) (for panel data, see their Section 2.6 on general loss functions) 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 have the same skill level 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 $\left([\beta, \lambda^*]', \sigma^2\right)$.

The flat prior (Flat) is specified as $f\left(\lambda_i\right) \propto 1$, an uninformative prior with the posterior mode being the MLE estimate. 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) combines cross-sectional information via a parametric distribution, such as a Gaussian distribution with unknown mean and variance, $\lambda_i \sim N\left(\mu, \omega^2\right)$. A normal-inverse-gamma hyperprior is further adopted for $\left(\mu, \omega^2\right)$. The parametric prior can be viewed as a limit case of the DPM prior when the scale parameter $\alpha \to 0$, so there is only one component, and

Figure 1: Alternative Predictors



Notes: For easier illustration, here I consider the baseline model with univariate λ_i and homoskedasticity. The black solid and teal dotted lines represent two draws from each prior (except NP-disc, where the teal one is also solid). Homog: Because λ^* is unknown *ex ante*, 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 the DP and how it relates to the DPM.

(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 nonparametric distribution 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, comparing with NP-disc 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* whether the parsimonious parametric predictors or the flexible semiparametric ones would perform better. Therefore, it is worthwhile to implement the Monte Carlo experiments and assess which predictor produces more accurate forecasts under which circumstances.

Table 1: Simulation Setup: Baseline Model with Random Effects

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it},\ u_{it} \sim N\left(0, \sigma^2\right)$
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	$\lambda_i \sim \frac{1}{9}N\left(2, \frac{1}{2}\right) + \frac{8}{9}N\left(-\frac{1}{4}, \frac{1}{2}\right)$, so $\mathbb{V}(\lambda_i) = 1$
Bimodal	$\lambda_i \sim (0.35N(0,1) + 0.65N(10,1))/\sqrt{1 + 10^2 \cdot 0.35 \cdot 0.65}$, so $\mathbb{V}(\lambda_i) = 1$

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 satisfies the moment condition in Assumption 5(3). 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. The first experiment features a degenerate λ_i distribution, where all firms have 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 third experiment incorporates 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 and discard the first 20,000 as burn-in. Based on graphical and statistical tests, the MCMC draws converge to a stationary distribution (see Appendix).

Table 2 shows the forecasting comparison across 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. Flat is considerably worse. When the λ_i distribution is non-degenerate, there is a substantial gain from employing NP-R. In the bimodal case, NP-R far exceeds all alternatives. In the skewed case, Flat and Param are second best, yet still significantly inferior to NP-R. Homog and NP-disc yield the poorest forecasts, which suggests that their discrete supports may not be able to approximate the continuous λ_i distribution in this case—even the nonparametric DP prior with countably infinite support may still be far from enough.

To investigate why we obtain better forecasts, Figure 2 plots the posterior distribution of the λ_i distribution for experiments Skewed and Bimodal. In the skewed case, NP-R better tracks the peak on the left and the tail on the right. In the bimodal case, NP-R nicely captures the M-shape. Therefore, the nonparametric prior flexibly approximates a vast set of distributions, which

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

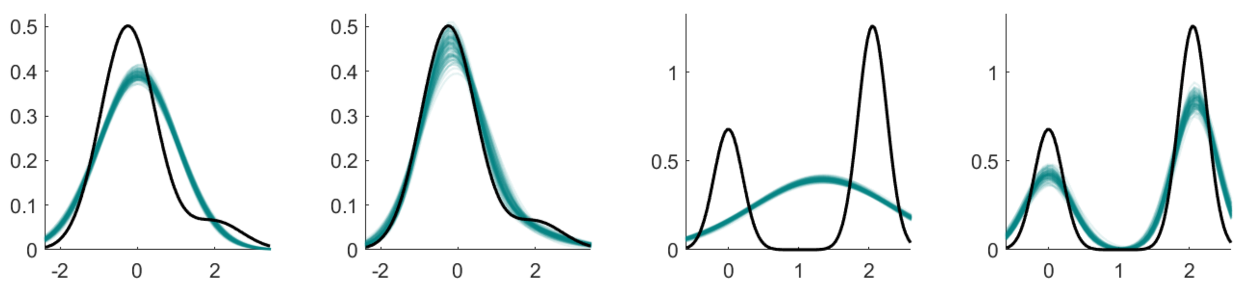
	Degenerate	Skewed	Bimodal
Oracle	*-725*	*-798*	*-766*
Homog	**-0.2*****	-193***	-424***
Flat	-102***	-7***	-38***
Param	-4	-1***	-34***
NP-disc	**-0.2*****	-206***	-40***
NP-R	-4	**-0.3**	**-6**

Notes: The density forecasts are assessed by the LPS and a test combining Amisano and Giacomini (2007) and Qu *et al.* (2020). 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.

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

(a) Skewed (b) Bimodal

Param NP-R Param NP-R



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, $\Pi_f\left(f\,|y_{1:N,0:T}\right)$.

provides more precise estimates of the underlying λ_i distributions and consequently more accurate density forecasts. This connection between distribution estimation and density forecasts reflects the theoretical results in Theorem 11.

4.3 General Model

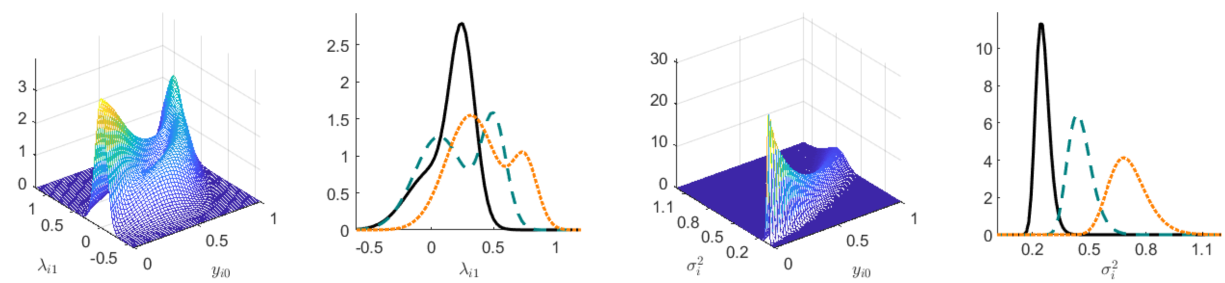
The general model accounts for three key features: multidimensional individual heterogeneity, cross-sectional heteroskedasticity, and correlated random coefficients. The exact specification is characterized and depicted in Table 3.

In terms of multidimensional individual heterogeneity, λ_i is now a 3-by-1 vector, and the corresponding covariates are composed of the intercept, time-specific $w_{t-1}^{(2)}$, and individual-time-specific $w_{i,t-1}^{(3)}$. In terms of correlated random coefficients, I adopt the conditional distribution following Dunson and Park (2008) and Norets and Pelenis (2014). They regard it as a challenging problem because this conditional distribution exhibits rapid changes in its shape, which considerably restricts

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

(a) $\lambda_{i1} \,|\, y_{i0}$ (b) $\sigma_i^2 \,|\, y_{i0}$



Notes: In 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 fourth panels, the black solid / teal dashed / 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. $f(\lambda_{i1} | y_{i0}) = f(\lambda_{i1} | y_{i0})\, q_0(y_{i0}) = f(\lambda_{i1}, y_{i0})$.

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 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 large but bounded support above zero, 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. In the latter case, the normal likelihood function is misspecified. The v_{it} distributions are standardized, i.e. $\mathbb{E}(v_{it}) = 0$ and $\mathbb{V}(v_{it}) = 1$, so we can identify σ_i^2.

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 Homosk-NP-C to examine whether it is practically relevant to model heteroskedasticity. The third column of Table 4 assesses the forecasting performance under correct

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

		f_λ	f_{σ^2} (or f_l)	Normal v_{it}	Skewed v_{it}
Oracle		*Known*	*Known*	*-974*	*-965*
Homog		$= \delta_{\lambda^*}$	$f_{\sigma^2} = \delta_{\sigma^{2*}}$	-407***	-417***
Homosk	NP-C	$\sim$ MGLR$_\text{x}$	$f_{\sigma^2} = \delta_{\sigma^{2*}}$	-134***	-146***
Heterosk	Flat	$\propto 1$	$f_{\sigma^2} \propto 1$	-384***	-366***
	Param	= Normal	f_{σ^2} = IG	-79***	-78***
	NP-disc	$\sim$ DP	$f_l \sim$ DP	-79***	-78***
	NP-R	$\sim$ DPM	$f_l \sim$ DPM	-229***	-224***
	NP-C	$\sim$ MGLR$_\text{x}$	$f_l \sim$ MGLR$_\text{x}$	**-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.

specification. Heterosk-NP-C is the most accurate density predictor. There are several messages if we compare density forecast performance across 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. Although there is no theoretical guarantee, the proposed semiparametric method could still be helpful in 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 (Robb and Seamans, 2014; Akcigit and Kerr, 2018). 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 young firms. See Robb *et al.* (2009) for further description of the survey design.

Table 5: Descriptive Statistics of Observables

	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

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 Zarutskie and Yang (2015) and Akcigit and Kerr (2018), 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 Great Recession may partly account for the jobless recovery afterward. Below, I focus on the following model specification,

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

where $\text{R\&D}_{it}$ is given by the ratio of a firm's R&D employment over its total employment. Other setups are discussed in the Appendix. An extension to a panel Tobit model as in Liu *et al.* (2019) could help accommodate firms' endogenous exit choice, which is left for future exploration.

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 (#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. The descriptive statistics for $\log \text{emp}_{it}$ and $\text{R\&D}_{it}$ are summarized in Table 5, and the corresponding densities are plotted in Figure 12 in the Appendix. Both distributions are right skewed and may be multimodal, 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}. Then, I adopt an MGLR_x prior on f_λ and a DPM prior on f_l for Heterosk-NP-C/R. The conditioning variable y_{i0} is further standardized, which ensures numerical stability as the conditioning variables enter exponentially into the covariance function of the Gaussian process.

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

		β		LPS*N
		Mean	SD	
Heterosk	*NP-C/R*	*0.50*	*0.02*	***-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*

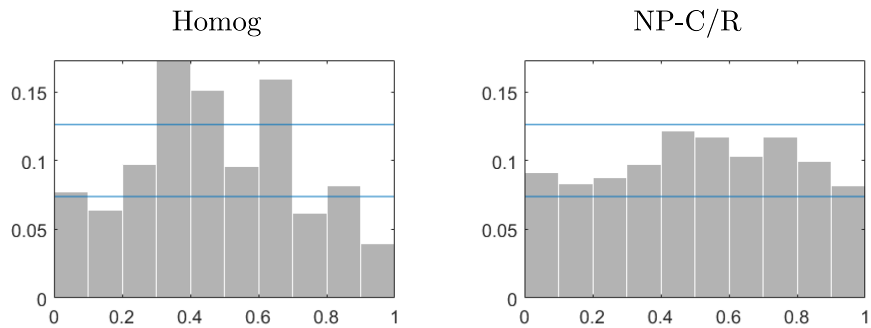
Notes: See the description of Table 2 for density forecast evaluation. Here Heterosk-NP-C/R is the benchmark for both normalization and significance tests. For Heterosk-NP-C/R, the table reports the exact values of LPS $\cdot\, N$. For other predictors, the table reports their differences from Heterosk-NP-C/R.

The first two columns in Table 6 characterize the posterior estimates of the common parameter β. In most 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 the persistence is not 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 may not be able to capture all individual effects, so these estimators may attribute the remaining individual effects to the 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 effect of the initial condition and misinterpret it as the persistence. In all scenarios, the posterior standard deviations are relatively small.

The last column in Table 6 compares density forecasting performance. The overall best is Heterosk-NP-C/R. 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 has these features. Intuitively, the odds, given by the exponential of the difference in the 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 the LPS characterizes the relative ranks of predictors, the PIT complements the 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 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

Figure 3: PIT



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

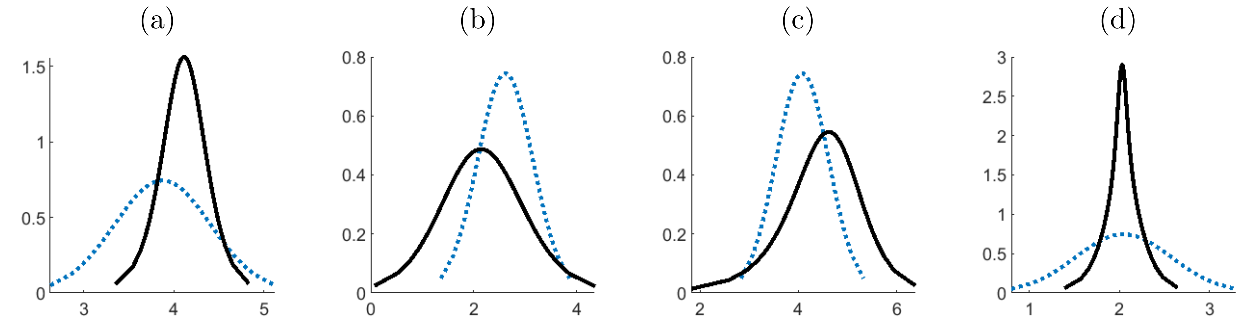
band, while other predictors yield apparent inverse-U shapes. The reason might be that the other predictors do not take correlated random coefficients into account but instead attribute their effects to the shock variance, which leads to more diffused predictive distributions.

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

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

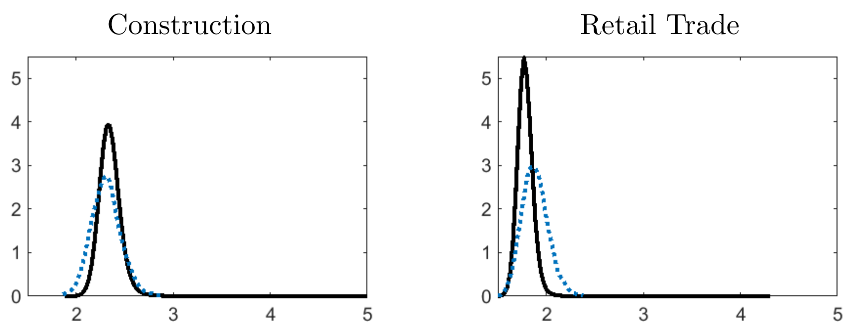
The latent heterogeneity structure is presented in Figure 6, which plots the joint distributions of the estimated individual effects and the conditional variable. For example, the pairwise relationship between λ_{i1} and the standardized y_{i0} is nonlinear and exhibits multiple components, which reassures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

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



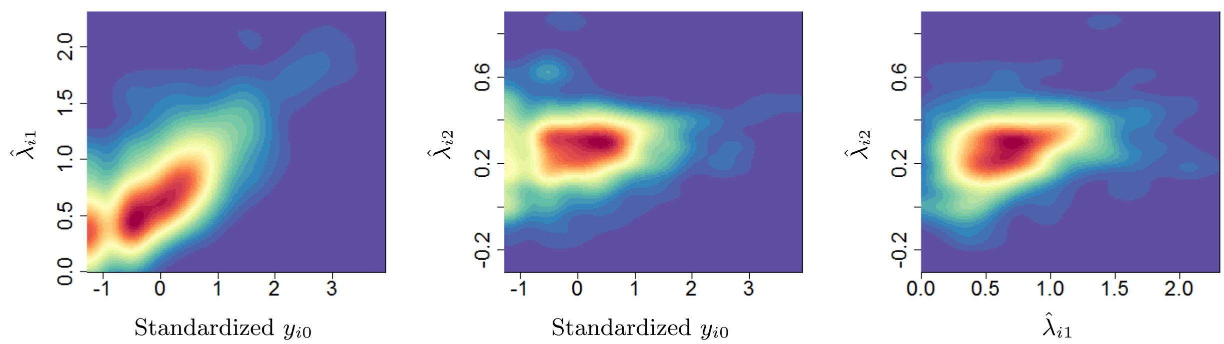
Notes: The black solid (teal dotted) lines are the predictive distributions via the NP-C/R (Homog). See Appendix for predictive distributions of all sectors.

pict pairwise joint distributions involving $\hat{\sigma}_i^2$ 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 combines information from the whole panel 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

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.

the oracle in cross-sectional homoskedastic cases. 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 Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, **9** (2), 291–312.

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

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

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

—, 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.

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.

Fisher, M. and Jensen, M. J. (2021). Bayesian nonparametric learning of how skill is distributed across the mutual fund industry. *Journal of Econometrics*, forthcoming.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pp. 1386–1403.

— (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, **36** (2), 454–456.

Galambos, J. and Simonelli, I. (2004). *Products of Random Variables: Applications to Problems of Physics and to Arithmetical Functions*. Marcel Dekker.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, **26** (2), 216–230.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R. *et al.* (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, **27** (1), 143–158.

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

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

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

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.

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. (2019). Forecasting with a panel tobit model. *NBER Working Papers* 26569.

—, — and — (2020). Forecasting with dynamic panel data models. *Econometrica*, **88** (1), 171–201.

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.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, **41** (1), 370–400.

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.

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.

Qu, R., Timmermann, A. and Zhu, Y. (2020). Comparing forecasting performance in cross-sections. *Journal of Econometrics*, forthcoming.

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.

SANTAMBROGIO, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, **55** (58-63), 94.

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.

SU, Y., BHATTACHARYA, A., ZHANG, Y., CHATTERJEE, N. and CARROLL, R. J. (2020). Nonparametric Bayesian deconvolution of a symmetric unimodal density. *arXiv preprint arXiv:2002.07255*.

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.

VILLANI, C. (2009). *Optimal Transport: Old and New*, vol. 338. Springer.

YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G. O. and HOLMES, C. (2011). Bayesian nonparametric 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"

Laura Liu

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 $\mathrm{Ga}(a,b)$ with pdf being $f_{\mathrm{Ga}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$. The according **inverse gamma distribution** is given by $\mathrm{IG}(a,b)$ with pdf being $f_{\mathrm{IG}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{-a-1}e^{-b/x}$. The $\Gamma(\cdot)$ in the denominators is the gamma function.

The **inverse Wishart distribution** is a generalization of the inverse gamma distribution to multi-dimensional setups. Let Ω be a $d\times d$ positive definite matrix following an inverse Wishart distribution $\mathrm{IW}(\Psi,\nu)$, then its pdf is $f_{\mathrm{IW}}(\Omega;\Psi,\nu)=\frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu d}{2}}\Gamma_d(\frac{\nu}{2})}|\Omega|^{-\frac{\nu+d+1}{2}}e^{-\frac{1}{2}tr(\boldsymbol{\Psi}\Omega^{-1})}$. 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 $\{c_1,c_2,\cdots,c_n\}$, $[\zeta(c_1),\zeta(c_2),\cdots,\zeta(c_n)]'$ has a joint Gaussian distribution with the mean vector being $[m(c_1),m(c_2),\cdots,m(c_n)]'$ and the i,j-th entry of the covariance matrix being $V(c_i,c_j)$, $i,j=1,\cdots,N$.

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

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

In the **panel data** setup, for a generic variable z, which can be v, w , x, or y, z_{it} is a $d_z\times 1$ vector, and $z_{i,t_1:t_2}=(z_{it_1},\cdots,z_{it_2})$ is a $d_z\times(t_2-t_1+1)$ matrix.

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

$D_{KL}(f_0\parallel f)=\int f_0\log\frac{f_0}{f}$ is the **KL divergence** of f from f_0.

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

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

$X \lesssim Y$ (or $X \gtrsim Y$) is an abbreviated form of inequality $X \leq CY$ (or $X \geq CY$) for some $C > 0$.

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, the proposed predictor would provide a sizable improvement as long as the time series for individual i is not informative enough to precisely recover its individual effects.

Dirichlet Process (DP). The DP is another candidate as a nonparametric prior, which is on a discrete support 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 $(A_1, \cdots, A_K)$,

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

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

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

which is a multivariate generalization of the Beta distribution.

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

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

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. Then, the DP scale parameter α controls the number of unique components in the mixture density and thus the flexibility of the mixture density. Let K^* denote the number of unique components. As derived in Antoniak (1974), we have

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

By definition, a draw from the 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 $z_i\,(=\lambda_i$ or $l_i)$ follows a continuous parametric distribution $f\left(z;\theta\right)$ where θ are the parameters, and adopt a DP prior for the distribution of θ. Then, the parameters θ are discrete while the individual heterogeneity z enjoys a continuous distribution. This additional layer of mixture leads to the DPM model.

Intuition: MGLR$_\mathbf{x}$ Prior. Here we give some intuition why the MGLR$_\text{x}$ prior is general enough to accommodate a broad class of conditional distributions.

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

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

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

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

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

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

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

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

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

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

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

In summary, the current setup is similar to approximating the conditional density via Bayes' theorem, but does not explicitly model the distribution of the conditioning variable c_0, and thus circumvents the difficulty associated with an uncountable set of conditional densities (Pati *et al.*, 2013).

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 similar manner. Let $\mathcal{T}_i = \{s_{i1}, s_{i2}, \cdots, s_{iT_i}\}$ be 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 $y_{i,T+1}$. 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 existing $y_{i,T+1}$.

(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 study). This assumption could also be relaxed depending on the model setup. For example,

in the baseline model, it may sometimes be reasonable to let $c_{i0} = y_{i,s_{i1}-1}$.

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

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

$$c_{i,t-1} = \left(c_{i0}, x^P_{i,\mathcal{S}_{i\tau}-1}, x^O_{i,\mathcal{T}_i-1}, w_{i,\mathcal{T}_i-1}\right),$$

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

Assumption 12. *(Identification: Unbalanced Panels) For all i,*

1. *c_{i0} is observed.*
2. *x_{iT} and w_{iT} are observed.*
3. *For all i, $w_{i,\mathcal{T}_i}$ has full rank d_w almost everywhere.*
4. *For $t = s_{i\tau}$, $\tau = 1, \cdots, T_i - d_w - 1$, let*

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

given by orthogonal forward differencing. 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 12.

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 $\left(\lambda_i, \sigma_i^2\right)$ 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 "$\left(\lambda_i, \sigma_i^2\right)$ are independent of c_{i0} and i.i.d. across i." Together with Assumption 1(1-d), it implies that $\left(\lambda_i, \sigma_i^2, c_{i0}\right)$ are mutually independent.

In principle, we could relax the conditional independence between λ_i and σ_i^2 and still achieve identification 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 $\mathrm{MGLR_x}$ prior on the vector of individual heterogeneity $h_i = (\lambda_i', l_i)'$, which combines the individual-specific coefficients λ_i and the transformed cross-sectional heteroskedasticity $l_i = \log \frac{\bar{\sigma}^2(\sigma_i^2 - \underline{\sigma}^2)}{\bar{\sigma}^2 - \sigma_i^2}$. Despite the possibility of this extension, I keep the conditional independence assumption in this paper considering that Appendix E.3 provides 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 $\left(\lambda_i, \sigma_i^2\right)$, (2) the distributions of v_{it}, $f_{v,t}\left(v_{it}|c_{i,t-1}\right)$, have known functional forms, such that $\mathbb{E}[v_{it}|c_{i,t-1}] = 0$, $\mathbb{V}[v_{it}|c_{i,t-1}] = 1$, and (3) the characteristic function of $v_{it}|c_{i,t-1}$ 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 standardization 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 in principle distinguish them. 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 13. *(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 of* $\lambda_i|y_{i0}$ *is non-vanishing almost everywhere.*
4. $T \geq 2$.

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 the 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 $\mathcal{F}_N$ that asymptotically approximates $\mathcal{F}$ and then regularize the asymptotic behavior of $\mathcal{F}_N$ instead of $\mathcal{F}$.

Lemma 14. *(Canale and De Blasi, 2017) Suppose we have:*

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

$$\Pi\left(f:\ D_{KL}\left(f_0 \,\|\, f\right) < \epsilon\right) > 0.$$

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

(b) For some $\gamma > 0$, $\sum_j \sqrt{\mathcal{N}\left(\epsilon, \mathcal{F}_{N,j}\right)\Pi\left(\mathcal{F}_{N,j}\right)} = o\left(\exp\left(\left(1-\gamma\right)N\epsilon^2\right)\right)$, where $\mathcal{N}\left(\epsilon, \mathcal{F}_{N,j}\right)$ is the covering number of $\mathcal{F}_{N,j}$ by balls with radius ϵ in the L_1-norm.[1]

Then, the posterior is strongly consistent at f_0, i.e. for all $\epsilon > 0$, as $N \to \infty$,

$$\Pi\left(f:\ \|f - f_0\|_1 < \epsilon \,|\, D\right) \to 1,$$

in probability with respect to the true DGP.

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

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

For the numerator, the sieve property ensures that the sieve expands to the 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

[1] As 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.

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) imposes more regularities on the prior Π and thus weaker assumptions on the true distribution f_0 (see Assumptions 6 and 9).

Lemma 14 establishes posterior consistency in a density estimation context. However, as mentioned in the introduction, there are a number of challenges in adapting to the dynamic panel data setting. The first challenge is, because we observe y_{it} rather than λ_i, to disentangle the uncertainty generated from unknown cross-sectional heterogeneity λ_i and from independent shocks u_{it}, i.e. a deconvolution problem.[2] The second is to incorporate an unknown shock size σ^2 in cross-sectional homoskedastic cases.[3] The third is to handle strictly exogenous and predetermined variables (including lagged dependent variables) as covariates. The fourth is to address correlated random coefficients by a flexible conditional density estimation.

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 $t = 1, \cdots, T - d_w$,

$$\tilde{y}_{it} = y_{it} - \sum_{s=t+1}^{T} y_{is} w'_{i,s-1} \left(\sum_{s=t+1}^{T} w_{i,s-1} w'_{i,s-1} \right)^{-1} w_{i,t-1}, \tag{8}$$

$$\tilde{x}_{i,t-1} = x_{i,t-1} - \sum_{s=t+1}^{T} x_{i,s-1} w'_{i,s-1} \left(\sum_{s=t+1}^{T} w_{i,s-1} w'_{i,s-1} \right)^{-1} w_{i,t-1}. \tag{9}$$

[2] Some previous studies (Amewou-Atisso *et al.*, 2003; Tokdar, 2006) estimate distributions of quantities that can be inferred from observables given common coefficients. For example, in the linear regression problems with an unknown error distribution, i.e. $y_i = \beta' x_i + u_i$, conditional on the regression coefficients β, $u_i = y_i - \beta' x_i$ is inferrable from the data. However, here the target λ_i intertwines with u_{it} and cannot be easily inferred from the observed y_{it}.

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

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

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

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

$$\begin{aligned} \tilde{u}_{it} &= \tilde{y}_{it} - \beta' \tilde{x}_{i,t-1}, \\ s_i^2 &= \sum_{t=1}^{T-d_w} \tilde{u}_{it}^2 = \sigma_i^2 k_i^2, \end{aligned} \tag{10}$$

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

Note that the Fourier transform (i.e. characteristic functions with sign reversal) is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004). For a generic variable z, the Mellin transform of $f(z)$ is specified as[4]

$$\mathcal{M}_z (\xi) = \int z^{i\xi} f(z)\, dz,$$

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

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

$$\mathcal{M}_{s^2} (\xi | c_{i0}) = \mathcal{M}_{\sigma^2} (\xi | c_{i0}) \mathcal{M}_{k^2} (\xi).$$

Note that a chi-square distribution has a non-vanishing Mellin transform, so it is legitimate to devide $\mathcal{M}_{k^2} (\xi | c_{i0})$ on both sides

$$\mathcal{M}_{\sigma^2} (\xi | c_{i0}) = \mathcal{M}_{s^2} (\xi | c_{i0}) / \mathcal{M}_{k^2} (\xi),$$

which recovers $\mathcal{M}_{\sigma^2} (\xi | c_{i0})$ 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 $\mathring{Y}_i = \mathring{y}_{i,1:T}$, $W_i = w'_{i,0:T-1}$, and $U_i = u_{i,1:T}$. Omitting subscript i, the above expression can be simplified as

$$\mathring{Y} = W\lambda + U.$$

Denote $\hat{f}_z$ as the Fourier transform of f_z, for $z = \mathring{Y}, \lambda, U$. Based on Assumption 1(2-a), $\hat{f}_\lambda (\cdot | c_{i0})$

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

and $\hat{f}_U(\cdot|c_{i0})$ are non-vanishing almost everywhere. Then, we obtain

$$\log \hat{f}_\lambda \left(W'\xi|c_{i0}\right) = \log \hat{f}_{\mathring{Y}}(\xi|c_{i0}) - \log \hat{f}_U(\xi|c_{i0}),$$

where $\hat{f}_{\mathring{Y}}$ is constructed from the observables and the common parameters identified in part 1, and $\hat{f}_U$ is based on f_{σ^2} identified in part 2. Note that W is non-random conditional on c_{i0}. Let $\zeta = W'\xi$ and $A_W = (W'W)^{-1} W'$, then the second derivative of $\log \hat{f}_\lambda(\zeta|c_{i0})$ is characterized by

$$\frac{\partial^2}{\partial\zeta\partial\zeta'} \log \hat{f}_\lambda(\zeta|c_{i0}) = A_W \left(\frac{\partial^2}{\partial\xi\partial\xi'}\left(\log \hat{f}_{\mathring{Y}}(\xi|c_{i0}) - \log \hat{f}_U(\xi|c_{i0})\right)\right) A_W'.$$

Moreover,

$$\begin{aligned} \log \hat{f}_\lambda(0|c_{i0}) &= 0, \\ \frac{\partial}{\partial\zeta} \log \hat{f}_\lambda(0|c_{i0}) &= -iA_W \mathbb{E}\left(\mathring{Y}\middle| c_{i0}\right), \end{aligned}$$

so we can pin down $\log \hat{f}_\lambda(\zeta|c_{i0})$ and then $f_\lambda(\lambda_i|c_{i0})$.

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. □

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 $N \to \infty$,

(1) for all $\delta > 0$, $\Pi_\vartheta\left(\vartheta \in \Theta_\delta^c \middle| D\right) \to 0$,

(2) for all $\epsilon > 0$, $\Pi_f\left(f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N^c \middle| D\right) \to 0$,

(3) for all $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that $\Pi\left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D\right) \to 0$,

in probability with respect to the true DGP (Lemmas 16, 17, and 18, 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

$$\begin{aligned} &\Pi\left(\vartheta \in \Theta_\delta^c \text{ or } f \in \mathcal{F}_\epsilon^c \middle| D\right) \\ \le &\Pi_\vartheta\left(\vartheta \in \Theta_{\delta'(\epsilon)}^c \middle| D\right) + \Pi_f\left(f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N^c \middle| D\right) + \Pi\left(\vartheta \in \Theta_{\delta'(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D\right) \to 0, \end{aligned} \tag{11}$$

as $N \to \infty$, in probability with respect to the true DGP. □

Lemma 15. *Suppose condition 1-a in Theorem 4 holds, then, for all $\eta > 0$, as $N \to \infty$,*

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

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 $N \to \infty$. $\square$

Lemma 16. *Suppose conditions 1-a and 2 in Theorem 4 hold, then, for all $\delta > 0$, as $N \to \infty$,*

$$\Pi_\vartheta\left(\vartheta \in \Theta_\delta^c \middle| D\right) \to 0,$$

almost surely with respect to the true DGP.

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

$$\begin{aligned}\Pi_\vartheta\left(\vartheta \in \Theta_\delta^c \middle| D\right) &= \frac{\int_{\Theta_\delta^c\times\mathcal{F}} R_N(D, \vartheta, f)\, d\Pi(\vartheta, f)}{\int_{\Theta\times\mathcal{F}} R_N(D, \vartheta, f)\, d\Pi(\vartheta, f)} \\ &\leq \varphi_N(D) + \frac{(1-\varphi_N(D)) \int_{\Theta_\delta^c\times\mathcal{F}} R_N(D, \vartheta, f)\, d\Pi(\vartheta, f)}{\int_{\Theta\times\mathcal{F}} R_N(D, \vartheta, f)\, d\Pi(\vartheta, f)}.\end{aligned} \tag{12}$$

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

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

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

$$\exp\left(C_\varphi N/2\right)\left(1-\varphi_N\left(D\right)\right)\int_{\Theta_\delta^c\times\mathcal{F}} R_N\left(D,\vartheta,f\right)d\Pi\left(\vartheta,f\right)\to 0, \tag{13}$$

almost surely with respect to the true DGP. For the denominator in the second term, condition 1-a in Theorem 4 ensures Lemma 15. If we let $\eta = C_\varphi/4$, then as $N \to \infty$,

$$\exp\left(C_\varphi N/4\right)\int_{\Theta\times\mathcal{F}} R_N\left(D,\vartheta,f\right)d\Pi\left(\vartheta,f\right)\to \infty, \tag{14}$$

almost surely with respect to the true DGP. Combining (12), (13), and (14), we prove the lemma. □

Lemma 17. *Suppose conditions 1-a and 3-a in Theorem 4 hold, then, for all $\epsilon > 0$, as $N \to \infty$,*

$$\Pi_f\left(f\in\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_N^c\,\middle|\,D\right)\to 0,$$

almost surely with respect to the true DGP.

Proof. Decompose the posterior probability as follows,

$$\Pi_f\left(f\in\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_N^c\,\middle|\,D\right)=\frac{\int_{\Theta\times\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_N^c} R_N\left(D,\vartheta,f\right)d\Pi\left(\vartheta,f\right)}{\int_{\Theta\times\mathcal{F}} R_N\left(D,\vartheta,f\right)d\Pi\left(\vartheta,f\right)}. \tag{15}$$

For the numerator,

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

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

$$\exp\left(\beta N/2\right)\left(1-\varphi_N\left(D\right)\right)\int_{\Theta\times\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_N^c} R_N\left(D,\vartheta,f\right)d\Pi\left(\vartheta,f\right)\to 0, \tag{16}$$

almost surely with respect to the true DGP. For the denominator, condition 1-a in Theorem 4 ensures Lemma 15. If we let $\eta = \beta/4$, then as $N \to \infty$,

$$\exp\left(\beta N/4\right) \int_{\Theta\times\mathcal{F}} R_N\left(D, \vartheta, f\right) d\Pi\left(\vartheta, f\right) \to \infty, \tag{17}$$

almost surely with respect to the true DGP. Combining (15), (16), and (17), we prove the lemma. $\square$

Lemma 18. *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 $N \to \infty$,*

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

in probability with respect to the true DGP.

Proof. **1. Inversion inequality.** Define $\epsilon' = \mathfrak{C}^{-1}(\epsilon)/9$ and

$$\delta(\epsilon) = \min\left(\epsilon'/\left(4C_g\right), \delta_\vartheta/2\right) = \min\left(\mathfrak{C}^{-1}(\epsilon)/\left(36C_g\right), \delta_\vartheta/2\right),$$

then for all $\|\vartheta - \vartheta_0\|_2 < \delta(\epsilon)$, $f \in \mathcal{F}_\epsilon^c$,

$$\begin{aligned} & \left\|g\left(D_i|\vartheta, f\right) - g\left(D_i|\vartheta_0, f_0\right)\right\|_1 \\ \geq & \left\|g\left(D_i|\vartheta_0, f\right) - g\left(D_i|\vartheta_0, f_0\right)\right\|_1 - \left\|g\left(D_i|\vartheta, f\right) - g\left(D_i|\vartheta_0, f\right)\right\|_1 \\ \geq & \mathfrak{C}^{-1}\left(W_2\left(f, f_0\right)\right) - C_g\left\|\vartheta - \vartheta_0\right\|_2 \\ \geq & 9\epsilon' - \frac{1}{4}\epsilon' > 8\epsilon'. \end{aligned} \tag{18}$$

The second line is given by the triangular inequality. The first and second terms in the third line follow conditions 1-c and 1-b in Theorem 4, respectively. 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

$$\mathbb{E}_{g_0}^N \tilde{\varphi}_N(D) \leq \sqrt{\frac{\gamma_2}{\gamma_1}} \mathcal{N}\left(\epsilon', G\right) \exp\left(-N\epsilon'^2\right) \text{ and } \sup_{g\in G} \mathbb{E}_g^N\left[1 - \tilde{\varphi}_N(D)\right] \leq \sqrt{\frac{\gamma_1}{\gamma_2}} \exp\left(-N\epsilon'^2\right). \tag{19}$$

[5] We let δ depend on ϵ here because Lemma 16 holds for all $\delta > 0$.

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

2. Sieve prorperty. For all $\|\vartheta_1 - \vartheta_2\|_2 < \delta_\vartheta$ and $f_1, f_2 \in \mathcal{F}$,

$$\begin{aligned}
&\|g(D_i|\,\vartheta_1, f_1) - g(D_i|\,\vartheta_2, f_2)\|_1 \\
\le &\|g(D_i|\,\vartheta_2, f_1) - g(D_i|\,\vartheta_2, f_2)\|_1 + \|g(D_i|\,\vartheta_1, f_1) - g(D_i|\,\vartheta_2, f_1)\|_1 \\
\le &\|f_1 - f_2\|_1 + C_g \|\vartheta_1 - \vartheta_2\|_2 .
\end{aligned}$$

The bound of the first term is based on the general "convolution" form of the individual likelihood in (7), and the bound of the second term follows Theorem 4(1-b). If f is an unconditional distribution,

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

If f is a conditional distribution,

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

Considering ϵ' and $\delta(\epsilon)$ defined in part 1, 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 $\|f_1 - f_2\|_1 < \epsilon'/2$,

$$\begin{aligned}
\|g(D_i|\,\vartheta_1, f_1) - g(D_i|\,\vartheta_2, f_2)\|_1 &\le \|f_1 - f_2\|_1 + C_g \|\vartheta_1 - \vartheta_2\|_2 \\
&< \frac{1}{2}\epsilon' + \frac{1}{2}\epsilon' = \epsilon'.
\end{aligned}$$

Let $\mathcal{G}$ be the space induced by $f \in \mathcal{F}$ and $\|\vartheta - \vartheta_0\|_2 < \delta(\epsilon)$ according to the likelihood function in (7), then for all $G \in \mathcal{G}$ induced by $F \in \mathcal{F}$, the covering number

$$\mathcal{N}\left(\epsilon', G\right) \le \mathcal{N}\left(\epsilon'/2, F\right). \tag{20}$$

3. Asymptotic analysis. Define

$$H_N = \left\{ D : \int R_N(D, \vartheta, f)\, d\Pi(\vartheta, f) \ge \exp\left(-\gamma_0 N \epsilon'^2\right) \right\}, \tag{21}$$

where $\gamma_0 \leq (3+\gamma)/4$ for γ in Theorem 4 condition 3-b. Lemma 15 implies that as $N \to \infty$, $\mathbb{P}_0^N(H_N) \to 1$, where $\mathbb{P}_0^N$ is characterized by the true DGP with the sample size being N. Hence,

$$\begin{aligned}
&\mathbb{E}_{g_0}^N\left[\Pi\left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D\right)\right] \\
=&\mathbb{E}_{g_0}^N\left[\Pi\left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_N \middle| D\right)\mathbf{1}(H_N)\right] + o(1) \\
=&\sum_j \mathbb{E}_{g_0}^N\left[\Pi\left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_{N,j} \middle| D\right)\mathbf{1}(H_N)\right] + o(1).
\end{aligned} \tag{22}$$

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. For each j, given (18), we have $\inf_{g \in \mathcal{G}_\epsilon^c \bigcap \mathcal{G}_{N,j}} \|g - g_0\|_1 \geq 8\epsilon'$, so there exists a $\tilde{\varphi}_{N,j}(D)$ for each $\mathcal{G}_\epsilon^c \bigcap \mathcal{G}_{N,j}$, and we can decompose the posterior probability as follows:

$$\begin{aligned}
&\mathbb{E}_{g_0}^N\left[\Pi\left(\vartheta \in \Theta_{\delta(\epsilon)} \text{ and } f \in \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_{N,j} \middle| D\right)\mathbf{1}(H_N)\right] \\
&\leq \mathbb{E}_{g_0}^N\tilde{\varphi}_{N,j}(D) + \mathbb{E}_{g_0}^N\left[(1 - \tilde{\varphi}_{N,j}(D))\int_{\Theta_{\delta(\epsilon)} \times \mathcal{F}_\epsilon^c \bigcap \mathcal{F}_{N,j}} R_N(D, \vartheta, f)\, d\Pi(\vartheta, f)\right]\exp\left(\gamma_0 N \epsilon'^2\right).
\end{aligned} \tag{23}$$

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,

$$\begin{aligned}
&\mathbb{E}_{g_0}^N\tilde{\varphi}_{N,j}(D) \\
\leq&\sqrt{\frac{\gamma_{2,j}}{\gamma_{1,j}}}\mathcal{N}\left(\epsilon', \mathcal{G}_\epsilon^c \bigcap \mathcal{G}_{N,j}\right)\exp\left(-N\epsilon'^2\right) \\
\leq&\sqrt{\mathcal{N}(\epsilon', \mathcal{G}_{N,j})\,\Pi_f(\mathcal{F}_{N,j})}\exp\left(-N\epsilon'^2\right),
\end{aligned} \tag{24}$$

where the second line is given by the test in (19). For the second term, note that

$$
\begin{aligned}
&\mathbb{E}_{g_0}^N\left[\left(1-\tilde{\varphi}_{N,j}(D)\right)\int_{\Theta_{\delta(\epsilon)}\times\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_{N,j}} R_N(D,\vartheta,f)\,d\Pi(\vartheta,f)\right] \qquad (25)\\
=&\int\left(1-\tilde{\varphi}_N(D)\right)\left[\int_{\Theta_{\delta(\epsilon)}\times\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_{N,j}} R_N(D,\vartheta,f)\,d\Pi(\vartheta,f)\right]\prod_{i=1}^N g\left(D_i|\,\vartheta_0,f_0\right)dD\\
=&\int_{\Theta_{\delta(\epsilon)}\times\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_{N,j}}\left[\int\left(1-\tilde{\varphi}_N(D)\right)\prod_{i=1}^N g\left(D_i|\,\vartheta,f\right)dD\right]d\Pi(\vartheta,f)\\
\leq&\sup_{g\in\mathcal{G}_\epsilon^c\bigcap\mathcal{G}_{N,j}}\mathbb{E}_g^N\left[1-\tilde{\varphi}_N(D)\right]\cdot\Pi\left(\Theta_{\delta(\epsilon)},\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_{N,j}\right)\\
\leq&\sup_{g\in\mathcal{G}_\epsilon^c\bigcap\mathcal{G}_{N,j}}\mathbb{E}_g^N\left[1-\tilde{\varphi}_N(D)\right]\cdot\Pi_f\left(\mathcal{F}_{N,j}\right)\\
\leq&\sqrt{\frac{\gamma_{1,j}}{\gamma_{2,j}}}\exp\left(-N\epsilon'^2\right)\cdot\Pi_f\left(\mathcal{F}_{N,j}\right)\\
\leq&\sqrt{\mathcal{N}\left(\epsilon',\mathcal{G}_{N,j}\right)\Pi_f\left(\mathcal{F}_{N,j}\right)}\exp\left(-N\epsilon'^2\right),
\end{aligned}
$$

where the second to last line is given by the test in (19). Combining (23), (24), and (25), as $N\to\infty$,

$$
\begin{aligned}
&\sum_j\mathbb{E}_{g_0}^N\left[\Pi\left(\vartheta\in\Theta_{\delta(\epsilon)}\text{ and }f\in\mathcal{F}_\epsilon^c\bigcap\mathcal{F}_{N,j}\middle|\,D\right)\mathbf{1}\left(H_N\right)\right]\\
\leq&\sum_j\sqrt{\mathcal{N}\left(\epsilon',\mathcal{G}_{N,j}\right)\Pi_f\left(\mathcal{F}_{N,j}\right)}\exp\left(-N\epsilon'^2\left(1-\gamma_0\right)\right)\\
\leq&\sum_j\sqrt{\mathcal{N}\left(\epsilon'/2,\mathcal{F}_{N,j}\right)\Pi_f\left(\mathcal{F}_{N,j}\right)}\exp\left(-N\epsilon'^2\left(1-\gamma_0\right)\right)\\
=&o\left(\exp\left(\left(1-\gamma\right)N\epsilon'^2/4\right)\exp\left(-N\epsilon'^2\left(1-\gamma_0\right)\right)\right)\\
=&o\left(\exp\left(-N\epsilon'^2\left(1-\gamma_0-\left(1-\gamma\right)/4\right)\right)\right)\\
\to&0.
\end{aligned}
$$

The third line converts 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 $\gamma_0\leq\left(3+\gamma\right)/4$. Then, according to (22), as $N\to\infty$,

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

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

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

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

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

C.3.1 Random Coefficients: Cross-sectional Homoskedasticity

Remark 19. To ensure condition 1 in Theorem 4, we consider space $\mathcal{F} = \left\{ f : \mathbb{E}_f \left\| \lambda \right\|_2^{2(1+\eta)} \leq M \right\}$ for some large $M > 0$. Given Assumption 6(1-e), f_0 satisfies this condition when M is large enough. Let $\bar{\mathcal{F}}$ be the space of all possible underlying distribution of individual heterogeneity f (with or without bounded $2(1+\eta)$-th moments). Then, $\mathcal{F} \subseteq \bar{\mathcal{F}}$. Let $\bar{\Pi}$ be the corresponding probability measure on $\bar{\mathcal{F}}$. According to Bayes' theorem, for any event A,

$$\Pi(A) = \bar{\Pi}(A | \mathcal{F}) = \frac{\bar{\Pi}(A \cap \mathcal{F})}{\bar{\Pi}_f(\mathcal{F})}.$$

As the denominator $0 \leq \bar{\Pi}_f(\mathcal{F}) \leq 1$, we have

$$\Pi(A) = \frac{\bar{\Pi}(A \cap \mathcal{F})}{\bar{\Pi}_f(\mathcal{F})} \geq \bar{\Pi}(A \cap \mathcal{F}).$$

Thus, to verify condition 1-a in Theorem 4, it suffices to prove that for all $\epsilon > 0$,

$$\bar{\Pi}\left((\vartheta, f) : \left\{ D_{KL}\left(g(D_i | \vartheta_0, f_0) \,\|\, g(D_i | \vartheta, f)\right) < \epsilon \right\} \cap \mathcal{F}\right) > 0, \tag{26}$$

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

$$\bar{\Pi}_f(\mathcal{F}) > 1 - \tau. \tag{27}$$

Therefore, we have

$$\Pi(A) = \frac{\bar{\Pi}(A \cap \mathcal{F})}{\bar{\Pi}_f(\mathcal{F})} < \frac{\bar{\Pi}(A \cap \mathcal{F})}{1 - \tau} \leq \frac{\bar{\Pi}(A)}{1 - \tau}.$$

It implies that to verify condition 3 in Theorem 4, it suffices to prove that for all $\epsilon > 0$ and for some $\beta, \gamma > 0$,

$$\bar{\Pi}_f(\mathcal{F}_N^c) = O(\exp(-\beta N)), \text{ and } \sum_j \sqrt{\mathcal{N}(\epsilon, \mathcal{F}_{N,j}) \bar{\Pi}_f(\mathcal{F}_{N,j})} = o\left(\exp\left((1-\gamma) N \epsilon^2\right)\right). \tag{28}$$

Remark 20. Here I demonstrate that if for all $\epsilon > 0$, $\bar{\Pi}_f \left(D_{KL} \left(f_0 \parallel f \right) < \epsilon \right) > 0$, then

$$\bar{\Pi}_f \left(f : \left\{ D_{KL} \left(f_0 \parallel f \right) < \epsilon \right\} \cap \mathcal{F} \right) > 0. \tag{29}$$

Let $\mathcal{F}_{KL,\epsilon} = \left\{ f \in \bar{\mathcal{F}} : D_{KL} \left(f_0 \parallel f \right) < \epsilon \right\}$. First, we can obtain (27) in Remark 19 based on Doss and Sellke (1982) and Egorov's Theorem. Then, there exists $\epsilon^* > 0$ such that $\bar{\Pi}_f \left(\mathcal{F}_{KL,\epsilon^*} \right) > \tau$, so we have

$$\bar{\Pi}_f \left(\mathcal{F}_{KL,\epsilon^*} \cap \mathcal{F} \right) > 0. \tag{30}$$

(1) If $\epsilon \geq \epsilon^*$, the above expression implies $\bar{\Pi}_f \left(\mathcal{F}_{KL,\epsilon} \cap \mathcal{F} \right) > 0$, which is equivalent to (29).
(2) If $\epsilon < \epsilon^*$, let $w = \epsilon / \epsilon^*$, then for all $f^* \in \mathcal{F}_{KL,\epsilon^*} \cap \mathcal{F}$, we can construct

$$f = w f^* + (1 - w) f_0. \tag{31}$$

Thus,

$$\begin{aligned} D_{KL} \left(f_0 \parallel f \right) &= \int f_0 \log \frac{f_0}{f} \\ &\leq w \int f_0 \log \frac{f_0}{f^*} + (1 - w) \int f_0 \log \frac{f_0}{f_0} \\ &< w \epsilon^* = \epsilon, \end{aligned} \tag{32}$$

where the second line is given by the convexity of $(-\log x)$. At the same time, when M is sufficiently large, $\mathbb{E}_{f_0} \left\| \lambda \right\|_2^{2(1+\eta)} \leq M$, then,

$$\int \left\| \lambda_i \right\|_2^{2(1+\eta)} f(\lambda_i) d\lambda_i = w \int \left\| \lambda_i \right\|_2^{2(1+\eta)} f^*(\lambda_i) d\lambda_i + (1 - w) \int \left\| \lambda_i \right\|_2^{2(1+\eta)} f_0(\lambda_i) d\lambda_i \leq M. \tag{33}$$

Combining (32) and (33), we obtain $f \in \mathcal{F}_{KL,\epsilon} \cap \mathcal{F}$. Also note that (31) is an invertible linear mapping from $\mathcal{F}_{KL,\epsilon^*} \cap \mathcal{F}$ to $\mathcal{F}_{KL,\epsilon} \cap \mathcal{F}$, i.e. an isomorphism. Therefore, considering (30) and the fact that $\bar{\Pi}_f$ has full support,[7] we have

$$\bar{\Pi}_f \left(\mathcal{F}_{KL,\epsilon} \cap \mathcal{F} \right) > 0.$$

Proof. **(Theorem 7)**

[7] More specifically, for all f with $\operatorname{supp}(f) \in \operatorname{supp}(G_0)$, we have $f \in \operatorname{supp}\left(\bar{\Pi}_f\right)$ (see Theorem 3.2.4 in Ghosh and Ramamoorthi (2003)). Especially, if G_0 has full support on Θ, then $\bar{\Pi}_f$ has full support on $\bar{\mathcal{F}}$. Here, G_0 has full support on $\mathbb{R}^{d_w} \times \mathcal{S}$, where $\mathcal{S}$ is the space of $d_w \times d_w$ positive definite matrices with the spectral norm (the spectral norm is induced by the L_2-norm on vectors, $\left\| \Omega \right\|_2 = \max_{x \neq 0} \frac{\left\| \Omega x \right\|_2}{\left\| x \right\|_2}$).

The individual-specific likelihood function is characterized as

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

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 $\epsilon > 0$,

$$\bar{\Pi}_f\left(f:\ D_{KL}\left(f_0 \parallel f\right) < \epsilon\right) > 0.$$

Then, Remark 20 shows that

$$\bar{\Pi}_f\left(f:\ \{D_{KL}\left(f_0 \parallel f\right) < \epsilon\} \cap \mathcal{F}\right) > 0. \tag{34}$$

Now, we need to establish an altered KL property specified on g (the distribution of observables) based on sufficient condition (26) in Remark 19. The KL divergence of $g\left(D_i|\,\vartheta, f\right)$ with respect to $g\left(D_i|\,\vartheta_0, f_0\right)$ can be decomposed as

$$\begin{aligned} 0 &\le \int g\left(D_i|\,\vartheta_0, f_0\right) \log \frac{g\left(D_i|\,\vartheta_0, f_0\right)}{g\left(D_i|\,\vartheta, f\right)} dD_i \\ &= \int g\left(D_i|\,\vartheta_0, f_0\right) \log \frac{g\left(D_i|\,\vartheta_0, f_0\right)}{g\left(D_i|\,\vartheta_0, f\right)} dD_i + \int g\left(D_i|\,\vartheta_0, f_0\right) \log \frac{g\left(D_i|\,\vartheta_0, f\right)}{g\left(D_i|\,\vartheta, f\right)} dD_i. \end{aligned} \tag{35}$$

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

$$\int g\left(D_i|\,\vartheta_0, f_0\right) \log \frac{g\left(D_i|\,\vartheta_0, f_0\right)}{g\left(D_i|\,\vartheta_0, f\right)} dD_i = \int g\left(D_i|\,\vartheta_0, f_0\right) \log \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i} dD_i.$$

We can apply the convolution property of the KL divergence in Lemma 29(1) to the integral over λ_i

$$\begin{aligned} &\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i \log \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i} \\ &\le \int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) \log \frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)} d\lambda_i. \end{aligned}$$

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

$$
\begin{aligned}
0 \leq & \int g\left(D_i \middle| \vartheta_0, f_0\right) \log \frac{g\left(D_i \middle| \vartheta_0, f_0\right)}{g\left(D_i \middle| \vartheta_0, f\right)} dD_i \\
= & \int g\left(D_i \middle| \vartheta_0, f_0\right) \log \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) f_0\left(\lambda_i\right) d\lambda_i} dD_i \\
\leq & \int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \left[\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) \log \frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)} d\lambda_i\right] dD_i \\
= & \int \left[\int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) dD_i\right] f_0\left(\lambda_i\right) \log \frac{f_0\left(\lambda_i\right)}{f\left(\lambda_i\right)} d\lambda_i \\
= & D_{KL}\left(f_0 \parallel f\right).
\end{aligned}
$$

According to the KL property on f in (34), define

$$
S_{f,\epsilon} = \left\{f : \left\{D_{KL}\left(f_0 \parallel f\right) < \frac{\epsilon}{3}\right\} \cap \mathcal{F}\right\},
$$

then $\bar{\Pi}_f\left(S_{f,\epsilon}\right) > 0$, and for all $f \in S_{f,\epsilon}$, the first term

$$
0 \leq \int g\left(D_i \middle| \vartheta_0, f_0\right) \log \frac{g\left(D_i \middle| \vartheta_0, f_0\right)}{g\left(D_i \middle| \vartheta_0, f\right)} dD_i < \frac{\epsilon}{3}. \tag{36}
$$

Second term: Given the bounds in (35) and (36), we have

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

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

$$
\begin{aligned}
-\frac{\epsilon}{3} &< \int g\left(D_i \middle| \vartheta_0, f_0\right) \log \frac{g\left(D_i \middle| \vartheta_0, f\right)}{g\left(D_i \middle| \vartheta, f\right)} dD_i \\
&= \int g\left(D_i \middle| \vartheta_0, f_0\right) \log \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) f\left(\lambda_i\right) d\lambda_i} dD_i \\
&= \int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i} \\
&\quad \cdot \int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i \log \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) f\left(\lambda_i\right) d\lambda_i} dD_i \\
&\le \int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \frac{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i}{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i} \\
&\quad \cdot \left[\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) \log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} d\lambda_i \right] dD_i \\
&= \int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i \\
&\quad \cdot \left[\int \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right)}{\int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f\left(\lambda_i\right) d\lambda_i} \log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} d\lambda_i \right] dD_i \\
&= \int \prod_t p\left(x_{i,t-1}^{P*} \middle| y_{i,t-1}, c_{i,0:t-2}\right) p\left(c_{i0}\right) \int \prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right) f_0\left(\lambda_i\right) d\lambda_i \\
&\quad \cdot \left[\int \frac{\phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right)}{\int \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right) d\lambda_i} \log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} d\lambda_i \right] dD_i,
\end{aligned}
\tag{37}
$$

where

$$
\begin{aligned}
m_i\left(\beta_0\right) &= \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1} \sum_t w_{i,t-1}\left(y_{it} - \beta_0' x_{i,t-1}\right), \\
\Sigma_i\left(\sigma_0^2\right) &= \sigma_0^2 \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1}.
\end{aligned}
\tag{38}
$$

The second line in (37) 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 29(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,

$$\begin{aligned}
&\log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} \\
=&\frac{T}{2}\left(\log \sigma^2 - \log \sigma_0^2\right) + \frac{1}{2}\sum_t \left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right) \\
&+ \sum_t \frac{\left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 - \left(y_{it} - \beta_0' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2}{2\sigma_0^2} \\
=&\frac{T}{2}\left(\log \sigma^2 - \log \sigma_0^2\right) + \frac{1}{2}\sum_t \left(y_{it} - \beta' x_{i,t-1} - \lambda_i' w_{i,t-1}\right)^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right) \\
&+ \sum_t \frac{\left(\beta' x_{i,t-1}\right)^2 - \left(\beta_0' x_{i,t-1}\right)^2 - 2\left(y_{it} - \lambda_i' w_{i,t-1}\right)\left(\beta - \beta_0\right)' x_{i,t-1}}{2\sigma_0^2}.
\end{aligned} \tag{39}$$

Define

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

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

$$\begin{aligned}
\left|\left(\beta' x_{i,t-1}\right)^2 - \left(\beta_0' x_{i,t-1}\right)^2\right| &\leq 2\|\beta_0\|_2 \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2 + \|\beta - \beta_0\|_2^2 \|x_{i,t-1}\|_2^2 \\
&\leq \left(2\|\beta_0\|_2 + \delta_\beta\right) \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2 \\
&\lesssim \|\beta - \beta_0\|_2 \|x_{i,t-1}\|_2^2.
\end{aligned} \tag{40}$$

At the same time,

$$\begin{aligned}
\left|2\left(y_{it} - \lambda_i' w_{i,t-1}\right)\left(\beta - \beta_0\right)' x_{i,t-1}\right| &\leq \|\beta - \beta_0\|_2 \cdot 2\|x_{i,t-1}\|_2 \left(|y_{it}| + \|\lambda_i\|_2 \|w_{i,t-1}\|_2\right) \\
&\leq \|\beta - \beta_0\|_2 \left(2\|x_{i,t-1}\|_2^2 + y_{it}^2 + \|\lambda_i\|_2^2 \|w_{i,t-1}\|_2^2\right) \\
&\lesssim \|\beta - \beta_0\|_2 \left(y_{it}^2 + \|x_{i,t-1}\|_2^2 + \|\lambda_i\|_2^2\right).
\end{aligned} \tag{41}$$

The last line follows that $w_{i,0:T-1}$ is bounded due to Assumption 5(1). Given that T is finite, combining (40) and (41), the last term in (39) is bounded as follows:

$$\begin{aligned}
&\sum_t \frac{\left(\beta' x_{i,t-1}\right)^2 - \left(\beta_0' x_{i,t-1}\right)^2 - 2\left(y_{it} - \lambda_i' w_{i,t-1}\right)\left(\beta - \beta_0\right)' x_{i,t-1}}{2\sigma_0^2} \\
\lesssim &\|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2\right) + \|\lambda_i\|_2^2\right].
\end{aligned} \tag{42}$$

Note that for all $f \in \mathcal{F}$, the second moment of λ_i is bouned by some $M_2 > 0$. We can treat f as a "prior," ϕ as a Gaussian "likelihood," and $\frac{\phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) f(\lambda_i)}{\int \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) f(\lambda_i) d\lambda_i}$ as the "posterior," then the second moment with respect to the "posterior"

$$\begin{aligned} &\int \frac{\phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right)}{\int \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right) d\lambda_i} \left\|\lambda_i\right\|_2^2 d\lambda_i \\ &\lesssim \left(\left\|m_i\left(\beta_0\right)\right\|_2^2 + \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right) + M_2\right) \lesssim \left[\sum_t \left(y_{it}^2 + \left\|x_{i,t-1}\right\|_2^2\right) + 1\right]. \end{aligned} \tag{43}$$

Assumption 5(2) ensures that $\operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right)$ is bounded above. Plugging (43) back to (42), we see that the expression in the brackets in the last line of (37) is bounded as follows:

$$\begin{aligned} &\int \frac{\phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right)}{\int \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right) d\lambda_i} \log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} d\lambda_i \\ &\lesssim \left\|\beta - \beta_0\right\|_2 \left[\sum_t \left(y_{it}^2 + \left\|x_{i,t-1}\right\|_2^2\right) + 1\right]. \end{aligned} \tag{44}$$

Assumptions 6(1-e) and 5(1,3) ensure that $\mathbb{E} y_{it}^2$ and $\mathbb{E}\left\|x_{i,t-1}\right\|_2^2$ exist, so the rest of the integration in (37) is bounded by $C_\beta \left\|\beta - \beta_0\right\|_2$. Define

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

then $\bar{\Pi}_\beta\left(S_{\beta,\epsilon}\right) > 0$, and for all $\beta \in S_{\beta,\epsilon}$, the integral associated with the last term in (39) is less than $\epsilon/3$, i.e.

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

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

$$\bar{\Pi}\left(\left(\vartheta, f\right) : \; D_{KL}\left(g\left(D_i | \vartheta_0, f_0\right) \| g\left(D_i | \vartheta, f\right)\right) < \epsilon\right) > 0.$$

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

$$
\begin{aligned}
&\left\| g\left(D_i \middle| \vartheta_1, f\right) - g\left(D_i \middle| \vartheta_2, f\right) \right\|_1 \\
&\leq \sum_{\tau} \int \prod_{t} p\left(x_{i,t-1}^{P*} \left| y_{i,t-1}, c_{i,0:t-2}\right.\right) p\left(c_{i0}\right) \left[\prod_{t=1}^{\tau-1} \phi\left(y_{it}; \beta_1' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_1^2\right) \prod_{t=\tau+1}^{T} \phi\left(y_{it}; \beta_2' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_2^2\right) \right. \\
&\quad \left. \cdot \left| \ \phi\left(y_{i\tau}; \beta_1' x_{i,\tau-1} + \lambda_i' w_{i,\tau-1}, \sigma_1^2\right) - \phi\left(y_{i\tau}; \beta_2' x_{i,\tau-1} + \lambda_i' w_{i,\tau-1}, \sigma_2^2\right) \ \right| \right] \cdot f\left(\lambda_i\right) d\lambda_i dD_i \\
&= \sum_{\tau} \int \prod_{t=2}^{\tau-1} p\left(x_{i,t-1}^{P*} \left| y_{i,t-1}, c_{i,0:t-2}\right.\right) p\left(c_{i0}\right) \left[\prod_{t=1}^{\tau-1} \phi\left(y_{it}; \beta_1' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_1^2\right) \right. \\
&\quad \left. \cdot \left| \ \phi\left(y_{i\tau}; \beta_1' x_{i,\tau-1} + \lambda_i' w_{i,\tau-1}, \sigma_1^2\right) - \phi\left(y_{i\tau}; \beta_2' x_{i,\tau-1} + \lambda_i' w_{i,\tau-1}, \sigma_2^2\right) \ \right| \right] \cdot f\left(\lambda_i\right) dy_{i,1:\tau} dx_{i,1:\tau-1}^{P*} dc_{i0} d\lambda_i.
\end{aligned}
$$

The last line is given by integrating out y_{it} and $x_{i,t-1}^{P*}$ iteratively for $t = T, T-1, \cdots, \tau+1$. According to Lemma 31 on L_1-distance between normal distributions,

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

The last line follows the facts that $\log(1+x) \leq x$ for all $x > -1$ and that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$. For the first term, note that $0 \leq x - \log(1+x) \leq \frac{x^2}{1+x}$ for all $x > -1$. Given condition 3-b in Theorem 7,

$$
\sqrt{\frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1} \leq \frac{\left|\sigma_1^2 - \sigma_2^2\right|^2}{\sigma_1^2 \sigma_2^2} \leq \frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\left(\underline{\sigma}^2\right)^2} \left|\sigma_1^2 - \sigma_2^2\right|.
$$

For the second term,

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

$\mathbb{E}_f \left\| x_{i,\tau-1} \right\|_2$ exists based on Assumption 5(3) and the fact that for all $f \in \mathcal{F}$, the second moment of λ_i is bouned by some $M_2 > 0$. Therefore, for all $\vartheta_1, \vartheta_2 \in \Theta$, there exists $C_g > 0$ not depending on f such that

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

3. Condition 1-c in Theorem 4. This part of the proof builds on the proofs of Theorem 2 in Nguyen (2013) and Lemma 1 in Su *et al.* (2020). For notation simplicity, let $\hat{f}$ denote the Fourier transform of f. Let K be an density on $\mathbb{R}$. Suppose that (1) K is symmetric and has a bounded $2(1+\eta)$ moment, where η is defined in Assumption 6(1-e); and (2) its Fourier transform $\hat{K}$ is continuous with $\operatorname{supp}\left(\hat{K}\right)=[-1,\ 1]$. Denote the mollifier $\mathcal{K}_\delta(\upsilon)=\delta^{-d_w}\prod_{j=1}^{d_w}K(\upsilon_j/\delta)$, where $\upsilon=(\upsilon_1,\upsilon_2,\cdots,\upsilon_{d_w})'\in\mathbb{R}^{d_w}$. Let $f*\mathcal{K}_\delta$ be the convolution of f and $\mathcal{K}_\delta$. Following the triangular inequality,

$$W_2^2(f,f_0)\leq W_2^2(f,f*\mathcal{K}_\delta)+W_2^2(f_0,f_0*\mathcal{K}_\delta)+W_2^2(f*\mathcal{K}_\delta,f_0*\mathcal{K}_\delta).$$

First and second terms: Consider coupling $(\lambda,\lambda+\upsilon)$, where the marginal distributions of λ and υ are f and $\mathcal{K}_\delta$, respectively. Then, by the definition of the Wasserstein metric,

$$\begin{aligned}W_2^2(f,f*\mathcal{K}_\delta)&\leq\int\|\lambda-(\lambda+\upsilon)\|_2^2 f(\lambda)\mathcal{K}_\delta(\upsilon)\,d\lambda d\upsilon\\&=\int\left[\int f(\lambda)\,d\lambda\right]\|\upsilon\|_2^2\mathcal{K}_\delta(\upsilon)\,d\upsilon\\&=\int\|\upsilon\|_2^2\cdot\frac{1}{\delta^{d_w}}\prod_{j=1}^{d_w}K\left(\frac{\upsilon_j}{\delta}\right)d\upsilon.\end{aligned}$$

Define $\tilde{\upsilon}=\upsilon/\delta$, then

$$\int\|\upsilon\|_2^2\cdot\frac{1}{\delta^{d_w}}\prod_{j=1}^{d_w}K\left(\frac{\upsilon_j}{\delta}\right)d\upsilon=\delta^2\int\|\tilde{\upsilon}\|_2^2\cdot\prod_{j=1}^{d_w}K(\tilde{\upsilon})\,d\tilde{\upsilon}\lesssim\delta^2.$$

The last inequality obtains as the second moment of K is bounded. Similarly, we have the second term $W_2^2(f_0,f_0*\mathcal{K}_\delta)\lesssim\delta^2$ as well.

Third term: Let z be a generic variable. According to Theorem 6.15 in Villani (2009),

$$\begin{aligned}&W_2^2(f*\mathcal{K}_\delta,f_0*\mathcal{K}_\delta)\\&\lesssim\int\|z\|_2^2\left|(f-f_0)*\mathcal{K}_\delta(z)\right|dz\\&=\int_{\|z\|_2\leq\mathcal{M}}\|z\|_2^2\left|(f-f_0)*\mathcal{K}_\delta(z)\right|dz+\int_{\|z\|_2>\mathcal{M}}\|z\|_2^2\left|(f-f_0)*\mathcal{K}_\delta(z)\right|dz,\end{aligned}$$

for some large $\mathcal{M}>0$ that could depend on $\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\|_1$.

Third term - first part: Similar to (38), define

$$m_i=\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)^{-1}\sum_t w_{i,t-1}\left(y_{it}-\beta_0'x_{i,t-1}\right),\quad\Sigma_i=\sigma_0^2\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)^{-1}.$$

Then,

$$m_i = \lambda_i + \bar{u}_i, \quad \bar{u}_i \sim \bar{\phi}(\bar{u}_i) = \int \phi(\bar{u}_i; 0, \Sigma_i)\, p(w_{i,0:T-1})\, dw_{i,0:T-1},$$

and the distribution of m_i is

$$\tilde{g}(m_i) = p(m_i | \vartheta_0, f) = f * \bar{\phi}(m_i),$$

Also, denote $\tilde{g}_0(m_i) = p(m_i | \vartheta_0, f_0)$. Let $\tilde{u}_{it} = \tilde{y}_{it} - \beta_0' \tilde{x}_{i,t-1}$ be the output from the orthogonal forward differencing in (10) evaluated at β_0, and consider the change of variables from $D_i = \left(y_{i,1:T}, x^{P*}_{i,1:T-1}, c_{i0}\right)$ to $\mathcal{D}_i = \left(m_i, \tilde{u}_{i,1:T-d_w}, x^{P*}_{i,1:T-1}, c_{i0}\right)$. Then, individual-specific likelihood function becomes

$$\begin{aligned}
& g_{\mathcal{D}}(\mathcal{D}_i | \vartheta_0, f) \\
= & \prod_t p\left(x^{P*}_{i,t-1} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) p(c_{i0}) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \int \phi(m_i; \lambda_i, \Sigma_i)\, f(\lambda_i)\, d\lambda_i.
\end{aligned}$$

Note that the L_1-norm is preserved under the change of variables, so we have

$$\begin{aligned}
& \left\| g(D_i | \vartheta_0, f) - g(D_i | \vartheta_0, f_0) \right\|_1 \\
= & \left\| g_{\mathcal{D}}(\mathcal{D}_i | \vartheta_0, f) - g_{\mathcal{D}}(\mathcal{D}_i | \vartheta_0, f_0) \right\|_1 \\
= & \int \left| \int \prod_t p\left(x^{P*}_{i,t-1} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) p(c_{i0}) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \right. \\
& \left. \cdot \phi(m_i; \lambda_i, \Sigma_i)\left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| d\mathcal{D}_i \\
= & \int \prod_t p\left(x^{P*}_{i,t-1} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) p(c_{i0}) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \\
& \cdot \left| \int \phi(m_i; \lambda_i, \Sigma_i)\left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i \right| d\mathcal{D}_i.
\end{aligned}$$

After iteratively integrating out
(1) $\int p\left(x^{P*}_{i,t-1} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) dx^{P*}_{i,t-1} = 1$ for $t = T, T-1, \cdots, 2$,
(2)$\int p(\tilde{c}_{i0})\, d\tilde{c}_{i0} = 1$ where $\tilde{c}_{i0} = c_{i0} \setminus w_{i,0:T-1}$,
(3) $\int \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) d\tilde{u}_{it} = 1$ for $t = 1, \cdots, T-d_w$,

we are left with

$$
\begin{aligned}
&\int\left|\int \phi\left(m_i;\lambda_i,\Sigma_i\right)\left(f\left(\lambda_i\right)-f_0\left(\lambda_i\right)\right)d\lambda_i\right| p\left(w_{i,0:T-1}\right)dw_{i,0:T-1}dm_i \qquad (45)\\
\geq&\int\left|\int\left[\int \phi\left(m_i;\lambda_i,\Sigma_i\right)p\left(w_{i,0:T-1}\right)dw_{i,0:T-1}\right]\left(f\left(\lambda_i\right)-f_0\left(\lambda_i\right)\right)d\lambda_i\right|dm_i\\
=&\int\left|\int \bar{\phi}\left(m_i-\lambda_i\right)\left(f\left(\lambda_i\right)-f_0\left(\lambda_i\right)\right)d\lambda_i\right|dm_i\\
=&\left\|\tilde{g}-\tilde{g}_0\right\|_1
\end{aligned}
$$

Define $\mathcal{K}_\delta^*$ such that $\mathcal{K}_\delta=\bar{\phi}*\mathcal{K}_\delta^*$. Then, its Fourier transform is $\hat{\mathcal{K}}_\delta^*=\hat{\mathcal{K}}_\delta\Big/\hat{\bar{\phi}}$. Following the Cauchy–Schwarz inequality,

$$
\begin{aligned}
&\int_{\|z\|_2\leq\mathcal{M}}\|z\|_2^2\left|\left(f-f_0\right)*\mathcal{K}_\delta\left(z\right)\right|dz \qquad (46)\\
\leq&\left(\int_{\|z\|_2\leq\mathcal{M}}\|z\|_2^4\,dz\int_{\|z\|_2\leq\mathcal{M}}\left|\left(f-f_0\right)*\mathcal{K}_\delta\left(z\right)\right|^2dz\right)^{1/2}\\
\leq&\mathcal{M}^{5/2}\left\|\left(f-f_0\right)*\mathcal{K}_\delta\right\|_2\\
=&\mathcal{M}^{5/2}\left\|\left(f-f_0\right)*\left(\bar{\phi}*\mathcal{K}_\delta^*\right)\right\|_2\\
=&\mathcal{M}^{5/2}\left\|\left(\tilde{g}-\tilde{g}_0\right)*\mathcal{K}_\delta^*\right\|_2\\
\leq&\mathcal{M}^{5/2}\left\|\tilde{g}-\tilde{g}_0\right\|_1\left\|\mathcal{K}_\delta^*\right\|_2.
\end{aligned}
$$

Based on the Plancherel theorem,

$$
\begin{aligned}
\left\|\mathcal{K}_\delta^*\right\|_2^2\lesssim\left\|\hat{\mathcal{K}}_\delta^*\right\|_2^2&=\int\left(\frac{\hat{\mathcal{K}}_\delta\left(\xi\right)}{\hat{\bar{\phi}}\left(\xi\right)}\right)^2d\xi \qquad (47)\\
&\lesssim\int_{\|\xi\|_2\leq\frac{1}{\delta}}\left(\hat{\bar{\phi}}\left(\xi\right)\right)^{-2}d\xi\\
&=\int_{\|\xi\|_2\leq\frac{1}{\delta}}\left(\int\phi\left(\xi;0,\Sigma_i^{-1}\right)p\left(w_{i,0:T-1}\right)dw_{i,0:T-1}\right)^{-2}d\xi\\
&\lesssim\int_{\|\xi\|_2\leq\frac{1}{\delta}}\exp\left(\frac{d_w\sigma_0^2}{m_w}\xi^2\right)d\xi\\
&\lesssim\exp\left(\frac{d_w\sigma_0^2}{m_w\delta^2}\right).
\end{aligned}
$$

The second line is obtained by construction as $\hat{K}$ is continuous with $\operatorname{supp}\left(\hat{K}\right)=[-1,\,1]$. The fourth line follows Assumption 5(1,2)—m_w is the lower bound of the eigenvalues of $\sum_t w_{i,t-1}w_{i,t-1}'$, and the upper bound of the eigenvalues of $\sum_t w_{i,t-1}w_{i,t-1}'$ exists due to the boundedness of $w_{i,0:T-1}$.

Combining (45), (46), and (47), the first part of the third term

$$\int_{\|z\|_2 \le \mathcal{M}} \|z\|_2^2 \left|(f-f_0) * \mathcal{K}_\delta(z)\right| dz \lesssim \mathcal{M}^{5/2} \exp\left(\frac{d_w \sigma_0^2}{2m_w \delta^2}\right) \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1 .$$

Third term - second part:

$$\begin{aligned}
&\int_{\|z\|_2 > \mathcal{M}} \|z\|_2^2 \left|(f-f_0) * \mathcal{K}_\delta(z)\right| dz \\
\le& \mathcal{M}^{-2\eta} \int_{\|z\|_2 > \mathcal{M}} \|z\|_2^{2(1+\eta)} \left|(f-f_0) * \mathcal{K}_\delta(z)\right| dz \\
\le& \mathcal{M}^{-2\eta} \int \|z\|_2^{2(1+\eta)} (f+f_0) * \mathcal{K}_\delta(z)\, dz \\
\lesssim& \mathcal{M}^{-2\eta} \int \left(\|z-\upsilon\|_2^{2(1+\eta)} + \|\upsilon\|_2^{2(1+\eta)}\right) (f+f_0)(z-\upsilon)\, \mathcal{K}_\delta(\upsilon)\, dz d\upsilon \\
=& \mathcal{M}^{-2\eta} \left(\int \|\lambda\|_2^{2(1+\eta)} (f+f_0)(\lambda)\, d\lambda + \int \|\upsilon\|_2^{2(1+\eta)} \mathcal{K}_\delta(\upsilon)\, d\upsilon\right) \\
\lesssim& \mathcal{M}^{-2\eta}.
\end{aligned}$$

Note that the $2(1+\eta)$-th moment of $\mathcal{K}_\delta$ exists by construction, the $2(1+\eta)$-th moment of f_0 exists based on Assumption 6(1-e), and the $2(1+\eta)$-th moment of f exists as we consider space $\mathcal{F}$.

In summary: We have

$$W_2^2(f, f_0) \lesssim \delta^2 + \mathcal{M}^{5/2} \exp\left(\frac{d_w \sigma_0^2}{2m_w \delta^2}\right) \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1 + \mathcal{M}^{-2\eta}. \tag{48}$$

When $\left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1 < 1$, we have $\log \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1 < 0$. We can choose

$$\begin{aligned}
\mathcal{M} &= \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1^{-\upsilon_1}, \\
\delta &= \sqrt{\frac{d_w \sigma_0^2}{2m_w}} \left(-\log\left(\mathcal{M}^{\frac{5}{2}+\upsilon_2} \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1\right)\right)^{-1/2} \\
&= \sqrt{\frac{d_w \sigma_0^2}{2m_w \left(1 - \frac{5}{2}\upsilon_1 - \upsilon_1 \upsilon_2\right)}} \left(-\log \left\| g\left(D_i \middle| \vartheta_0, f\right) - g\left(D_i \middle| \vartheta_0, f_0\right)\right\|_1\right)^{-1/2},
\end{aligned}$$

for some $v_1, v_2 > 0$ and $\frac{5}{2}v_1 + v_1 v_2 < 1$. Then, the three terms in (48) become

$$\begin{aligned}
&\delta^2 \lesssim \left(-\log \|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1\right)^{-1}, \\
&\mathcal{M}^{5/2} \exp\left(\frac{d_w \sigma_0^2}{2 m_w \delta^2}\right) \|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1 \\
&\quad = \|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1^{v_1 v_2}, \\
&\mathcal{M}^{-2\eta} = \|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1^{2\eta v_1}.
\end{aligned}$$

The second and third terms are dominated by the first term. Therefore, there exists $C_W > 0$ such that

$$\mathfrak{C}\left(\|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1\right) = C_W \cdot \left(-\log \|g(D_i|\,\vartheta_0, f) - g(D_i|\,\vartheta_0, f_0)\|_1\right)^{-1/2} \geq 0$$

is an increasing function with $\lim_{x\to 0} \mathfrak{C}(x) = 0$ satisfying condition 1-c in Theorem 4.

4. Condition 2 in Theorem 4. After orthogonal forward differencing in (8) and (9), we can estimate

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

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

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

Under the null hypothesis,

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

Assumption 5 ensures the existence of these asymptotic variances. Then,

$$
\begin{aligned}
\mathbb{E}_{\vartheta_0, f_0} \varphi_N(D) &= \mathbb{P}_0^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \text{ or } \left| \hat{\sigma}^2_{GMM} - \sigma_0^2 \right| > \frac{\Delta'}{2} \right) \\
&\leq \mathbb{P}_0^N \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2} \right) + \mathbb{P}_0^N \left(\left| \hat{\sigma}^2_{GMM} - \sigma_0^2 \right| > \frac{\Delta'}{2} \right) \\
&\leq \sum_{j=1}^{d_x} \mathbb{P}_0^N \left(\left| \hat{\beta}_{GMM} - \beta_{0,j} \right| > \frac{\Delta}{2\sqrt{d_x}} \right) + \mathbb{P}_0^N \left(\left| \hat{\sigma}^2_{GMM} - \sigma_0^2 \right| > \frac{\Delta'}{2} \right) \\
&\leq \frac{2 d_x \phi \left(\frac{\Delta}{2\sqrt{d_x}} \Big/ \sqrt{\frac{\sigma_0^2}{N} \Lambda^{-1}_{\min,xx}} \right)}{\frac{\Delta}{2\sqrt{d_x}}} + \frac{2\phi \left(\frac{\Delta'}{2} \Big/ \sqrt{\frac{2\sigma_0^2}{N(T-d_w)-d_x}} \right)}{\frac{\Delta'}{2}} \\
&= \frac{4 d_x^{3/2}}{\Delta} \phi \left(\frac{\Delta}{2} \sqrt{\frac{\Lambda_{\min,xx} N}{d_x \sigma_0^2}} \right) + \frac{4}{\Delta'} \phi \left(\frac{\Delta'}{2} \sqrt{\frac{N(T-d_w)-d_x}{2\sigma_0^2}} \right),
\end{aligned} \tag{49}
$$

where $\Lambda_{\min,xx}$ is the smallest eigenvalue of $\mathbb{E}\big[\sum_t \tilde{x}_{i,t-1} \tilde{x}'_{i,t-1}\big]$. The third line is given by the fact that $\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 > \frac{\Delta}{2}$ implies that $\left| \hat{\beta}_{GMM} - \beta_{0,j} \right| > \frac{\Delta}{2\sqrt{d_x}}$ for at least one $j = 1, \cdots, d_x$. The fourth line follows the bound of the tail of a standard normal distribution in Lemma 32. Under the alternative hypothesis,

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

Then,

$$
\begin{aligned}
\mathbb{E}_{\vartheta, f}\left[1 - \varphi_N(D)\right] &= \mathbb{P}^N_{\vartheta,f} \left(\left\| \hat{\beta}_{GMM} - \beta_0 \right\|_2 \leq \frac{\Delta}{2} \text{ and } \left| \hat{\sigma}^2_{GMM} - \sigma_0^2 \right| \leq \frac{\Delta'}{2} \right) \\
&\leq \mathbb{P}^N_{\vartheta,f} \left(\left\| \hat{\beta}_{GMM} - \beta \right\|_2 > \frac{\Delta}{2} \text{ and } \left| \hat{\sigma}^2_{GMM} - \sigma^2 \right| > \frac{\Delta'}{2} \right) \\
&\leq \mathbb{P}^N_{\vartheta,f} \left(\left\| \hat{\beta}_{GMM} - \beta \right\|_2 > \frac{\Delta}{2} \text{ or } \left| \hat{\sigma}^2_{GMM} - \sigma^2 \right| > \frac{\Delta'}{2} \right) \\
&\leq \frac{4 d_x^{3/2}}{\Delta} \phi \left(\frac{\Delta}{2} \sqrt{\frac{\Lambda_{\min,xx} N}{d_x \sigma^2}} \right) + \frac{4}{\Delta'} \phi \left(\frac{\Delta'}{2} \sqrt{\frac{N(T-d_w)-d_x}{2\sigma^2}} \right).
\end{aligned} \tag{50}
$$

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 (49) and (50) and set

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

which leads to

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

5. 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, $c_3 > 0$, $r > (d_w - 1)/2$, and $\kappa > d_w(d_w-1)$, for sufficiently large $\lambda_* > 0$,

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

where Λ_1 and Λ_{d_w} are the largest and smallest eigenvalues of Ω^{-1}, respectively. Then, we can establish the sieve property in terms of $\bar{\Pi}_f$ based on Theorem 2 in Canale and De Blasi (2017). It further leads to the sieve property in terms of Π_f according to sufficient condition (28) in Remark 19. $\square$

C.3.2 Correlated Random Coefficients: Cross-sectional Homoskedasticity

The proofs of correlated random coefficients models build on Pati *et al.* (2013)'s work on univariate conditional density estimation, and the current proof introduces two major extensions: multivariate conditional density estimation based on location-scale mixture, and deconvolution and dynamic panel data structures. For conditional distributions, let $f(h, c_0) = f(h|c_0)\, q_0(c_0)$, where q_0 is true marginal density of c_0. Then, the induced q_0-integrated L_1-distance is defined as

$$\begin{aligned}
\|f - f_0\|_1 &= \|f(h|c_0)\, q_0(c_0) - f_0(h|c_0)\, q_0(c_0)\|_1 \\
&= \int\left[\int \left|f(\lambda|c_0) - f_0(\lambda|c_0)\right| d\lambda\right] q_0(c_0)\, dc_0,
\end{aligned}$$

the induced q_0-integrated KL divergence is

$$\begin{aligned}
D_{KL}(f_0 \,\|\, f) &= D_{KL}\left(f(h|c_0)\, q_0(c_0) \,\|\, f_0(h|c_0)\, q_0(c_0)\right) \\
&= \int\left[\int f_0(\lambda|c_0)\log\frac{f_0(\lambda|c_0)}{f(\lambda|c_0)} d\lambda\right] q_0(c_0)\, dc_0,
\end{aligned}$$

and the induced second Wasserstein distance is

$$
\begin{aligned}
W_2(f, f_0) &= W_2\left(f(h|c_0)\, q_0(c_0), f_0(h|c_0)\, q_0(c_0)\right) \\
&\leq \left(\int W_2^2\left(f_0(\lambda|c_0), f(\lambda|c_0)\right) q_0(c_0)\, dc_0\right)^{1/2}.
\end{aligned}
$$

Proof. **(Theorem 10)**
The individual-specific likelihood function is characterized as

$$
g(D_i|\vartheta, f) = \prod_t p\left(x_{i,t-1}^{P*}|y_{i,t-1}, c_{i,0:t-2}\right) \int \prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right) f(\lambda_i|c_{i0})\, q_0(c_{i0})\, d\lambda_i.
$$

1. Condition 1-a in Theorem 4. Assumptions 8 and 9(1,2,3-a) ensure the induced q_0-integrated KL property on f, i.e. for all $\epsilon > 0$,

$$
\bar{\Pi}_f\left(f : D_{KL}(f_0 \parallel f) < \epsilon\right) > 0.
$$

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 case in Appendix C.3.1, except for changing $f(\lambda_i)$ and $p(c_{i0})$ to $f(\lambda_i|c_{i0})$ and $q_0(c_{i0})$, respectively, and modifying (43) and (44) for the second term: let

$$
M_{2c}(c_0) = \int \|\lambda\|_2^2 f_\lambda(\lambda|c_0)\, d\lambda,
$$

then, as we consider space $\mathcal{F}$, for some $M_2 > 0$,

$$
\int M_{2c}(c_0)\, q_0(c_0)\, dc_0 \leq M_2. \tag{51}
$$

Now (43) becomes

$$
\begin{aligned}
&\int \frac{\phi\left(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)\right) f(\lambda_i|c_{i0})}{\int \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i(\sigma_0^2)\right) f(\lambda_i|c_{i0})\, d\lambda_i} \|\lambda_i\|_2^2\, d\lambda_i \\
&\lesssim \|m_i(\beta_0)\|_2^2 + \mathrm{tr}\left(\Sigma_i(\sigma_0^2)\right) + M_{2c}(c_{i0}) \lesssim \sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2\right) + M_{2c}(c_{i0}) + 1,
\end{aligned}
$$

and (44) turns to be

$$\int \frac{\phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right)}{\int \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) f\left(\lambda_i\right) d\lambda_i} \log \frac{\prod_t \phi\left(y_{it}; \beta_0' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma_0^2\right)}{\prod_t \phi\left(y_{it}; \beta' x_{i,t-1} + \lambda_i' w_{i,t-1}, \sigma^2\right)} d\lambda_i$$
$$\lesssim \|\beta - \beta_0\|_2 \left[\sum_t \left(y_{it}^2 + \|x_{i,t-1}\|_2^2\right) + M_{2c}\left(c_{i0}\right) + 1\right].$$

Consider (51), once we integrate out c_{i0}, there still exists a constant $C_\beta > 0$ as on page A-23.

2. Condition 1-b in Theorem 4. Similar to the random coefficients case in Appendix C.3.1, except changing $f\left(\lambda_i\right)$ and $p\left(c_{i0}\right)$ to $f\left(\lambda_i\left|c_{i0}\right.\right)$ and $q_0\left(c_{i0}\right)$, respectively.

3. Condition 1-c in Theorem 4. This part of the proof is similar to the random coefficients case in Appendix C.3.1. Denote $W_2\left(f_0, f|c_0\right) = W_2\left(f_0\left(\lambda|c_0\right), f\left(\lambda|c_0\right)\right).$ According to the q_0-induced Wasserstein metric,

$$\begin{aligned} W_2\left(f, f_0\right) &= W_2\left(f\left(h|c_0\right) q_0\left(c_0\right), f_0\left(h|c_0\right) q_0\left(c_0\right)\right) \\ &\leq \left(\int W_2^2\left(f_0\left(\lambda|c_0\right), f\left(\lambda|c_0\right)\right) q_0\left(c_0\right) dc_0\right)^{1/2}. \end{aligned}$$

Following the triangular inequality,

$$\begin{aligned} &\int W_2^2\left(f, f_0|c_0\right) q_0\left(c_0\right) dc_0 \\ &\leq \int \left(W_2^2\left(f, f * \mathcal{K}_\delta|c_0\right) + W_2^2\left(f_0, f_0 * \mathcal{K}_\delta|c_0\right) + W_2^2\left(f * \mathcal{K}_\delta, f_0 * \mathcal{K}_\delta|c_0\right)\right) q_0\left(c_0\right) dc_0. \end{aligned}$$

First and second terms: Consider coupling $(\lambda, \lambda + \upsilon)$, where the marginal distributions of λ and υ are $f\left(\lambda|c_0\right)$ and $\mathcal{K}_\delta$, respectively. Then, as the second moment of K is bounded, we have the first two terms

$$\int \left(W_2^2\left(f, f * \mathcal{K}_\delta|c_0\right) + W_2^2\left(f_0, f_0 * \mathcal{K}_\delta|c_0\right)\right) q_0\left(c_0\right) dc_0 \lesssim \delta^2.$$

Third term: Let z be a generic variable. According to Theorem 6.15 in Villani (2009),

$$\begin{aligned} &\int W_2^2\left(f * \mathcal{K}_\delta, f_0 * \mathcal{K}_\delta|c_0\right) q_0\left(c_0\right) dc_0 \\ &\lesssim \int \|z\|_2^2 \left|\left(f - f_0\right) * \mathcal{K}_\delta\left(z|c_0\right)\right| q_0\left(c_0\right) dz dc_0 \\ &= \int_{\|z\|_2 \leq \mathcal{M}} \|z\|_2^2 \left|\left(f - f_0\right) * \mathcal{K}_\delta\left(z|c_0\right)\right| q_0\left(c_0\right) dz dc_0 \\ &\quad + \int_{\|z\|_2 > \mathcal{M}} \|z\|_2^2 \left|\left(f - f_0\right) * \mathcal{K}_\delta\left(z|c_0\right)\right| q_0\left(c_0\right) dz dc_0, \end{aligned}$$

for some large $\mathcal{M} > 0$ that could depend on $\left\|g\left(D_i\right|\vartheta_0, f) - g\left(D_i\right|\vartheta_0, f_0)\right\|_1$.

Third term - first part: Define

$$m_i = \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1} \sum_t w_{i,t-1} \left(y_{it} - \beta_0' x_{i,t-1}\right), \quad \Sigma_i = \sigma_0^2 \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1}.$$

Conditional on c_{i0}, we have

$$m_i = \lambda_i + \bar{u}_i, \quad \bar{u}_i \sim \phi_{\Sigma_i}\left(\bar{u}_i\right) = \phi\left(m_i; 0, \Sigma_i\right),$$

and the (conditional) distribution of m_i is

$$\tilde{g}\left(m_i | c_{i0}\right) = p\left(m_i | c_{i0}, \vartheta_0, f\right) = f * \phi_{\Sigma_i}\left(m_i | c_{i0}\right),$$

Also, denote $\tilde{g}_0\left(m_i | c_{i0}\right) = f_0 * \phi_{\Sigma_i}\left(m_i | c_{i0}\right)$. Again, consider the change of variables from $D_i = \left(y_{i,1:T}, x_{i,1:T-1}^{P*}, c_{i0}\right)$ to $\mathcal{D}_i = \left(m_i, \tilde{u}_{i,1:T-d_w}, x_{i,1:T-1}^{P*}, c_{i0}\right)$. Then, individual-specific likelihood function becomes

$$\begin{aligned}
& g_{\mathcal{D}}\left(\mathcal{D}_i | \vartheta_0, f\right) \\
= & \prod_t p\left(x_{i,t-1}^{P*} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \int \phi\left(m_i; \lambda_i, \Sigma_i\right) f\left(\lambda_i | c_{i0}\right) q_0\left(c_{i0}\right) d\lambda_i.
\end{aligned}$$

Note that the L_1-norm is preserved under the change of variables, so we have

$$\begin{aligned}
& \left\| g\left(D_i | \vartheta_0, f\right) - g\left(D_i | \vartheta_0, f_0\right)\right\|_1 \\
= & \left\| g_{\mathcal{D}}\left(\mathcal{D}_i | \vartheta_0, f\right) - g_{\mathcal{D}}\left(\mathcal{D}_i | \vartheta_0, f_0\right)\right\|_1 \\
= & \int \left| \int \prod_t p\left(x_{i,t-1}^{P*} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \right. \\
& \left. \cdot \phi\left(m_i; \lambda_i, \Sigma_i\right) \left(f\left(\lambda_i | c_{i0}\right) - f_0\left(\lambda_i | c_{i0}\right)\right) d\lambda_i \right| q_0\left(c_{i0}\right) d\mathcal{D}_i \\
= & \int \prod_t p\left(x_{i,t-1}^{P*} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) \prod_{t=1}^{T-d_w} \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) \\
& \cdot \left| \int \phi\left(m_i; \lambda_i, \Sigma_i\right) \left(f\left(\lambda_i | c_{i0}\right) - f_0\left(\lambda_i | c_{i0}\right)\right) d\lambda_i \right| q_0\left(c_{i0}\right) d\mathcal{D}_i.
\end{aligned}$$

After iteratively integrating out
(1) $\int p\left(x_{i,t-1}^{P*} | m_i, \tilde{u}_{i,1:T-d_w}, c_{i,0:t-2}\right) dx_{i,t-1}^{P*} = 1$ for $t = T, T-1, \cdots, 2$,
(2) $\int \phi\left(\tilde{u}_{it}; 0, \sigma_0^2\right) d\tilde{u}_{it} = 1$ for $t = 1, \cdots, T - d_w$,

we are left with

$$\begin{aligned}
&\int\left|\int\phi\left(m_i;\lambda_i,\Sigma_i\right)\left(f\left(\lambda_i|c_{i0}\right)-f_0\left(\lambda_i|c_{i0}\right)\right)d\lambda_i\right|q_0\left(c_{i0}\right)dc_{i0}dm_i \qquad (52)\\
=&\int\left\|\tilde{g}\left(\cdot|c_{i0}\right)-\tilde{g}_0\left(\cdot|c_{i0}\right)\right\|_1 q_0\left(c_{i0}\right)dc_{i0}.
\end{aligned}$$

Given $w_{i,0:T-1}$ satisfying Assumptions 5(2) and 8, we define $\mathcal{K}^*_{\delta,w}$ such that $\mathcal{K}_\delta=\phi_{\Sigma_i}*\mathcal{K}^*_{\delta,w}$, where w in the subscript indicates that $\mathcal{K}^*_{\delta,w}$ depends on $w=w_{i,0:T-1}$. Then, the Fourier transform of $\mathcal{K}^*_{\delta,w}$ is $\hat{\mathcal{K}}^*_{\delta,w}=\hat{\mathcal{K}}_\delta\Big/\hat{\phi}_{\Sigma_i}$. Following the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\int_{\|z\|_2\le\mathcal{M}}\|z\|_2^2\left|(f-f_0)*\mathcal{K}_\delta\left(z|c_0\right)\right|dz \qquad (53)\\
\le&\left(\int_{\|z\|_2\le\mathcal{M}}\|z\|_2^4\,dz\int_{\|z\|_2\le\mathcal{M}}\left|(f-f_0)*\mathcal{K}_\delta\left(z|c_0\right)\right|^2dz\right)^{1/2}\\
\le&\mathcal{M}^{5/2}\left\|(f-f_0)*\mathcal{K}_\delta\left(\cdot|c_0\right)\right\|_2\\
=&\mathcal{M}^{5/2}\left\|(f-f_0)*\left(\phi_{\Sigma_i}*\mathcal{K}^*_{\delta,w}\right)\left(\cdot|c_0\right)\right\|_2\\
=&\mathcal{M}^{5/2}\left\|\left(\tilde{g}\left(\cdot|c_0\right)-\tilde{g}_0\left(\cdot|c_0\right)\right)*\mathcal{K}^*_{\delta,w}\left(\cdot|w\right)\right\|_2\\
\le&\mathcal{M}^{5/2}\left\|\tilde{g}\left(\cdot|c_0\right)-\tilde{g}_0\left(\cdot|c_0\right)\right\|_1\left\|\mathcal{K}^*_{\delta,w}\left(\cdot|w\right)\right\|_2.
\end{aligned}$$

Based on the Plancherel theorem,

$$\begin{aligned}
\left\|\mathcal{K}^*_{\delta,w}\left(\cdot|w\right)\right\|_2^2\lesssim\left\|\hat{\mathcal{K}}^*_{\delta,w}\left(\cdot|w\right)\right\|_2^2=&\int\left(\frac{\hat{\mathcal{K}}_\delta(\xi)}{\hat{\phi}_{\Sigma_i}(\xi)}\right)^2d\xi \qquad (54)\\
\lesssim&\int_{\|\xi\|_2\le\frac{1}{\delta}}\left(\hat{\phi}_{\Sigma_i}(\xi)\right)^{-2}d\xi\\
\lesssim&\int_{\|\xi\|_2\le\frac{1}{\delta}}\exp\left(\frac{d_w\sigma_0^2}{m_w}\xi^2\right)d\xi\\
\lesssim&\exp\left(\frac{d_w\sigma_0^2}{m_w\delta^2}\right).
\end{aligned}$$

The second line is obtained by construction as $\hat{K}$ is continuous with supp $\left(\hat{K}\right)=[-1,\ 1]$. The fourth line follows Assumptions 5(2) and 8—m_w is the lower bound of the eigenvalues of $\sum_t w_{i,t-1}w'_{i,t-1}$, and the upper bound of the eigenvalues of $\sum_t w_{i,t-1}w'_{i,t-1}$ exists due to the compactedness of $\mathcal{C}$. Combining (52), (53), and (54), the first part of the third term

$$\int_{\|z\|_2\le\mathcal{M}}\|z\|_2^2\left|(f-f_0)*\mathcal{K}_\delta\left(z|c_0\right)\right|q_0\left(c_0\right)dzdc_0\lesssim\mathcal{M}^{5/2}\exp\left(\frac{d_w\sigma_0^2}{2m_w\delta^2}\right)\left\|g\left(D_i|\,\vartheta_0,f\right)-g\left(D_i|\,\vartheta_0,f_0\right)\right\|_1.$$

Third term - second part:

$$
\begin{aligned}
&\int_{\|z\|_2>\mathcal{M}} \|z\|_2^2 \left|(f-f_0)*\mathcal{K}_\delta(z|c_0)\right| q_0(c_0)\, dz dc_0 \\
\leq& \mathcal{M}^{-2\eta}\int_{\|z\|_2>\mathcal{M}} \|z\|_2^{2(1+\eta)} \left|(f-f_0)*\mathcal{K}_\delta(z|c_0)\right| q_0(c_0)\, dz dc_0 \\
\leq& \mathcal{M}^{-2\eta}\int \|z\|_2^{2(1+\eta)} (f+f_0)*\mathcal{K}_\delta(z|c_0)\, q_0(c_0)\, dz dc_0 \\
\lesssim& \mathcal{M}^{-2\eta}\int \left(\|z-\upsilon\|_2^{2(1+\eta)}+\|\upsilon\|_2^{2(1+\eta)}\right)(f+f_0)(z-\upsilon|c_0)\,\mathcal{K}_\delta(\upsilon)\, q_0(c_0)\, dz d\upsilon dc_0 \\
=& \mathcal{M}^{-2\eta}\left(\int \|\lambda\|_2^{2(1+\eta)}(f+f_0)(\lambda|c_0)\, q_0(c_0)\, d\lambda dc_0+\int \|\upsilon\|_2^{2(1+\eta)}\mathcal{K}_\delta(\upsilon)\, d\upsilon\right) \\
\lesssim& \mathcal{M}^{-2\eta}.
\end{aligned}
$$

Note that the $2(1+\eta)$-th moment of $\mathcal{K}_\delta$ exists by construction, the unconditional $2(1+\eta)$-th moment of f_0 exists based on Assumption 9(1-e), and the unconditional $2(1+\eta)$-th moment of f exists as we consider space $\mathcal{F}$.

In summary: We have

$$
\begin{aligned}
W_2(f,f_0) &= W_2\left(f(h|c_0)\,q_0(c_0), f_0(h|c_0)\,q_0(c_0)\right) \\
&\leq \left(\int W_2^2(f_0,f|c_0)\, q_0(c_0)\, dc_0\right)^{1/2} \\
&\lesssim \left(\delta^2+\mathcal{M}^{5/2}\exp\left(\frac{d_w\sigma_0^2}{2m_w\delta^2}\right)\left\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\right\|_1+\mathcal{M}^{-2\eta}\right)^{1/2}.
\end{aligned} \tag{55}
$$

Similar to the random coefficients case in Appendix C.3.1, we can choose

$$
\begin{aligned}
\mathcal{M} &= \left\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\right\|_1^{-\upsilon_1}, \\
\delta &= \sqrt{\frac{d_w\sigma_0^2}{2m_w\left(1-\frac{5}{2}\upsilon_1-\upsilon_1\upsilon_2\right)}}\left(-\log\left\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\right\|_1\right)^{-1/2},
\end{aligned}
$$

for some $\upsilon_1,\upsilon_2>0$ and $\frac{5}{2}\upsilon_1+\upsilon_1\upsilon_2<1$. Then, last two terms in (55) are dominated by the first term. Therefore, there exists $C_W>0$ such that

$$
\mathfrak{C}\left(\left\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\right\|_1\right)=C_W\cdot\left(-\log\left\|g(D_i|\,\vartheta_0,f)-g(D_i|\,\vartheta_0,f_0)\right\|_1\right)^{-1/2}\geq 0
$$

is an increasing function with $\lim_{x\to 0}\mathfrak{C}(x)=0$ satisfying condition 1-c in Theorem 4.

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

5. 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. Assumption 9(2) resembles the random coefficients 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), Assumption 9(2) here imposes weaker tail conditions on G_0 and hence is able to accommodate multivariate-normal-inverse-Wishart components. Assumption 9(3-b,c) on the stick breaking process directly follows Remark 5.12 and Lemma 5.15 in Pati *et al.* (2013). □

C.4 Density Forecasts: General Semiparametric Model

Let $p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)$ be the individual-specific likelihood of $y_{i,1:T}$, $p\left(y_{i,T+1}\left|h_i, \vartheta, D_i\right.\right)$ be 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 (3), and $\mathfrak{A}\left(h_i, \vartheta, D_i\right) = \mathbb{E}\left[y_{i,T+1}^2\left|h_i, \vartheta, D_i\right.\right] p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)$.

Theorem 21. (Density Forecasts: General Semiparametric Model) *Given i, suppose we have:*

1. *Posterior consistency: conditions in Theorem 4.*
2. *Distribution of individual heterogeneity: For some $M_{\lambda,i}, M_{2,i} > 0$:*
   (a) *f_0 is bounded above by $M_{\lambda,i}$.*
   (b) *$0 < \mathbb{E}_{f_0}\left[\left.\|h_i\|_2^2\right| c_{i0}\right] \le M_{2,i}$.*
   (c) *If f is a conditional distribution, $q_0\left(c_{i0}\right)$ is continuous, and $q_0\left(c_{i0}\right) > 0$ for all $c_{i0} \in \mathcal{C}$.*
3. *Likelihood and predictive distribution: For some $M_{l,i}, M_{\mathfrak{A},\mathrm{i}} > 0$:*
   (a) *$p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)$ is continuous in h_i, and $0 < p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right) \le M_{l,i}$.*
   (b) *There exists $\delta'_\vartheta > 0$ such that for all $\|\vartheta - \vartheta_0\|_2 < \delta'_\vartheta$, $\mathfrak{A}\left(h_i, \vartheta, D_i\right)$ is continuous in h_i and bounded by $M_{\mathfrak{A},\mathrm{i}}$.*
4. *Differences:*
   For $z = l, p, h, \mathfrak{A}$, there exist increasing functions $C_{z,i}(\cdot): \mathbb{R}_{\ge 0} \mapsto \mathbb{R}_{\ge 0}$ with $\lim_{x\to 0} C_{z,i}(x) = 0$ such that:
   (a) *$\int \left|p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right) - p\left(y_{i,1:T}\left|h_i, \vartheta_0, D_i \backslash y_{i,1:T}\right.\right)\right| dh_i \le C_{l,i}\left(\|\vartheta - \vartheta_0\|_2\right)$.*
   (b) *$\left\|p\left(y_{i,T+1}\left|h_i, \vartheta, D_i\right.\right) - p\left(y_{i,T+1}\left|h_i, \vartheta_0, D_i\right.\right)\right\|_1 \le C_{p,i}\left(\|\vartheta - \vartheta_0\|_2\right)$.*
   (c) *$\left\|p\left(y_{i,T+1}\left|h_i, \vartheta, D_i\right.\right) - p\left(y_{i,T+1}\left|\tilde{h}_i, \vartheta, D_i\right.\right)\right\|_1 \le C_{h,i}\left(\left\|h_i - \tilde{h}_i\right\|_2\right)$.*
   (d) *$\int \left|\mathfrak{A}\left(h_i, \vartheta, D_i\right) - \mathfrak{A}\left(h_i, \vartheta_0, D_i\right)\right| dh_i \le C_{\mathfrak{A},i}\left(\|\vartheta - \vartheta_0\|_2\right)$.*

All quantities with subscript i can depend on D_i. Then, density forecasts converge to the oracle, i.e. given i, for all $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(W_2\left(f_{i,T+1}^{cond}, f_{i,T+1}^{oracle}\right) < \epsilon \,\middle|\, D\right) \to 1,$$

in probability with respect to the true DGP.

Proof. **(Theorem 21)**

According to Theorem 5.11 in Santambrogio (2015), convergence in the W_2 metric is equivalent to weak convergence plus convergence of the second moment. Thus, the posterior consistency conditions in Theorem 4 implies that for all continuous bounded functions $\psi(\cdot)$, and $\epsilon_\psi, \epsilon_2 > 0$, as $N \to \infty$, if f is an unconditional distribution,

$$\mathbb{P}\left(\left|\int \psi(h_i)(f(h_i) - f(h_i))\, dh_i\right| < \epsilon_\psi \middle| D\right) \to 1, \tag{56}$$
$$\mathbb{P}\left(\left|\int \|h_i\|_2^2 (f(h_i) - f(h_i))\, dh_i\right| < \epsilon_2 \middle| D\right) \to 1,$$

if f is a conditional distribution,

$$\mathbb{P}\left(\left|\int \psi(h_i, c_{i0})(f(h_i|c_{i0}) - f(h_i|c_{i0}))\, q_0(c_{i0})\, dh_i dc_{i0}\right| < \epsilon_\psi \middle| D\right) \to 1, \tag{57}$$
$$\mathbb{P}\left(\left|\int \left(\|h_i\|_2^2 + \|c_{i0}\|_2^2\right)(f(h_i|c_{i0}) - f(h_i|c_{i0}))\, q_0(c_{i0})\, dh_i dc_{i0}\right| < \epsilon_2 \middle| D\right) \to 1.$$

All above convergence results are in probability with respect to the true DGP. Also, to prove the convergence of density forecasts to the oracle in the W_2 metric, it is equivalent to prove that given i, for all continuous bounded functions $\psi(\cdot)$, and $\epsilon_\psi, \epsilon_2 > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\left|\int \psi(y)\left(f_{i,T+1}^{cond}(y|\vartheta, f, D_i) - f_{i,T+1}^{oracle}(y|D_i)\right) dy\right| < \epsilon_\psi \middle| D\right) \to 1,$$
$$\mathbb{P}\left(\left|\int y^2\left(f_{i,T+1}^{cond}(y|\vartheta, f, D_i) - f_{i,T+1}^{oracle}(y|D_i)\right) dy\right| < \epsilon_2 \middle| D\right) \to 1,$$

in probability with respect to the true DGP. Let $\tilde{\psi}(y)$ be either $\psi(y)$ or y^2. Following the definitions in Sections 2.2 and 3.3,

$$\begin{aligned}
&\left|\int \tilde{\psi}(y)\left(f_{i,T+1}^{cond}(y|\vartheta, f, D_i) - f_{i,T+1}^{oracle}(y|D_i)\right) dy\right| \\
=&\left|\int \tilde{\psi}(y)\left(\int p(y|\, h_i, \vartheta, D_i)\, p(h_i\,|\vartheta, f, D_i)\, dh_i - \int p(y|\, h_i, \vartheta_0, D_i)\, p(h_i\,|\vartheta_0, f_0, D_i)\, dh_i\right) dy\right| \\
=&\left|\int \tilde{\psi}(y)\left(\frac{\int p(y|\, h_i, \vartheta, D_i)\, p(y_{i,1:T}\,|h_i, \vartheta, D_i \backslash y_{i,1:T})\, f(h_i|c_{i0})\, dh_i}{\int p(y_{i,1:T}\,|h_i, \vartheta, D_i \backslash y_{i,1:T})\, f(h_i|c_{i0})\, dh_i}\right.\right. \\
&\left.\left.-\frac{\int p(y|\, h_i, \vartheta_0, D_i)\, p(y_{i,1:T}\,|h_i, \vartheta_0, D_i \backslash y_{i,1:T})\, f_0(h_i|c_{i0})\, dh_i}{\int p(y_{i,1:T}\,|h_i, \vartheta_0, D_i \backslash y_{i,1:T})\, f_0(h_i|c_{i0})\, dh_i}\right) dy\right|.
\end{aligned} \tag{58}$$

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, $f(h_i|c_{i0}) = f(h_i)$. Let

$$A_i = \int p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right) f(h_i|c_{i0})\, dh_i,$$

$$B_i\left(y;\tilde{\psi}\right) = \tilde{\psi}(y) \cdot \int p\left(y|\, h_i, \vartheta, D_i\right) p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right) f(h_i|c_{i0})\, dh_i,$$

with A_{i0} and $B_{i0}\left(y;\tilde{\psi}\right)$ being the counterparts for the oracle predictor. Then,

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

and it is sufficient to establish the following six statements (Lemmas 22, 23, 24, 25, 27, and 28). Note that A_i and A_{i0} are non-negative by definition, so we get rid of $|\cdot|$ for these terms. □

Lemma 22. *Suppose conditions 1, 2-a, 2-c, 3-a, and 4-a in Theorem 21 hold, then for all $\epsilon > 0$, as $N \to \infty$,*

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

in probability with respect to the true DGP.

Proof. Note that

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

First term: Theorem 21(3-a) ensures that $p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)$ is a continuous bounded function of h_i. If f is an unconditional distribution, let $\psi(h_i) = p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)$ and $\epsilon_\psi = \epsilon/2$. Then, the posterior consistency of f implies (56), which in turn implies the convergence of the first term. If f is a conditional distribution, let $\psi(h_i, c_{i0}) = p\left(y_{i,1:T}\left|h_i, \vartheta, D_i \backslash y_{i,1:T}\right.\right)/\, q_0(c_{i0})$ and $\epsilon_\psi = \epsilon/2$. Given Theorem 21(2-c), $\psi(h_i, c_{i0})$ is a continuous bounded function of (h_i, c_{i0}). Again, the posterior consistency of f implies (57), which in turn implies the convergence of the first term.

Combining both cases, we prove that as $N \to \infty$, the first term

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

in probability with respect to the true DGP.

Second term:

$$\begin{aligned}
&\int \left|p\left(y_{i,1:T}\left|h_i, \vartheta,\, D_i \backslash\, y_{i,1:T}\right.\right) - p\left(y_{i,1:T}\left|h_i, \vartheta_0,\, D_i \backslash\, y_{i,1:T}\right.\right)\right| f_0\left(h_i|c_{i0}\right) dh_i \\
\leq & M_{\lambda,i} \int \left|p\left(y_{i,1:T}\left|h_i, \vartheta,\, D_i \backslash\, y_{i,1:T}\right.\right) - p\left(y_{i,1:T}\left|h_i, \vartheta_0,\, D_i \backslash\, y_{i,1:T}\right.\right)\right| dh_i \\
\leq & M_{\lambda,i} C_{l,i}\left(\|\vartheta - \vartheta_0\|_2\right).
\end{aligned} \tag{59}$$

The second and third lines follow conditions 2-a and 4-a in Theorem 21, respectively. Since ϑ enjoys posterior consistency, as $N \to \infty$, the second term

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

in probability with respect to the true DGP. □

Lemma 23. *Let $\tilde{\psi}(y) = \psi(y)$. Suppose conditions 1, 2-a, 2-c, 3-a, 4-a, 4-b, and 4-c in Theorem 21 hold, then for all $\epsilon > 0$, as $N \to \infty$,*

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

in probability with respect to the true DGP.

Proof. $\psi(y)$ is a continuous bounded function. Suppose $|\psi(y)| \leq M_\psi$, and M_ψ could depend on the specific ψ. Note that

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

First term:

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

$\mathbb{E}\left[\psi(y)\left|h_i, \vartheta, D_i\right.\right]$ is bounded by definition,

$$\begin{aligned}
\mathbb{E}\left[\psi(y)\left|h_i, \vartheta, D_i\right.\right] &= \int \psi(y)\, p\left(y|\, h_i, \vartheta, D_i\right) dy \\
&\leq M_\psi \int p\left(y|\, h_i, \vartheta, D_i\right) dy = M_\psi .
\end{aligned}$$

Moreover, $\mathbb{E}\left[\psi(y)\left|h_i, \vartheta, D_i\right.\right]$ is continuous in h_i based on Theorem 21(4-c): for all $\epsilon > 0$, there exists $\delta_h = C_{h,i}^{-1}\left(\epsilon / M_\psi\right) > 0$ such that for all $\left\|h_i - \tilde{h}_i\right\|_2 < \delta_h$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\psi(y)\left|h_i, \vartheta, D_i\right.\right] - \mathbb{E}\left[\psi(y)\left|\tilde{h}_i, \vartheta, D_i\right.\right]\right| \\
\leq &\int |\psi(y)| \left|p\left(y|\, h_i, \vartheta, D_i\right) - p\left(y|\, \tilde{h}_i, \vartheta, D_i\right)\right| dy \\
\leq &M_\psi \int \left|p\left(y|\, h_i, \vartheta, D_i\right) - p\left(y|\, \tilde{h}_i, \vartheta, D_i\right)\right| dy \\
\leq &M_\psi C_{h,i}\left(\left\|h_i - \tilde{h}_i\right\|_2\right) < \epsilon.
\end{aligned}$$

Then, $\mathbb{E}\left[\psi(y)\left|h_i, \vartheta, D_i\right.\right] p\left(y_{i,1:T}\left|h_i, \vartheta,\, D_i \backslash\, y_{i,1:T}\right.\right)$ is a continuous and bounded function of h_i, and we can proceed as the proof of the first term in Lemma 22.

Second term: The second term can be reduced to (59) in the proof of Lemma 22,

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

Third term: Conditions 3-a and 4-b in Theorem 21 bound the third term,

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

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

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

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

Lemma 24. *Let $\tilde{\psi}(y) = y^2$. Suppose conditions 1, 2-a, 2-c, 3-b, and 4-d in Theorem 21 hold, then for all $\epsilon > 0$, as $N \to \infty$,*

$$\mathbb{P}\left(\left|\int \left(B_i\left(y; y^2\right) - B_{i0}\left(y; y^2\right)\right) dy\right| < \epsilon \right| D\right) \to 1,$$

in probability with respect to the true DGP.

Proof. Note that

$$\begin{aligned}
&\left|\int \left(B_i\left(y; y^2\right) - B_{i0}\left(y; y^2\right)\right) dy\right| \\
=&\left|\int y^2 p(y|\, h_i, \vartheta, D_i)\, p(y_{i,1:T}\,|h_i, \vartheta, D_i \backslash\, y_{i,1:T})\, f(h_i|c_{i0})\, dh_i dy\right. \\
&\left. - \int y^2 p(y|\, h_i, \vartheta_0, D_i)\, p(y_{i,1:T}\,|h_i, \vartheta_0, D_i \backslash\, y_{i,1:T})\, f_0(h_i|c_{i0})\, dh_i dy\right| \\
=&\left|\int \left(\mathfrak{A}(h_i, \vartheta, D_i)\, f(h_i|c_{i0}) - \mathfrak{A}(h_i, \vartheta_0, D_i)\, f_0(h_i|c_{i0})\right) dh_i\right| \\
\leq&\left|\int \mathfrak{A}(h_i, \vartheta, D_i)\left(f(h_i|c_{i0}) - f_0(h_i|c_{i0})\right) dh_i\right| \\
&+ \int \left|\mathfrak{A}(h_i, \vartheta, D_i) - \mathfrak{A}(h_i, \vartheta_0, D_i)\right| f_0(h_i|c_{i0})\, dh_i.
\end{aligned}$$

First term: According to Theorem 21(3-b), for $\|\vartheta - \vartheta_0\|_2 < \delta'_\vartheta$, $\mathfrak{A}(h_i, \vartheta, D_i)$ is continuous in h_i and bounded by $M_{\mathfrak{A},i}$. Then, we can proceed as the proof of the first term in Lemma 22. Since ϑ enjoys posterior consistency, as $N \to \infty$, the first term

$$\mathbb{P}\left(\left|\int \mathfrak{A}(h_i, \vartheta, D_i)\left(f(h_i|c_{i0}) - f_0(h_i|c_{i0})\right) dh_i\right| < \frac{\epsilon}{2}\right| D\right) \to 1,$$

in probability with respect to the true DGP.

Second term:

$$\begin{aligned}
&\int \left|\mathfrak{A}(h_i, \vartheta, D_i) - \mathfrak{A}(h_i, \vartheta_0, D_i)\right| f_0(h_i|c_{i0})\, dh_i \\
\leq& M_{\lambda,i} \int \left|\mathfrak{A}(h_i, \vartheta, D_i) - \mathfrak{A}(h_i, \vartheta_0, D_i)\right| dh_i. \\
\leq& M_{\lambda,i} C_{\mathfrak{A},i}\left(\|\vartheta - \vartheta_0\|_2\right).
\end{aligned}$$

The second and third lines follow conditions 2-a and 4-d in Theorem 21, respectively. Since ϑ enjoys posterior consistency, as $N \to \infty$, the second term

$$\mathbb{P}\left(\int \left| \mathfrak{A}\left(h_i, \vartheta, D_i\right) - \mathfrak{A}\left(h_i, \vartheta_0, D_i\right) \right| f_0\left(h_i | c_{i0}\right) dh_i < \frac{\epsilon}{2} \middle| D \right) \to 1,$$

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

Lemma 25. *Suppose conditions 2-b, 2-c, and 3-a in Theorem 21 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 21(2-b) ensures the existence of the (conditional) second moment. Following Chebyshev's inequality, let d_h be the dimension of h_i, we have

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

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

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

Based on Theorem 21(3-a) and the extreme value theorem, the minimum exists and is positive. Intuitively, since the domains of h_i in $p\left(y_{i,1:T} \left| h_i, \vartheta_0, D_i \backslash y_{i,1:T}\right.\right)$ and $f_0\left(h_i | c_{i0}\right)$ overlap, the integral is bounded below by some positive number. $\square$

Remark 26. Moreover, together with Lemma 22, let $\epsilon = \underline{A}_i/2$, then as $N \to \infty$,

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

in probability with respect to the true DGP.

Lemma 27. *Let $\tilde{\psi}(y) = \psi(y)$. Suppose condition 3-a in Theorem 21 holds, then there exists $\bar{B}_{\psi,i} > 0$, which could depend on the specific ψ, such that*

$$\left| \int B_{i0}\left(y; \psi\right) dy \right| < \bar{B}_{\psi,i}.$$

Proof. $\psi(y)$ is a continuous bounded function. Suppose $|\psi(y)| \leq M_\psi$, and M_ψ could depend on the specific ψ. We have

$$\begin{aligned}
&\left|\int B_{i0}(y;\psi)\,dy\right| \\
&= \left|\int \psi(y)\, p\left(y\middle|\, h_i,\vartheta_0,D_i\right) p\left(y_{i,1:T}\,\middle|h_i,\vartheta_0,\,D_i\backslash\, y_{i,1:T}\right) f_0\left(h_i|c_{i0}\right) dh_i dy\right| \\
&\leq M_\psi \int \left[\int p\left(y\middle|\, h_i,\vartheta_0,D_i\right) dy\right] p\left(y_{i,1:T}\,\middle|h_i,\vartheta_0,\,D_i\backslash\, y_{i,1:T}\right) f_0\left(h_i|c_{i0}\right) dh_i \\
&= M_\psi \int p\left(y_{i,1:T}\,\middle|h_i,\vartheta_0,\,D_i\backslash\, y_{i,1:T}\right) f_0\left(h_i|c_{i0}\right) dh_i \\
&\leq M_\psi M_{l,i} \int f_0\left(h_i|c_{i0}\right) dh_i \\
&= M_\psi M_{l,i} \overset{\text{def}}{=} \bar{B}_{\psi,i}.
\end{aligned}$$

The second to last line follows condition 3-a in Theorem 21. □

Lemma 28. *Let $\tilde{\psi}(y) = y^2$. Suppose conditions 1 and 3-b in Theorem 21 hold, then there exists $\bar{B}_{2,i} > 0$ such that as $N \to \infty$,*

$$\mathbb{P}\left(\int B_{i0}\left(y;y^2\right) dy < \bar{B}_{2,i} \,\middle|\, D\right) \to 1,$$

in probability with respect to the true DGP.[8]

Proof. According to Theorem 21(3-b), for $\|\vartheta - \vartheta_0\|_2 < \delta'_\vartheta$, $\mathfrak{A}(h_i,\vartheta,D_i)$ is bounded $M_{\mathfrak{A},i}$. Then, we have

$$\begin{aligned}
&\int B_{i0}\left(y;y^2\right) dy \\
&= \int y^2 p\left(y\middle|\, h_i,\vartheta_0,D_i\right) p\left(y_{i,1:T}\,\middle|h_i,\vartheta_0,\,D_i\backslash\, y_{i,1:T}\right) f_0\left(h_i|c_{i0}\right) dh_i dy \\
&= \int \mathfrak{A}\left(h_i,\vartheta_0,D_i\right) f_0\left(h_i|c_{i0}\right) dh_i \\
&\leq M_{\mathfrak{A},i} \int f_0\left(h_i|c_{i0}\right) dh_i \\
&= M_{\mathfrak{A},i} \overset{\text{def}}{=} \bar{B}_{2,i}.
\end{aligned}$$

Since ϑ enjoys posterior consistency, as $N \to \infty$,

$$\mathbb{P}\left(\int B_{i0}\left(y;y^2\right) dy < \bar{B}_{2,i} \,\middle|\, D\right) \to 1,$$

[8]As $B_{i0}\left(y;y^2\right)$ is non-negative by definition, we get rid of $|\cdot|$.

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

C.5 Density Forecasts: (Correlated) Random Coefficients Model with Cross-sectional Homoskedasticity

Proof. **(Theorem 11)**

1. Condition 2 in Theorem 21. In the random coefficient case, conditions 2-a,b in Theorem 21 are given by Assumption 6(1-b,e). In the correlated random coefficient case, condition 2-a in Theorem 21 is given by Assumption 9(1-b); condition 2-b in Theorem 21 is satisfied because we consider individuals i with finite $\mathbb{E}_{f_0}\left[\left.\|\lambda\|_2^2\right| c_{i0}\right]$; for condition 2-c in Theorem 21, the continuity of $q_0\left(c_{i0}\right)$ is assumed in Theorem 11, and $q_0\left(c_{i0}\right) > 0$ follows Assumption 8.

2. Condition 3-a in Theorem 21.

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

where

$$\begin{aligned} m_i\left(\beta\right) &= \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1} \sum_t w_{i,t-1}\left(y_{it} - \beta' x_{i,t-1}\right), \qquad (60)\\ \Sigma_i\left(\sigma^2\right) &= \sigma^2\left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1}, \\ C_i\left(\beta, \sigma^2\right) &= \frac{1}{\sqrt{(2\pi)^{T-d_w}\left|\sum_t w_{i,t-1} w_{i,t-1}'\right|}}\left(\sigma^2\right)^{-\frac{T-d_w}{2}} \exp\left(-\frac{b_i\left(\beta\right)}{2\sigma^2}\right), \\ b_i\left(\beta\right) &= \sum_t\left(y_{it} - \beta' x_{i,t-1}\right)^2 \\ &\quad - \left(\sum_t w_{i,t-1}\left(y_{it} - \beta' x_{i,t-1}\right)\right)'\left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-1}\left(\sum_t w_{i,t-1}\left(y_{it} - \beta' x_{i,t-1}\right)\right) \\ &= \left(y_{i,1:T} - \beta' x_{i,0:T-1}\right) M_{w,i}\left(y_{i,1:T} - \beta' x_{i,0:T-1}\right)', \\ M_{w,i} &= I_{d_w} - w_{i,0:T-1}'\left(w_{i,0:T-1} w_{i,0:T-1}'\right)^{-1} w_{i,0:T-1}. \end{aligned}$$

$y_{i,1:T}$, $x_{i,0:T-1}$, 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 $C_i\left(\beta, \sigma^2\right)$ can be cancelled in the numerator and denominator of (58), we can replace $p\left(y_{i,1:T}\left|h_i, \vartheta, D_i\backslash y_{i,1:T}\right.\right)$ by

$$p\left(h_i\left|\vartheta, D_i\right.\right) = \phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right).$$

Given the boundedness and rank condition on $w_{i,t-1}$, the continuity part of Theorem 21(3-a) is satisfied. For the boundedness part of Theorem 21(3-a), as $\sigma^2 \in \left[\underline{\sigma}^2,\ \bar{\sigma}^2\right]$,

$$0 < \phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) \leq \sqrt{\left(2\pi\sigma^2\right)^{-d_w}\left|\sum_t w_{i,t-1}w_{i,t-1}'\right|} \tag{61}$$
$$\leq \sqrt{\left(2\pi\underline{\sigma}^2\right)^{-d_w}\left|\sum_t w_{i,t-1}w_{i,t-1}'\right|} \stackrel{\text{def}}{=} M_{l,i}.$$

3. Condition 3-b in Theorem 21. After cancelling $C_i\left(\beta,\sigma^2\right)$ in the numerator and denominator of (58), we can reduce $\mathfrak{A}\left(h_i,\vartheta,D_i\right)$ to

$$\begin{aligned}\tilde{\mathfrak{A}}\left(h_i,\vartheta,D_i\right) &= \mathbb{E}\left[y_{i,T+1}^2\left|h_i,\vartheta,D_i\right.\right]\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right)\\ &= \left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2\right)\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right).\end{aligned}$$

Then, $\tilde{\mathfrak{A}}\left(h_i,\vartheta,D_i\right)$ is continuous in h_i. For boundedness,

$$\begin{aligned}\tilde{\mathfrak{A}}\left(h_i,\vartheta,D_i\right) \leq& \left(2\left(\left\|\beta\right\|_2^2\left\|x_{iT}\right\|_2^2 + \left\|\lambda_i\right\|_2^2\left\|w_{iT}\right\|_2^2\right) + \sigma^2\right)\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right)\\ =& \left(2\left\|\beta\right\|_2^2\left\|x_{iT}\right\|_2^2 + \sigma^2\right)\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) + 2\left\|\lambda_i\right\|_2^2\left\|w_{iT}\right\|_2^2\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right).\end{aligned}$$

For $\left\|\vartheta-\vartheta_0\right\|_2 < \delta_\vartheta'$, the first term is bounded by $\left(2\left(\left\|\beta_0\right\|_2 + \delta_\vartheta'\right)^2\left\|x_{iT}\right\|_2^2 + \bar{\sigma}^2\right)M_{l,i}$. For the second term, note that $\max_x x^2\exp\left(-cx^2\right) = \frac{1}{ce}$, so we have

$$\begin{aligned}&\left\|\lambda_i\right\|_2^2\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right)\\ \leq& \frac{2}{e}\text{tr}\left(\Sigma_i\left(\sigma^2\right)\right)\Big/\sqrt{\left(2\pi\right)^{d_w}\left|\Sigma_i\left(\sigma^2\right)\right|}\\ \leq& \frac{2\text{tr}\left(\left(\sum_t w_{i,t-1}w_{i,t-1}'\right)^{-1}\right)\sqrt{\left|\sum_t w_{i,t-1}w_{i,t-1}'\right|}}{e\left(2\pi\right)^{d_w/2}}\sigma^{2(1-d_w/2)}\\ \leq& \frac{2d_w}{em_w}\left(\frac{\Lambda_{\max,ww,i}}{2\pi}\right)^{d_w/2}\max\left(\underline{\sigma}^{2(1-d_w/2)},\ \bar{\sigma}^{2(1-d_w/2)}\right).\end{aligned}$$

where $\Lambda_{\max,ww,i}$ is the largest eigenvalues of $\sum_t w_{i,t-1}w_{i,t-1}'$. Given Assumption 5(2), the eigenvalues of $\sum_t w_{i,t-1}w_{i,t-1}'$ are bounded below by $m_w > 0$, so the whole term is finite. Thus, there exists $\delta_\vartheta' > 0$ such that for all $\left\|\vartheta-\vartheta_0\right\|_2 < \delta_\vartheta'$, $\mathfrak{A}\left(h_i,\vartheta,D_i\right)$ is continuous in h_i and bounded by some $M_{\mathfrak{A},\mathrm{i}} > 0$.

4. Conditions 4-a,b,c in Theorem 21. These conditions are established via Lemma 31 on L_1-distance between normal distributions. Also, based on condition 3 in Theorem 21, we have $\sigma_0^2 \in \left(\underline{\sigma}^2,\ \bar{\sigma}^2\right)$

and $\sigma^2 \in [\underline{\sigma}^2,\ \bar{\sigma}^2]$. For **condition 4-a**, similar to the argument on page A-24,

$$\begin{aligned}
&\int \left|\phi\left(\lambda_i; m\left(\beta\right), \Sigma\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m\left(\beta_0\right), \Sigma\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &\sqrt{d_w\left(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln\frac{\sigma^2}{\sigma_0^2}\right) + \sigma_0^{-2}\left(\beta - \beta_0\right)' M_{xw,i}\left(\beta - \beta_0\right)} \\
\leq &\sqrt{d_w}\frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\sigma_0^2\underline{\sigma}^2}\left|\sigma^2 - \sigma_0^2\right| + \sqrt{\frac{\Lambda_{\max,xw,i}}{\sigma_0^2}}\left\|\beta - \beta_0\right\|_2 \\
\leq &\max\left(\sqrt{d_w}\frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\sigma_0^2\underline{\sigma}^2}, \sqrt{\frac{\Lambda_{\max,xw,i}}{\sigma_0^2}}\right)\cdot\left\|\vartheta - \vartheta_0\right\|_2 \stackrel{\text{def}}{=} C_{l,i}\left(\left\|\vartheta - \vartheta_0\right\|_2\right),
\end{aligned} \tag{62}$$

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 $\Lambda_{\max,xw,i}$ is the largest eigenvalue of $M_{xw,i}$. Given Assumption 5(2), $\sum_t w_{i,t-1}w'_{i,t-1}$ is non-degenerate, so we have $\Lambda_{\max,xw,i} < \infty$.

The term in **conditions 4-b,c** is

$$p\left(y|\, h_i, \vartheta, D_i\right) = \phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right).$$

Similarly, using Lemma 31 to bound the L_1-distance between normal distributions, condition 4-b is given by

$$\begin{aligned}
&\int \left|\phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) - \phi\left(y; \beta_0' x_{iT} + \lambda_i' w_{iT}, \sigma_0^2\right)\right| dy \\
\leq &\sqrt{\frac{\sigma^2}{\sigma_0^2} - 1 - \ln\frac{\sigma^2}{\sigma_0^2} + \sigma_0^{-2}\left(\beta - \beta_0\right)' x_{iT}x_{iT}'\left(\beta - \beta_0\right)} \\
\leq &\frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\sigma_0^2\underline{\sigma}^2}\left|\sigma^2 - \sigma_0^2\right| + \frac{1}{\sqrt{\sigma_0^2}}\left\|\beta - \beta_0\right\|_2\left\|x_{iT}\right\|_2 \\
\leq &\max\left(\frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\sigma_0^2\underline{\sigma}^2}, \frac{\left\|x_{iT}\right\|_2}{\sqrt{\sigma_0^2}}\right)\cdot\left\|\vartheta - \vartheta_0\right\|_2 \stackrel{\text{def}}{=} C_{p,i}\left(\left\|\vartheta - \vartheta_0\right\|_2\right)
\end{aligned}$$

and so is condition 4-c

$$\begin{aligned}
&\int \left|\phi\left(y; \beta' x_{iT} + \lambda_i' w_{iT}, \sigma^2\right) - \phi\left(y; \beta' x_{iT} + \tilde{\lambda}_i' w_{iT}, \sigma^2\right)\right| dy \\
\leq &\sqrt{\sigma^{-2}\left(\lambda_i - \tilde{\lambda}_i\right)' w_{iT}w_{iT}'\left(\lambda_i - \tilde{\lambda}_i\right)} \\
\leq &\frac{\left\|w_{iT}\right\|_2}{\sqrt{\underline{\sigma}^2}}\cdot\left\|\lambda_i - \tilde{\lambda}_i\right\|_2 \stackrel{\text{def}}{=} C_{h,i}\left(\left\|\lambda_i - \tilde{\lambda}_i\right\|_2\right).
\end{aligned}$$

5. Condition 4-d in Theorem 21. Again, these conditions are established via Lemma 31 on L_1-

distance between normal distributions, as well as that $\sigma_0^2 \in \left(\underline{\sigma}^2,\ \bar{\sigma}^2\right)$ and $\sigma^2 \in \left[\underline{\sigma}^2,\ \bar{\sigma}^2\right]$.

$$
\begin{aligned}
&\int \left|\tilde{\mathfrak{A}}\left(h_i, \vartheta, D_i\right) - \tilde{\mathfrak{A}}\left(h_i, \vartheta_0, D_i\right)\right| dh_i \\
=&\int \left|\left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2\right) \phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \left(\left(\beta_0' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma_0^2\right) \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &\int \left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2\right) \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
&+ \int \left|\left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2\right) - \left(\left(\beta_0' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma_0^2\right)\right| \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right) d\lambda_i.
\end{aligned}
$$

First term:

$$
\begin{aligned}
&\int \left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2\right) \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &\left(2\left\|\beta\right\|_2^2 \left\|x_{iT}\right\|_2^2 + \sigma^2\right) \int \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
&+ 2\left\|w_{iT}\right\|_2^2 \int \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &\left(2\left(\left\|\beta_0\right\|_2 + \left\|\beta - \beta_0\right\|_2\right)^2 \left\|x_{iT}\right\|_2^2 + \bar{\sigma}^2\right) \cdot C_{l,i}\left(\left\|\vartheta - \vartheta_0\right\|_2\right) \\
&+ 2\left\|w_{iT}\right\|_2^2 \int \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i.
\end{aligned}
$$

The second inequality follows (62), which builds on Lemma 31 on L_1-distance between normal distributions. Moreover, let $M_h = \left\|\vartheta - \vartheta_0\right\|_2^{-1/4}$,

$$
\begin{aligned}
&\int \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &\int_{\left\|\lambda_i\right\|_2 \leq M_h} \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
&+ \int_{\left\|\lambda_i\right\|_2 > M_h} \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i.
\end{aligned}
$$

For the first part,

$$
\begin{aligned}
&\int_{\left\|\lambda_i\right\|_2 \leq M_h} \left\|\lambda_i\right\|_2^2 \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &M_h^2 \int \left|\phi\left(\lambda_i; m_i\left(\beta\right), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i\left(\beta_0\right), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq &M_h^2 C_{l,i}\left(\left\|\vartheta - \vartheta_0\right\|_2\right) \\
= &\max\left(\sqrt{d_w}\frac{\bar{\sigma}^2 - \underline{\sigma}^2}{\sigma_0^2 \underline{\sigma}^2}, \sqrt{\frac{\Lambda_{\max,xw,i}}{\sigma_0^2}}\right) \cdot \left\|\vartheta - \vartheta_0\right\|_2^{1/2}.
\end{aligned}
$$

The second inequality follows (62). For the second part,

$$
\begin{aligned}
&\int_{\|\lambda_i\|_2 > M_h} \|\lambda_i\|_2^2 \left|\phi\left(\lambda_i; m_i(\beta), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq & M_h^{-2} \int \|\lambda_i\|_2^4 \left|\phi\left(\lambda_i; m_i(\beta), \Sigma_i\left(\sigma^2\right)\right) - \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right)\right| d\lambda_i \\
\leq & M_h^{-2} \left(\int \|\lambda_i\|_2^4 \phi\left(\lambda_i; m_i(\beta), \Sigma_i\left(\sigma^2\right)\right) d\lambda_i + \int \|\lambda_i\|_2^4 \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) d\lambda_i \right) \\
\leq & M_h^{-2} \left(\|m_i(\beta)\|_2^4 + 6\|m_i(\beta)\|_2^2 \operatorname{tr}\left(\Sigma_i\left(\sigma^2\right)\right) + \operatorname{tr}\left(\Sigma_i\left(\sigma^2\right)\right)^2 + 2\operatorname{tr}\left(\Sigma_i\left(\sigma^2\right)^2\right) \right. \\
& \left. + \|m_i(\beta_0)\|_2^4 + 6\|m_i(\beta_0)\|_2^2 \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right) + \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right)^2 + 2\operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)^2\right) \right) \\
\leq & M_h^{-2} \left(\|m_i(\beta_0)\|_2^4 + 6\|m_i(\beta_0)\|_2^2 \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right) + \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right)^2 + 2\operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)^2\right) \right. \\
& + 8\|m_i(\beta_0)\|_2^4 + 8\left(\Lambda_{\max,xw2,i}\right)^2 \|\beta - \beta_0\|_2^4 + 6\operatorname{tr}\left(\Sigma_i\left(\bar{\sigma}^2\right)\right)\left(2\|m_i(\beta_0)\|_2^2 + 2\Lambda_{\max,xw2,i}\|\beta - \beta_0\|_2^2\right) \\
& \left. + \operatorname{tr}\left(\Sigma_i\left(\bar{\sigma}^2\right)\right)^2 + 2\operatorname{tr}\left(\Sigma_i\left(\bar{\sigma}^2\right)^2\right) \right) \\
= & \|\vartheta - \vartheta_0\|_2^{1/2} \left(C_{\mathfrak{A}1,i}^{(0)} + C_{\mathfrak{A}1,i}^{(2)} \|\beta - \beta_0\|_2^2 + C_{\mathfrak{A}1,i}^{(4)} \|\beta - \beta_0\|_2^4 \right),
\end{aligned}
$$

where $\Lambda_{\max,xw2,i}$ is the largest eigenvalue of $\sum_t x_{i,t-1} w_{i,t-1}' \left(\sum_t w_{i,t-1} w_{i,t-1}'\right)^{-2} \sum_t w_{i,t-1} x_{i,t-1}'$. Given Assumption 5(2), $\sum_t w_{i,t-1} w_{i,t-1}'$ is non-degenerate, so we have $\Lambda_{\max,xw2,i} < \infty$.

Second term:

$$
\begin{aligned}
& \left| \left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2 \right) - \left(\left(\beta_0' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma_0^2 \right) \right| \\
\leq & \|\beta - \beta_0\|_2 \|x_{iT}\|_2 \left(\|\beta - \beta_0\|_2 \|x_{iT}\|_2 + 2\|\beta_0\|_2 + 2\|\lambda_i\|_2 \|w_{iT}\|_2 \right) + \left|\sigma^2 - \sigma_0^2\right| \\
\leq & \|\beta - \beta_0\|_2 \|x_{iT}\|_2 \left(\|\beta - \beta_0\|_2 \|x_{iT}\|_2 + 2\|\beta_0\|_2 + \left(\|\lambda_i\|_2^2 + 1\right) \|w_{iT}\|_2 \right) + \left|\sigma^2 - \sigma_0^2\right| \\
\stackrel{\text{def}}{=} & C_{\mathfrak{A}2,i}\left(\|\vartheta - \vartheta_0\|_2\right) \left(\|\lambda_i\|_2^2 + 1\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
& \int \left| \left(\left(\beta' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma^2 \right) - \left(\left(\beta_0' x_{iT} + \lambda_i' w_{iT}\right)^2 + \sigma_0^2 \right) \right| \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) d\lambda_i \\
\leq & C_{\mathfrak{A}2,i}\left(\|\vartheta - \vartheta_0\|_2\right) \left(\int \|\lambda_i\|_2^2 \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) d\lambda_i + \int \phi\left(\lambda_i; m_i(\beta_0), \Sigma_i\left(\sigma_0^2\right)\right) d\lambda_i \right) \\
= & C_{\mathfrak{A}2,i}\left(\|\vartheta - \vartheta_0\|_2\right) \cdot \left(\|m_i(\beta_0)\|_2^2 + \operatorname{tr}\left(\Sigma_i\left(\sigma_0^2\right)\right) + 1 \right).
\end{aligned}
$$

Therefore, there exists an increasing function $C_{\mathfrak{A},\mathrm{i}}\left(\|\vartheta - \vartheta_0\|_2\right) \geq 0$ with $\lim_{x \to 0} C_{\mathfrak{A},\mathrm{i}}(x) = 0$ satisfying condition 4-d in Theorem 21. $\square$

C.6 Useful Lemmas

Lemma 29. (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,*

$$D_{KL}(g_0 \parallel g) \leq D_{KL}(f_0 \parallel f).$$

2. *(Independence) The KL divergence is addictive for independent distributions. If $f_{x,0}$ and $f_{y,0}$ are independent distributions with joint distribution $f_0(x,y) = f_{x,0}(x) f_{y,0}(y)$, and similarly $f(x,y) = f_x(x) f_y(y)$, then,*

$$D_{KL}(f_0 \parallel f) = D_{KL}(f_{x,0} f_{y,0} \parallel f_x f_y) = D_{KL}(f_{x,0} \parallel f_x) + D_{KL}(f_{y,0} \parallel f_y).$$

3. *(Invertible Transformation) The KL divergence is invariant under invertible transformation. If $y = h(x)$, where h is a invertible function,*

$$D_{KL}(f_{y,0} \parallel f_y) = D_{KL}(f_{x,0} \parallel f_x).$$

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

$$\begin{aligned} g_0(y) \log \frac{g_0(y)}{g(y)} &= g(y)\, \ell\left(\frac{g_0(y)}{g(y)}\right) \\ &= g(y)\, \ell\left(\int \frac{p(y-x) f(x)}{\int p(y-x) f(x)\, dx} \cdot \frac{f_0(x)}{f(x)} dx\right) \\ &\leq g(y) \int \frac{p(y-x) f(x)}{\int p(y-x) f(x)\, dx} \cdot \ell\left(\frac{f_0(x)}{f(x)}\right) dx \\ &= \int p(y-x) f(x)\, \ell\left(\frac{f_0(x)}{f(x)}\right) dx \\ &= \int p(y-x) f_0(x) \log \frac{f_0(x)}{f(x)} dx, \end{aligned} \tag{63}$$

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

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

where the inequality follow the above derivation (63).

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

Remark 30. We can extend Property 1 to a more general "convolution" form. Let $u = h\left(x,y\right)$, where $h\left(x,y\right)$ is invertible in y for all x, then $\left|\frac{\partial h(x,y)}{\partial y}\right| > 0$ for all $\left(x,y\right)$.[9] Given

$$\begin{aligned} g_0\left(y\right) &= \int \left|\frac{\partial h\left(x,y\right)}{\partial y}\right| p\left(h\left(x,y\right)\right)f_0\left(x\right)dx, \\ g\left(y\right) &= \int \left|\frac{\partial h\left(x,y\right)}{\partial y}\right| p\left(h\left(x,y\right)\right)f\left(x\right)dx, \end{aligned}$$

we can obtain $D_{KL}\left(g_0 \parallel g\right) \leq D_{KL}\left(f_0 \parallel f\right)$ in a similar manner.

Lemma 31. (L_1-Distance between Normal Distributions) *Suppose we have two multivariate normal distributions $\phi\left(x;\mu_1,\Sigma_1\right)$ and $\phi\left(x;\mu_2,\Sigma_2\right)$, where x is a $d_x \times 1$ vector, then*

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

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

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

and then plug in the formula of the KL divergence between multivariate normals. □

Lemma 32. (Tail of Normal Distribution) *If x follows a standard normal distribution, $x \sim N\left(0,1\right)$, then for $x^* > 0$,*

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

Proof. See Feller (1968).

[9] In Property 1 above, $h\left(x,y\right) = y - x$.

□

D Algorithms

D.1 Random Coefficients Model

For the random coefficients model, I adopt 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, and the data augmentation improves numerical convergence and leads to faster code.

To check the robustness regarding the truncation, I also implement the more sophisticated yet complicated slice-retrospective sampler (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015), which does not truncate the number of components at a predetermined K. The estimates and forecasts of the two samplers are almost indistinguishable, so I will only show the results generated from the simpler truncation sampler.

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 (5). 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 34(2). For individual heterogeneity $z=\lambda,l$, let $\gamma_{z,i}$ be individual i's component affiliation, which can take values $\{1,\cdots,K_z\}$, $J_{z,k}$ be the set of individuals in component k, i.e. $J_{z,k}=\{i: \gamma_{z,i}=k\}$, and $n_{z,k}$ be the number of individuals in component k, i.e. $n_{z,k}=\#J_{z,k}$. Then, the (data-augmented) joint posterior for the model parameters is given by

$$
\begin{aligned}
& p\left(\left\{\alpha_z,\left\{p_{z,k},\mu_{z,k},\Omega_{z,k}\right\},\left\{\gamma_{z,i},z_i\right\}\right\},\beta\middle|\, D\right) \\
& =\prod_{i,t} p\left(y_{it}\left|\lambda_i,l_i,\beta,w_{i,t-1},x_{i,t-1}\right.\right)\cdot\prod_{z,i} p\left(z_i\left|\mu_{z,\gamma_{z,i}},\Omega_{z,\gamma_{z,i}}\right.\right)p\left(\gamma_{z,i}\left|\{p_{z,k}\}\right.\right) \\
& \quad\cdot\prod_{z,k} p\left(\mu_{z,k},\Omega_{z,k}\right)p\left(p_{z,k}|\alpha_z\right)\cdot p\left(\alpha_z\right)\cdot p\left(\beta\right),
\end{aligned}
$$

where $z=\lambda,l$, $k=1,\cdots,K_z$, $i=1,\cdots N$, and $t=1,\cdots,T$. The first block links observations to model parameters $\{\lambda_i,l_i\}$ and β. The second block links the individual heterogeneity z_i to the underlying distribution f_z. The last block formulates the prior on (β,f).[10]

The proposed Gibbs sampler cycles over the following blocks of parameters (in order): (1) component probabilities, $\alpha_z,\{p_{z,k}\}$; (2) component parameters, $\{\mu_{z,k},\Omega_{z,k}\}$; (3) component memberships, $\{\gamma_{z,i}\}$; (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, step 5 alone would be sufficient to recover the common parameters. If the mixture structure of f_z were known (i.e. if $(p_{z,k},\mu_{z,k},\Omega_{z,k})$ 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 33. (Random Coefficients with Cross-sectional Heteroskedasticity)[11] *For each iteration* $s=1,\cdots,n_{sim}$,

1. *Component probabilities: For* $z=\lambda,l$,
   (a) *Draw* $\alpha_z^{(s)}$ *from a gamma distribution* $p\left(\alpha_z^{(s)}\middle|\, p_{z,K_z}^{(s-1)}\right)$:

$$
\alpha_z^{(s)}\sim Ga\left(a_{\alpha_z,0}+K_z-1,b_{\alpha_z,0}-\log p_{z,K_z}^{(s-1)}\right).
$$

   (b) *For* $k=1,\cdots,K_z$, *draw* $p_{z,k}^{(s)}$ *from the truncated stick-breaking process* $p\left(\left\{p_{z,k}^{(s)}\right\}\middle|\,\alpha_z^{(s)},\left\{n_{z,k}^{(s-1)}\right\}\right)$:

$$
p_{z,k}^{(s)}\sim TSB\left(1+n_{z,k}^{(s-1)},\alpha_z^{(s)}+\sum_{j=k+1}^{K_z} n_{z,j}^{(s-1)},K_z\right).
$$

[10] The hyperparameters are chosen in a relatively ignorant sense without inferring much from the data except aligning the scale with the variance of the data. See Appendix D.3 for the details of the baseline model with random effects.

[11] Below, I present the formulas for the key nonparametric Bayesian steps, and leave the details of standard posterior sampling procedures, such as drawing from a normal-inverse-gamma distribution or a linear regression, to Appendix D.5.

2. *Component parameters: For $z=\lambda, l$, and $k=1,\cdots,K_z$, draw $\left(\mu_{z,k}^{(s)},\Omega_{z,k}^{(s)}\right)$ from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a scalar) $p\left(\mu_{z,k}^{(s)},\Omega_{z,k}^{(s)}\left|\left\{z_i^{(s-1)}\right\}_{i\in J_{z,k}^{(s-1)}}\right.\right)$.*
3. *Component memberships: For $z=\lambda,l$, amd $i=1,\cdots N$, draw $\gamma_{z,i}^{(s)}$ from a multinomial distribution $p\left(\left\{\gamma_{z,i}^{(s)}\right\}\left|\left\{p_{z,k}^{(s)},\mu_{z,k}^{(s)},\Omega_{z,k}^{(s)}\right\},z_i^{(s-1)}\right.\right)$:*

$$\gamma_{z,i}^{(s)}=k,\ \text{with probability}\, p_{ik}\propto p_{z,k}^{(s)}\phi\left(z_i^{(s-1)};\ \mu_{z,k}^{(s)},\Omega_{z,k}^{(s)}\right),\quad \sum_{k=1}^{K_z}p_{ik}=1.$$

4. *Individual-specific parameters:*
   (a) *For $i=1,\cdots,N$, draw $\lambda_i^{(s)}$ from a multivariate normal distribution (or a normal distribution if λ is a scalar) $p\left(\lambda_i^{(s)}\left|\mu_{\lambda,\gamma_{\lambda,i}}^{(s)},\Omega_{\lambda,\gamma_{\lambda,i}}^{(s)},\left(\sigma_i^2\right)^{(s-1)},\beta^{(s-1)},D_i\right.\right)$.*
   (b) *For $i=1,\cdots,N$, draw $l_i^{(s)}$ via the random-walk Metropolis-Hastings approach,*

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

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

Remark 34. (1) With the above prior specification, all steps enjoy closed-form conditional posterior distributions except step 4-b for σ_i^2. 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), 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 36 for details.
(2) In cross-sectional homoskedastic cases, the algorithm would need the following changes: (a) in steps 1 to 4, only λ_i is considered, and (b) in step 5, $\left(\beta^{(s)},\left(\sigma^2\right)^{(s)}\right)$ are drawn from a linear regression model with an "unknown" variance, $p\left(\beta^{(s)},\left(\sigma^2\right)^{(s)}\left|\left\{\lambda_i^{(s)}\right\},D\right.\right)$.

D.2 Correlated Random Coefficients Model

To account for the conditional structure in the correlated random coefficients model, I implement a multivariate MGLR$_\mathrm{x}$ 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 in Section 2.1 for balanced panels or 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 retain 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 $A_k = C_k B_k$. Define $\eta = \beta/3$. According to the expressions in Assumption 9(3), we can let B_k^η follow the standard exponential distribution, i.e. $p\left(B_k^\eta\right) = \exp\left(-B_k^\eta\right)$, and 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 the sieve property in Theorem 4(3).[12] Furthermore, it is helpful to introduce a set of auxiliary stochastic functions $\xi_k(c_{i0})$, $k = 1, 2, \cdots$, such that

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

Note that the probit stick-breaking process defined in (6) can be recovered by marginalizing over $\{\xi_k(c_{i0})\}$. Finally, I combine the MGLR$_\mathrm{x}$ prior with Ishwaran and James (2001, 2002) truncation approximation to simplify the numerical procedure while still retaining reliable results.

Let $N\times 1$ vectors $\boldsymbol{\zeta}_k = \left[\zeta_k(c_{10}), \zeta_k(c_{20}), \cdots, \zeta_k(c_{N0})\right]'$ and $\boldsymbol{\xi}_k = \left[\xi_k(c_{10}), \xi_k(c_{20}), \cdots, \xi_k(c_{N0})\right]'$, as well as an $N\times N$ matrix $\boldsymbol{V}_k = \tilde{V}(A_k)$ with the i,j-th element being $(\boldsymbol{V}_k)_{ij} = \exp\left(-A_k \|c_{i0} - c_{j0}\|_2^2\right)$. The next algorithm extends Algorithm 33 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 33.

Algorithm 35. (Correlated Random Coefficients with Cross-sectional Heteroskedasticity)[13] *For each iteration $s = 1, \cdots, n_{sim}$,*

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

---

[12] In 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 the Monte Carlo simulations and the empirical application, and the results are robust across a range of C_*, e.g. from 1 to 10.

[13] See Remark 34(2) for the adaption to cross-sectional homoskedastic models.

(a) *For $k = 1, \cdots, K_z - 1$, draw $A_{z,k}^{(s)}$ via the random-walk Metropolis-Hastings approach,*[14]

$$p\left(A_{z,k}^{(s)} \middle| \zeta_{z,k}^{(s-1)}, \{c_{i0}\} \right) \propto \left(A_{z,k}^{(s)} \right)^{\eta-1} \exp\left(-\left(\frac{A_{z,k}^{(s)}}{C_k} \right)^{\eta} \right) \cdot \phi\left(\zeta_{z,k}^{(s-1)};\ 0, \tilde{V}\left(A_{z,k}^{(s)} \right) \right).$$

Then, calculate $\boldsymbol{V}_{z,k}^{(s)} = \tilde{V}\left(A_{z,k}^{(s)} \right)$.

(b) *For $k = 1, \cdots, K_z-1$, and $i = 1, \cdots, N$, draw $\xi_{z,k}^{(s)}(c_{i0})$ from a truncated normal distribution $p\left(\xi_{z,k}^{(s)}(c_{i0}) \middle| \zeta_{z,k}^{(s-1)}(c_{i0}), \gamma_{z,i}^{(s-1)} \right)$:*

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

(c) *For $k = 1, \cdots, K_z-1$, $\zeta_{z,k}^{(s)}$, draw from a multivariate normal distribution $p\left(\zeta_{z,k}^{(s)} \middle| \boldsymbol{V}_{z,k}^{(s)}, \xi_{z,k}^{(s)} \right)$:*

$$\zeta_{z,k}^{(s)} \sim N\left(m_{\zeta,k}, \Sigma_{\zeta,k} \right), \text{ where } \Sigma_{\zeta,k} = \left[\left(\boldsymbol{V}_{z,k}^{(s)} \right)^{-1} + I_N \right]^{-1} \text{ and } m_{\zeta,k} = \Sigma_{\zeta,k} \xi_{z,k}^{(s)}.$$

(d) *For $k = 1, \cdots, K_z$, and $i = 1, \cdots, N$, the component probabilities $p_{z,k}^{(s)}(c_{i0})$ are fully determined by $\zeta_{z,k}^{(s)}$:*

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

2. *Component parameters: For $z = \lambda, l$, and $k = 1, \cdots, K_z$,*
   (a) *Draw $\mathrm{vec}\left(\mu_{z,k}^{(s)} \right)$ from a multivariate normal distribution $p\left(\mu_{z,k}^{(s)} \middle| \Omega_{z,k}^{(s-1)}, \left\{ z_i^{(s-1)}, c_{i0} \right\}_{i \in J_{z,k}^{(s-1)}} \right)$.*
   (b) *Draw $\Omega_{z,k}^{(s)}$ from an inverse Wishart distribution (or an inverse gamma distribution if z is a scalar) $p\left(\Omega_{z,k}^{(s)} \middle| \mu_{z,k}^{(s)}, \left\{ z_i^{(s-1)}, c_{i0} \right\}_{i \in J_{z,k}^{(s-1)}} \right)$.*
3. *Component memberships: For $z = \lambda, l$, and $i = 1, \cdots N$, draw $\gamma_{z,i}^{(s)}$ from a multinomial distribution $p\left(\left\{ \gamma_{z,i}^{(s)} \right\} \middle| \left\{ p_{z,k}^{(s)}, \mu_{z,k}^{(s)}, \Omega_{z,k}^{(s)} \right\}, z_i^{(s-1)}, c_{i0} \right)$:*

$$\gamma_{z,i}^{(s)} = k, \text{ with probability } p_{ik} \propto p_{z,k}^{(s)}(c_{i0}) \, \phi\left(z_i^{(s-1)};\ \mu_{z,k}^{(s)} \left[1, c_{i0}' \right]', \Omega_{z,k}^{(s)} \right), \quad \sum_{k=1}^{K_z} p_{ik} = 1.$$

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

4. *Individual-specific parameters:*
   (a) *For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a multivariate normal distribution (or a normal distribution if* λ *is a scalar)* $p\left(\lambda_i^{(s)} \middle| \mu_{\lambda,\gamma_{\lambda,i}}^{(s)}, \Omega_{\lambda,\gamma_{\lambda,i}}^{(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i\right)$.
   (b) *For* $i = 1, \cdots, N$, *draw* $l_i^{(s)}$ *via the random-walk Metropolis-Hastings approach* $p\left(l_i^{(s)} \middle| \mu_{l,\gamma_{l,i}}^{(s)}, \Omega_{l,\gamma_{l,i}}^{(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i\right)$, *then calculate* $\left(\sigma_i^2\right)^{(s)}$ *based on* $\sigma^2(l)$.
5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model with a "known" variance,* $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right)$.

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 norma-inverse-gamma form,

$$\left(\beta, \sigma^2\right) \sim N\left(m_{\beta,0}, \psi_{\beta,0}\sigma^2\right) \mathrm{IG}\left(a_{\sigma^2,0}, b_{\sigma^2,0}\right).$$

The hyperparameters are chosen in a relatively ignorant sense without inferring much from the data except aligning the scale with the variance of the data.

$$a_{\sigma^2,0} = 2, \tag{64}$$

$$b_{\sigma^2,0} = \hat{\mathbb{E}}\left(\hat{\mathbb{V}}_i\left(y_{it}\right)\right) \cdot \left(a_{\sigma^2,0} - 1\right) = \hat{\mathbb{E}}\left(\hat{\mathbb{V}}_i\left(y_{it}\right)\right), \tag{65}$$

$$m_{\beta,0} = 0.5, \tag{66}$$

$$\psi_{\beta,0} = \frac{1}{b_{\sigma^2,0} / \left(a_{\sigma^2,0} - 1\right)} = \frac{1}{\hat{\mathbb{E}}\left(\hat{\mathbb{V}}_i\left(y_{it}\right)\right)}. \tag{67}$$

In (65) here and (68) below, $\hat{\mathbb{E}}_i$ and $\hat{\mathbb{V}}_i$ stand for the sample mean and variance for firm i over $t = 1, \cdots, T$, and $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ further take the sample mean and variance over the cross-section $i = 1, \cdots, N$. Equation (65) ensures that on average the prior and the data have a similar scale. Equation (66) conjectures that the young firm dynamics are likely to be 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 (64) characterizes a rather fat-tailed prior on σ^2 with infinite variance, and (67) assumes that the prior variance of β is equal to 1 on average.

The hyperpriors for the DPM prior are specified as:

$$G_0\left(\mu_k, \omega_k^2\right) = N\left(m_{\lambda,0}, \psi_{\lambda,0}\omega_k^2\right) \mathrm{IG}\left(a_{\lambda,0}, b_{\lambda,0}\right),$$
$$\alpha \sim \mathrm{Ga}\left(a_{\alpha,0}, b_{\alpha,0}\right).$$

Similarly, the hyperparameters are chosen to be:

$$
\begin{aligned}
a_{\lambda,0} &= 2,\ b_{\lambda,0} = \hat{\mathbb{V}}\left(\hat{\mathbb{E}}_i\left(y_{it}\right)\right)\cdot\left(a_{\lambda,0}-1\right) = \hat{\mathbb{V}}\left(\hat{\mathbb{E}}_i\left(y_{it}\right)\right), && (68)\\
m_{\lambda,0} &= 0,\ \psi_{\lambda,0} = 1,\\
a_{\alpha,0} &= 2,\ b_{\alpha,0} = 2. && (69)
\end{aligned}
$$

where $b_{\lambda,0}$ is selected to match the scale, while $a_{\lambda,0}$, $m_{\lambda,0}$, and $\psi_{\lambda,0}$ yields a relatively ignorant and diffuse prior. Following Ishwaran and James (2001, 2002), the hyperparameters for the DP scale parameter α in (69) allow for a flexible component structure with a wide range of component numbers. The truncated number of components is set to be $K = 50$, so that the approximation error is uniformly bounded by Ishwaran and James (2001) Theorem 2:

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

at the prior mean of α ($a_{\alpha,0}/b_{\alpha,0} = 1$) 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), who adaptively adjust the random walk step size in order to keep acceptance rates around a certain desirable percentage.

Algorithm 36. (Adaptive RWMH) *Let us consider a generic variable θ. For each iteration $s = 1, \cdots, n_{sim}$,*

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

$$
a.r.(\tilde{\theta}|\theta^{(s-1)}) = \min\left(1, \frac{p(\tilde{\theta}|\cdot)}{p(\theta^{(s-1)}|\cdot)}\right),
$$

   where $p(\theta|\cdot)$ is the conditional posterior distribution of interest.
3. *Accept the proposal and set $\theta^{(s)} = \tilde{\theta}$ with probability $a.r.(\tilde{\theta}|\theta^{(s-1)})$. Otherwise, reject the proposal and set $\theta^{(s)} = \theta^{(s-1)}$.*
4. *Update the random-walk step size for the next iteration,*

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

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

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

with $\bar{x} > 0$ *being a very large number.*

Remark 37. (1) In step 1, since the algorithms in this paper only consider the RWMH on conditionally independent scalar variables, Σ is simply taken to be 1.
(2) In step 4, I choose $c = 0.55$, a.r.$^{\star} = 30\%$ in the numerical exercises.

D.5 Details on Posterior Samplers

D.5.1 Step 2: Component Parameters

Random Coefficients Model For $z = \lambda, l$, and $k = 1, \cdots, K_z$, draw $\left(\mu_{z,k}^{(s)}, \Omega_{z,k}^{(s)}\right)$ from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a scalar) $p\left(\mu_{z,k}^{(s)}, \Omega_{z,k}^{(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{z,k}^{(s-1)}}\right)$:

$$\begin{aligned}
\left(\mu_{z,k}^{(s)}, \Omega_{z,k}^{(s)}\right) &\sim N\left(m_{z,k}, \psi_{z,k}\Omega_{z,k}^{(s)}\right) \mathrm{IW}\left(\Psi_{z,k}, \nu_{z,k}\right), \\
\psi_{z,k} &= \left((\psi_{z,0})^{-1} + n_{z,k}^{(s-1)}\right)^{-1}, \\
m_{z,k} &= \psi_{z,k}\left((\psi_{z,0})^{-1} m_{z,0} + \sum_{i \in J_{z,k}^{(s-1)}} z_i^{(s-1)}\right), \\
\nu_{z,k} &= \nu_{z,0} + n_{z,k}^{(s-1)}, \\
\Psi_{z,k} &= \Psi_{z,0} + \sum_{i \in J_{z,k}^{(s-1)}} \left(z_i^{(s-1)}\right)^2 + m_{z,0}'(\psi_{z,0})^{-1} m_{z,0} - m_{z,k}'(\psi_{z,k})^{-1} m_{z,k}.
\end{aligned}$$

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

(a) Draw $\text{vec}\left(\mu_{z,k}^{(s)}\right)$ from a multivariate normal distribution $p\left(\mu_{z,k}^{(s)} \middle| \Omega_{z,k}^{(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i\in J_{z,k}^{(s-1)}}\right)$:

$$\begin{aligned}
\text{vec}\left(\mu_{z,k}^{(s)}\right) &\sim N\left(\text{vec}\left(m_{z,k}\right), \psi_{z,k}\right), \\
\hat{m}_{z,k}^{zc} &= \sum_{i\in J_{z,k}^{(s-1)}} z_i^{(s-1)}\left[1, c_{i0}'\right], \\
\hat{m}_{z,k}^{cc} &= \sum_{i\in J_{z,k}^{(s-1)}} \left[1, c_{i0}'\right]'\left[1, c_{i0}'\right], \\
\hat{m}_{z,k} &= \hat{m}_{z,k}^{zc}\left(\hat{m}_{z,k}^{cc}\right)^{-1}, \\
\psi_{z,k} &= \left[\left(\psi_{z,0}\right)^{-1} + \hat{m}_{z,k}^{cc}\otimes\left(\Omega_{z,k}^{(s-1)}\right)^{-1}\right]^{-1}, \\
\text{vec}\left(m_{z,k}\right) &= \psi_{z,k}\left[\left(\psi_{z,0}\right)^{-1}\text{vec}\left(m_{z,0}\right) + \left(\hat{m}_{z,k}^{cc}\otimes\left(\Omega_{z,k}^{(s-1)}\right)^{-1}\right)\text{vec}\left(\hat{m}_{z,k}\right)\right].
\end{aligned}$$

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

$$\begin{aligned}
\Omega_{z,k}^{(s)} &\sim \text{IW}\left(\Psi_{z,k}, \nu_{z,k}\right), \\
\nu_{z,k} &= \nu_{z,0} + n_{z,k}^{(s-1)}, \\
\Psi_{z,k} &= \Psi_{z,0} + \sum_{i\in J_{z,k}^{(s-1)}}\left(z_i^{(s-1)} - \mu_{z,k}^{(s)}\left[1, c_{i0}'\right]'\right)\left(z_i^{(s-1)} - \mu_{z,k}^{(s)}\left[1, c_{i0}'\right]'\right)'.
\end{aligned}$$

D.5.2 Step 4: Individual-specific Parameters

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

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

where the conditional "prior" mean is characterized by

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

D.5.3 Step 5: Common parameters

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

$$
\begin{aligned}
\left(\beta^{(s)}, \left(\sigma^2\right)^{(s)}\right) \sim N\left(m_\beta, \psi_\beta \left(\sigma^2\right)^{(s)}\right) \mathrm{IG}\left(a_{\sigma^2}, b_{\sigma^2}\right), \\
\psi_\beta &= \left(\left(\psi_{\beta,0}\right)^{-1} + \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t-1} x_{i,t-1}'\right)^{-1}, \\
m_\beta &= \psi_\beta \left(\left(\psi_{\beta,0}\right)^{-1} m_{\beta,0} + \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right), \\
a_{\sigma^2} &= a_{\sigma^2,0} + \frac{NT}{2}, \\
b_{\sigma^2} &= b_{\sigma^2,0} + \frac{1}{2}\left(\sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)^2 + m_{\beta,0}' \left(\psi_{\beta,0}\right)^{-1} m_{\beta,0} - m_\beta' \left(\psi_\beta\right)^{-1} m_\beta\right).
\end{aligned}
$$

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

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

Remark 38. For unbalanced panels, the summations and products in steps 4 and 5 (Subsections D.5.2 and D.5.3) are instead over $t \in s_{i,1:T_i-1}$, where $s_{i,1:T_i-1}$ is the observed periods of individual i used for estimation.

D.6 Parametric Specification of Heteroskedasticity

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

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

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

$$\begin{aligned}
& b \sim \mathrm{Ga}\left(a_{b,0}, b_{b,0}\right), \\
& p\left(a \mid b, a_{a,0}, b_{a,0}, c_{a,0}\right) \propto \frac{\left(a_{a,0}\right)^{-1-a}(b)^{a c_{a,0}}}{\Gamma(a)^{b_{a,0}}}.
\end{aligned} \tag{70}$$

Following Llera and Beckmann (2016), the hyperparameters are chosen as $a_{a,0}=1$, $b_{a,0}=c_{a,0}=a_{b,0}=b_{b,0}=0.01$, which specifies relatively uninformative priors for a and b. The corresponding segment of the posterior sampler is given as follows.

Algorithm 39. (Parametric Specification: Cross-sectional Heteroskedasticity) *For each iteration $s=1,\cdots,n_{sim}$,*

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

$$p\left(a^{(s)} \middle| b^{(s-1)}, \left\{\left(\sigma_i^2\right)^{(s-1)}\right\}\right) = p\left(a^{(s)} | b^{(s-1)}, a_a, b_a, c_a\right),$$

   which is characterized by the same kernel form as expression (70) with

$$\begin{aligned}
\log(a_a) &= \log(a_{a,0}) + \sum_{i=1}^{N} \log\left(\left(\sigma_i^2\right)^{(s-1)}\right), \\
b_a &= b_{a,0} + N, \\
c_a &= c_{a,0} + N.
\end{aligned}$$

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

$$\begin{aligned}
b^{(s)} &\sim Ga\left(a_b, b_b\right), \\
a_b &= a_{b,0} + N a^{(s)}, \\
b_b &= b_{b,0} + \sum_{i=1}^{N} \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1}.
\end{aligned}$$

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

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

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

	Degenerate	Skewed	Bimodal
Oracle	*0.250*	*0.289*	*0.270*
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 the MSE and the Qu *et al.* (2020) test. For the oracle predictor, the table reports the exact values of the 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.

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

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

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 Qu *et al.* (2020) 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 predictors outperform Heterosk-NP-C in terms of point forecasts, though

Table 8: Point Forecast Evaluation: General Model

		Normal v_{it}	Skewed v_{it}
Oracle		*0.492*	*0.486*
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: See the description in Table 7 for point forecast evaluation. Here the tests are conducted with respect to Heterosk-NP-C.

Table 9: Point Forecast Evaluation: Young Firm Dynamics

		MSE
Heterosk	*NP-C/R*	*0.197*
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 the MSE. For other predictors, the table reports their differences from Heterosk-NP-C/R.

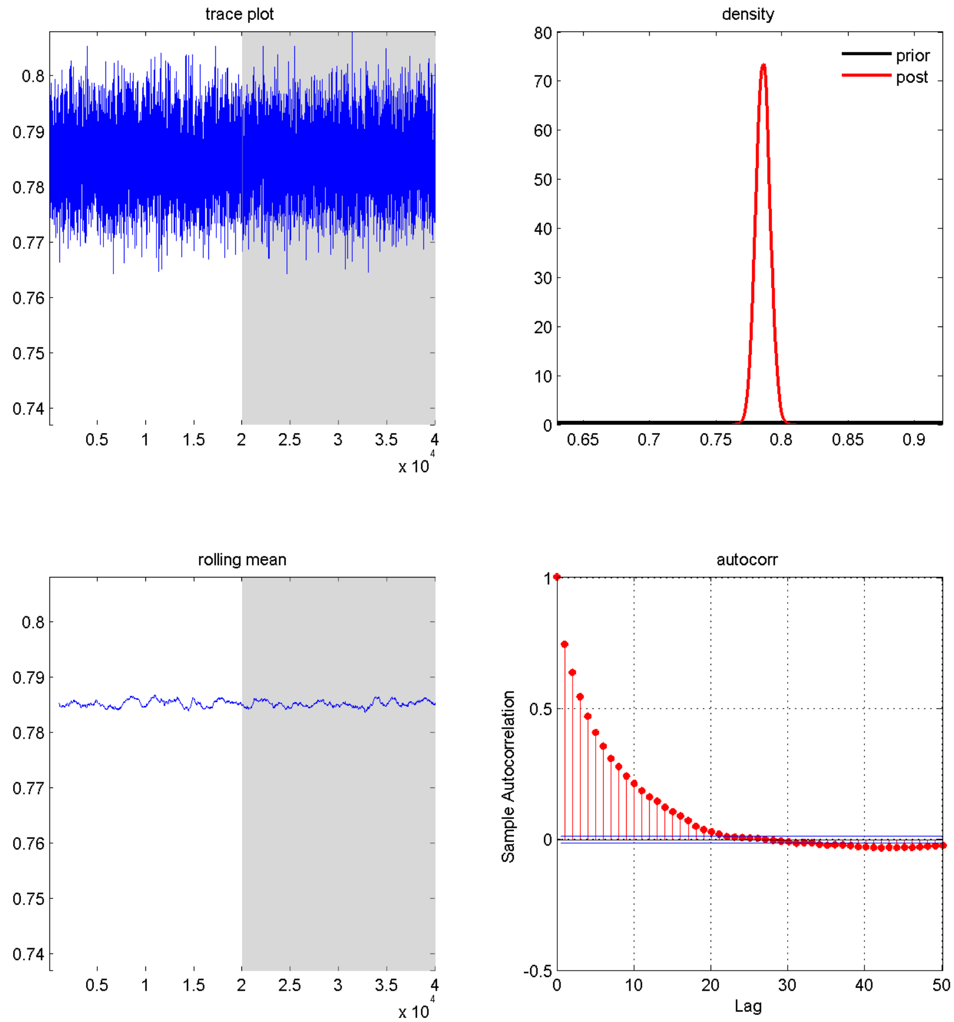
Heterosk-NP-C is correctly specified while the parsimonious ones are not.

Empirical Application. Most predictors are comparable according to the 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 $\left(\lambda_i, \sigma_i^2\right)$. Conditional on the common parameter β, the Flat estimator of $\left(\lambda_i, \sigma_i^2\right)$ is a Bayesian analog to individual-specific MLE/OLS that incorporates only firm i's own history, which is inadmissible under fixed T (Robbins, 1956; James and Stein, 1961; Efron, 2012).

E.2 Baseline Model with Random Effects

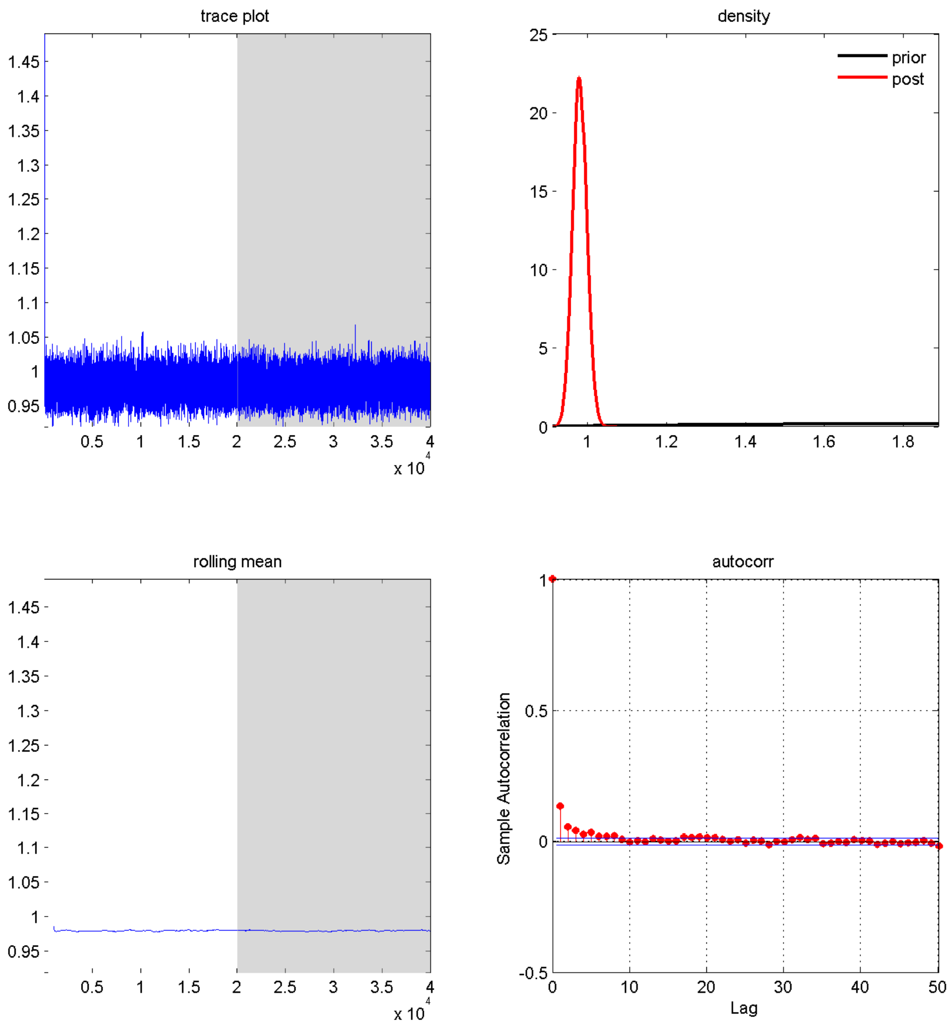
MCMC convergence. Both the Brook-Draper diagnostic and the Raftery-Lewis diagnostic yield desirable MCMC accuracy. Figures 7 to 10 show trace plots, prior/posterior distributions, rolling means, and autocorrelations of β, σ^2, α, and λ_i ($i = 1$).

Figure 7: Convergence Diagnostics: β



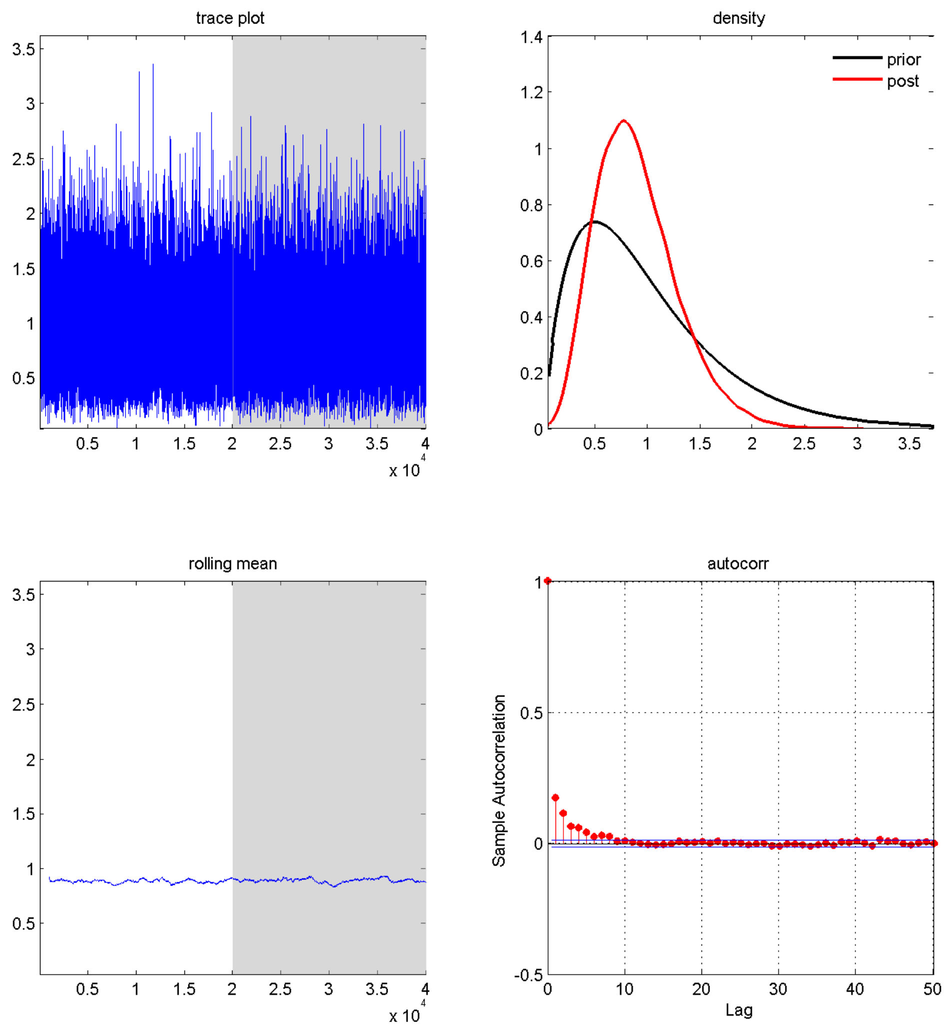
Notes: For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 8: Convergence Diagnostics: σ^2



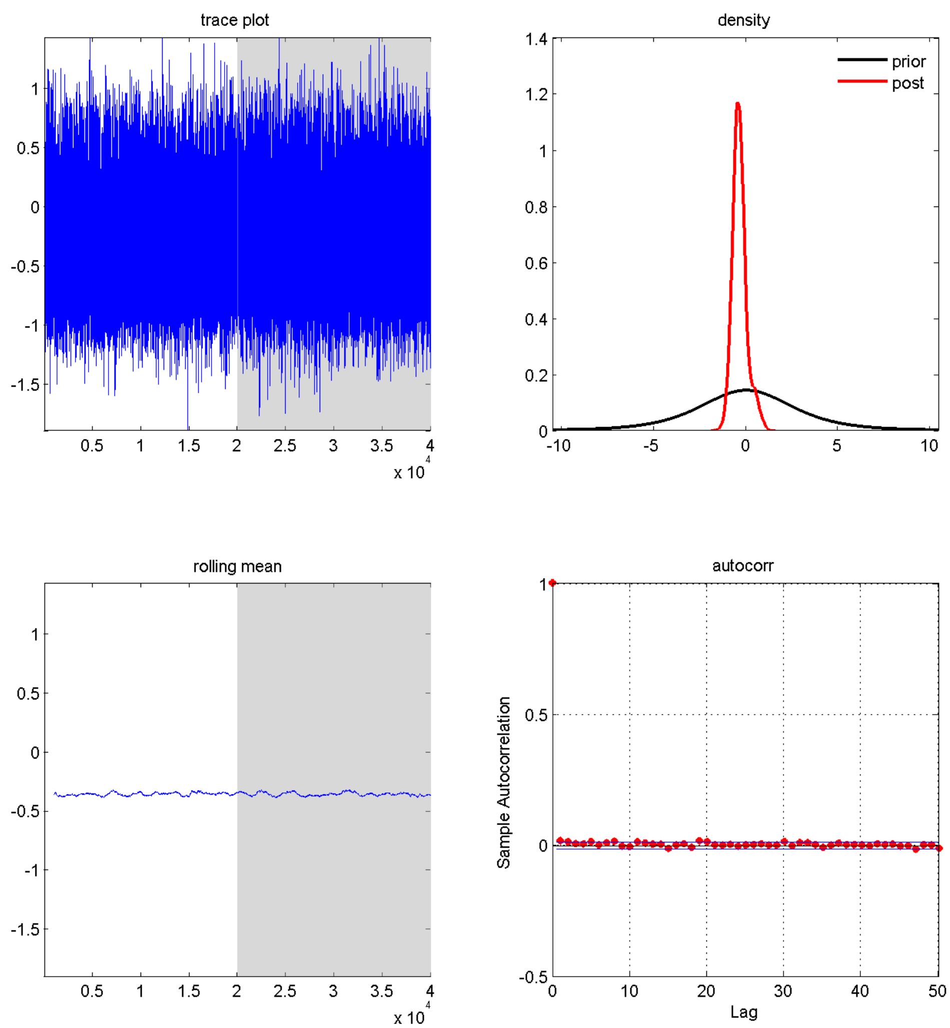
Notes: For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure 9: Convergence Diagnostics: α



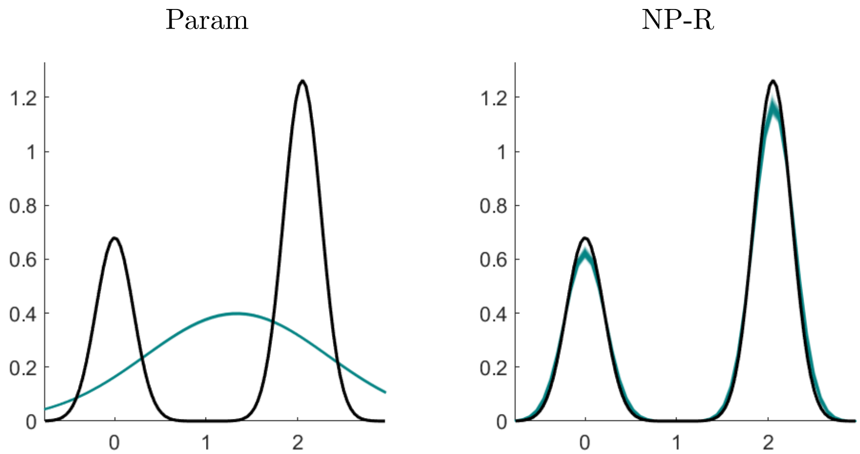
Notes: For each iteration s, rolling mean is calculated over the most recent 1000 draws.

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 $\Pi_f\left(f\,|y_{1:N,0:T}\right)$: 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, $\Pi_f\left(f\,|y_{1:N,0:T}\right)$.

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, and different underlying λ_i distributions (such as a normal distribution and a fat tail distribution). In general, NP-R 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 Param is only marginally better than NP-R in terms of both point and density forecasts. Intuitively, in the language of young firm dynamics, NP-R is preferable when the time series for a specific firm i is not informative enough to reveal its skill but the whole panel can help recover the skill distribution and hence firm i's uncertainty due to heterogenous skill. That is, NP-R works generally 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 overlap the true f_0 (see Figure 11), which resonates the posterior consistency result.

In terms of nonparametric Bayesian priors, 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 a DPM with skew normal components which better accommodates asymmetric DGPs. They provide minor improvement in the corresponding situations, but call for extra computational efforts.

E.3 Empirical Application

E.3.1 Additional Figures and Tables.

Below are additional figures and tables that supplement the main results in the text.

Figure 12: Distributions of Observables

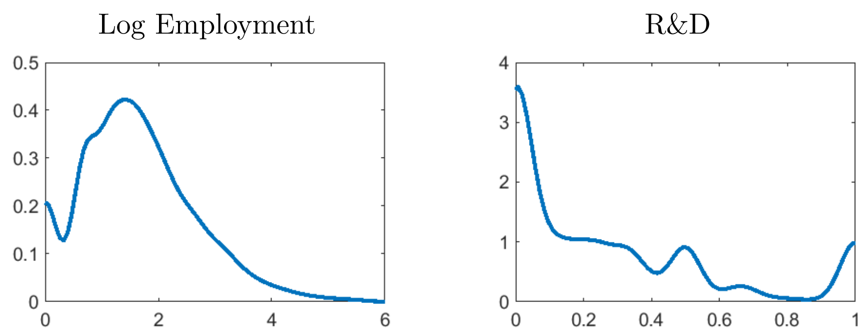
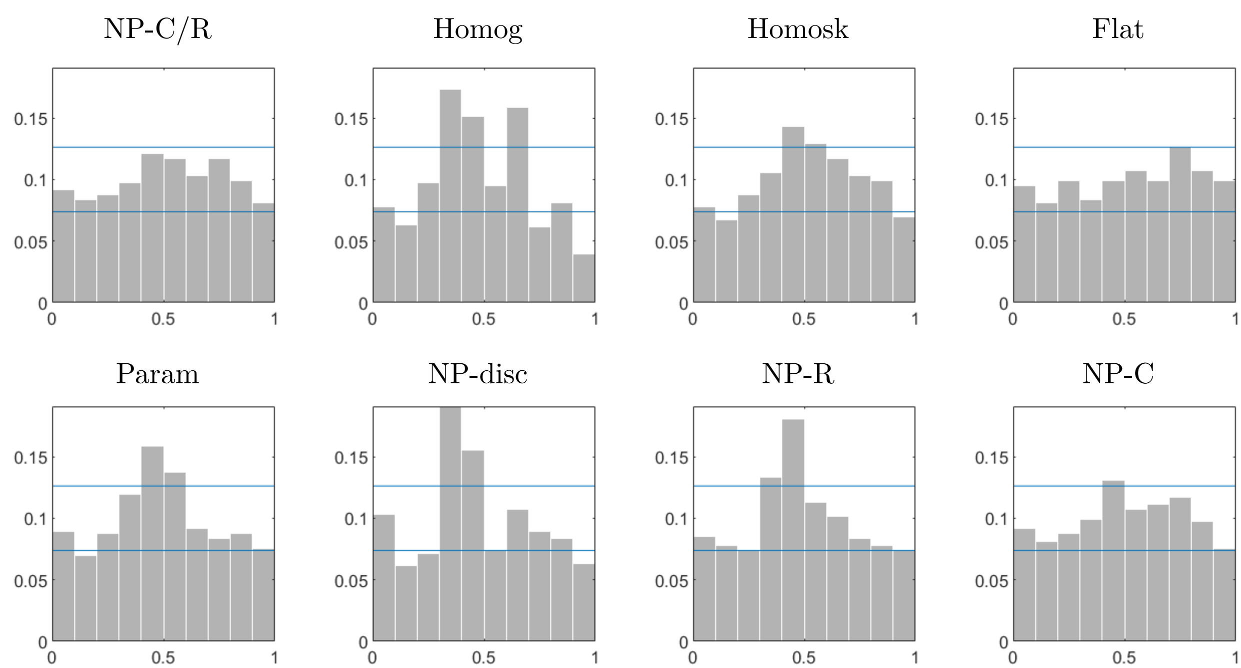
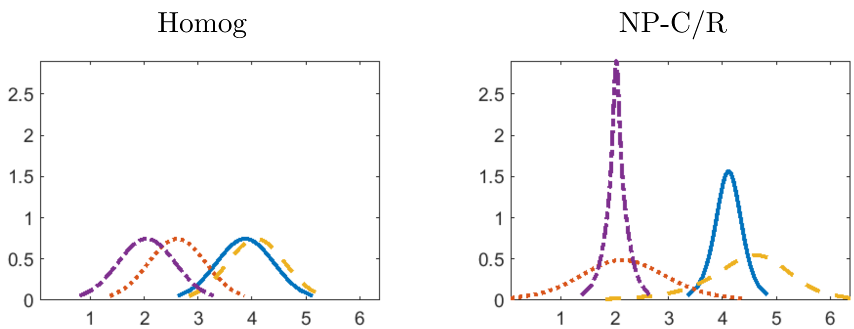


Figure 13: PIT: All Predictors



Notes: Teal lines indicate the confidence interval.

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

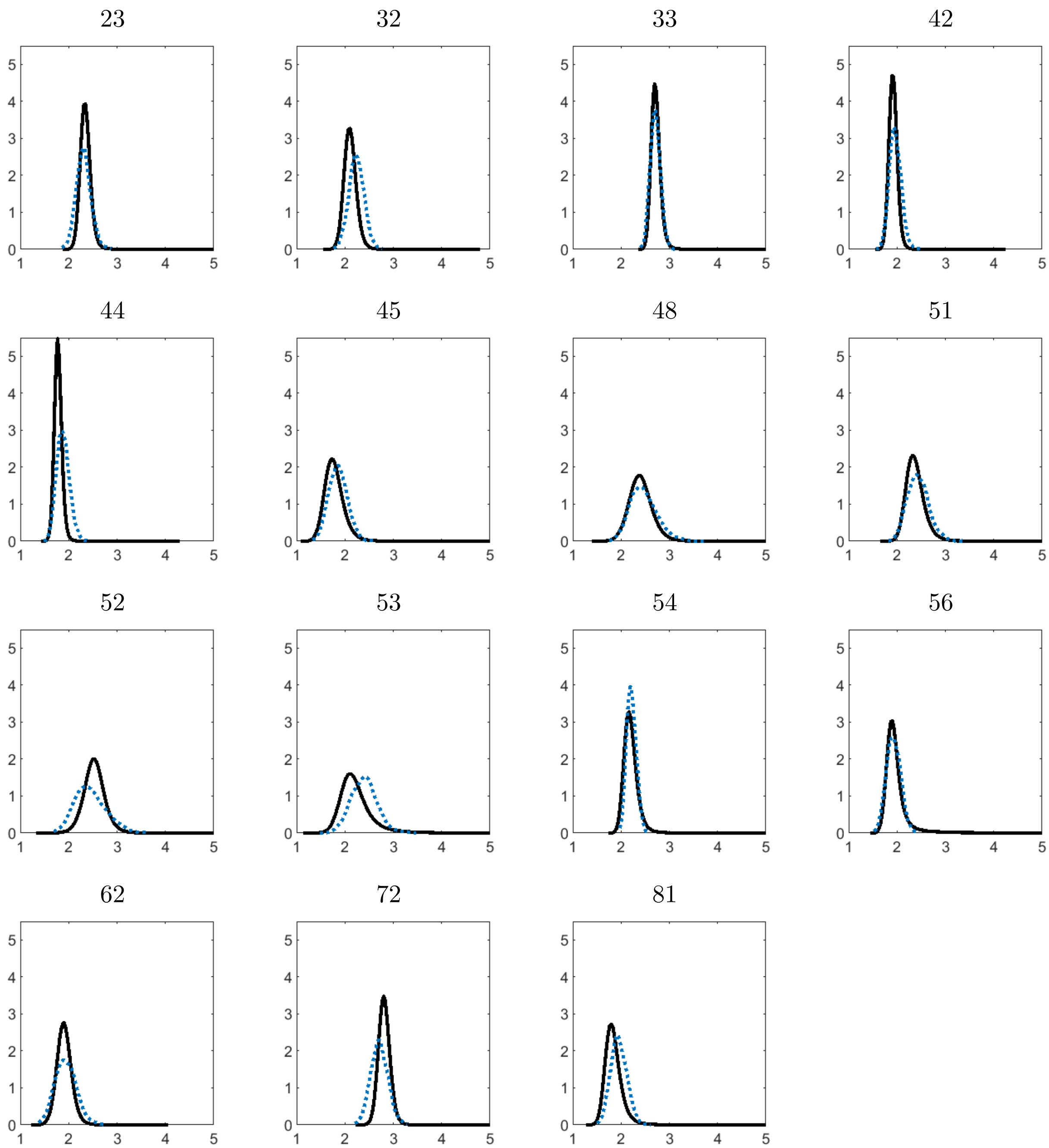


Notes: Predictive distributions are regrouped according to predictors. The blue solid / orange dotted / yellow dashed / purple dash-dot lines are the predictive distributions of typical firms a/b/c/d in Figure 4 in the main text, respectively.

Table 10: Two-digit NAICS Codes

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

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 log employment. I adopt log employment instead of the employment growth rate, as the latter significantly reduces the cross-sectional sample size due to the rank requirement. $\text{R\&D}_{it}$ is given by the ratio of a firm's R&D employment over its total employment considering that R&D employment has more complete observations compared with other innovation intensity gauges.

I have also explored other measures of firm performance (e.g. log revenue) and innovation activities (e.g. a binary variable on whether the firm has any R&D expenditure, and a discrete variable on 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 in the main text:

(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.[15]

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 predictors are comparable according to the MSE, with only Flat performing poorly in all three setups. For density forecasts, the overall best across all three setups is Heterosk-NP-C/R in the R&D setup. Comparing across setups, the one with the recession dummy 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.

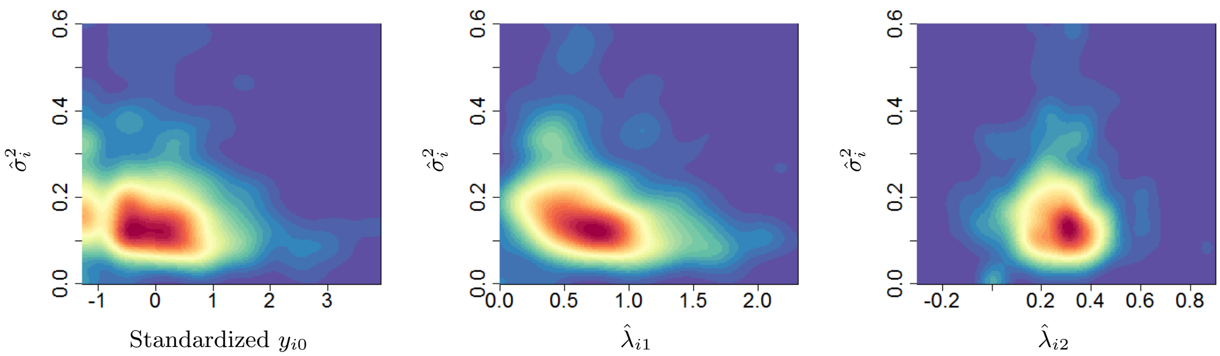
β Estimates in the Literature. Compared to the literature, the closest setup is Zarutskie and Yang (2015) using traditional panel data methods, where the estimated persistence of log employment is 0.824 and 0.816 without firm fixed effects (their Table 2) which is close to Homog, and 0.228 with firm fixed effects estimated via OLS (their Table 4) which is close 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 the 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 uncondtional correlations between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ and between $\hat{\sigma}_i^2$ and y_{i0} are roughly insignificant.

[15] I do not jointly incorporate recession and R&D because this 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.

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}
$\hat{\lambda}_{i1}$	—			
$\hat{\lambda}_{i2}$	**0.33 (0.00)**	—		
$\hat{\sigma}_i^2$	-0.08 (0.06)	0.02 (0.62)	—	
y_{i0}	**0.70 (0.00)**	**0.10 (0.02)**	**-0.10 (0.02)**	—

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.

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 between $\hat{\lambda}_i$ and $\hat{\sigma}_i^2$ conditioning on y_{i0}. The estimated $\hat{b}_1$ is -0.02 with the 95% interval being $[-0.09, 0.05]$, and $\hat{b}_2$ is 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 $\left(\hat{\lambda}_{i1}, \hat{\sigma}_i^2\right)\Big|\left(\hat{\lambda}_{i2}, y_{i0}\right)$, $\left(\hat{\lambda}_{i2}, \hat{\sigma}_i^2\right)\Big|\left(\hat{\lambda}_{i1}, y_{i0}\right)$, $\left(\hat{\lambda}_{i1}, \hat{\sigma}_i^2\right)\Big| y_{i0}$, and $\left(\hat{\lambda}_{i2}, \hat{\sigma}_i^2\right)\Big| y_{i0}$ 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$_\text{x}$ prior on $h_i = (\lambda_i', l_i)'$ mentioned in Appendix B.2. However, the density forecasts significantly deteriorate

Table 12: Density Forecast Evaluation: Robustness Checks

	Empirical	Monte Carlo
NP-C/R	-195	-1193
Cond. Correlated $\left(\lambda_i, \sigma_i^2\right)$	-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 in the general model with normal v_{it}.

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 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 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 as in the empirical application (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.

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