Forecasting with Bayesian Grouped Random Effects in Panel Data

Boyuan Zhang[†]

University of Pennsylvania

First Version: June 30, 2020

This Version: August 26, 2020

Abstract

In this paper, we estimate and leverage latent constant group structure to generate the point, set, and density forecasts for short dynamic panel data. We implement a nonparametric Bayesian approach to simultaneously identify coefficients and group membership in the random effects which are heterogeneous across groups but fixed within a group. This method allows us to flexibly incorporate subjective prior knowledge on the group structure that potentially improves the predictive accuracy. In Monte Carlo experiments, we demonstrate that our Bayesian grouped random effects (BGRE) estimators produce accurate estimates and score predictive gains over standard panel data estimators. With a data-driven group structure, the BGRE estimators exhibit comparable accuracy of clustering with the *Kmeans* algorithm and outperform a two-step Bayesian grouped estimator whose group structure relies on *Kmeans*. In the empirical analysis, we apply our method to forecast the investment rate across a broad range of firms and illustrate that the estimated latent group structure facilitates forecasts relative to standard panel data estimators.

JEL CLASSIFICATION: C11, C14, C23, C53, G31

KEY WORDS: Panel Data; Group Heterogeneity; Random Effects; Dirichlet Process; Set Forecast; Density forecast; Investment

[†]Department of Economics, Perelman Center for Political Science and Economics, University of Pennsylvania, 133 S. 36th St., Philadelphia, PA 19104-6297. Email: boyuanz@sas.upenn.edu. We would like to thank Karun Adusumilli, Francis Diebold, Philippe Goulet Coulombe, Akshay Malhotra, Frank Schorfheide for helpful comments and suggestions.

To-do list and Comments:

Frank Diebold:

(i) **[DONE]** Try different K, T in DGP 1 2 and 3.(fix T = 10, N = 100, K = 2, 4, 10)

(ii) Overlapping groups

(iii) **[DONE]** Check Askanasi, Diebold, Schorfheide and Shin 2018 JTSA. Issue with set forecast. Answer: (1) HPDI is the shortest well-calibrate interval but it is an undesirable inmplication as researcher doesn't receieve penalty for undercoverage and longer interval.

(iv) **[DONE]** The negative results in the paper suggests that abandoning set forecast might be benificially and, instead, focus on density forecast as densty forecast is richer. Conclusion: hard to compare set forecast across estimators.

(v) **[DONE]** See notes for the density forecast evaluation.

– "We also find that the Kmeans algorithm severely underestimates the number of groups under all data generating processes" is a very strong statement, presumably with far wider-ranging implications than your paper. Maybe you should write a second paper on the poor properties of kmeans as an unsupervised learning procedure. That is the fundamental issue. Your use of kmeans is just one particular application.

– Continuing on the above issue, later in the paper you take a much more measured tone, saying only that "Regarding clustering, the BGRE estimators generate comparable performance with Kmeans algorithm." You need to make up your mind, and defend your view.

– You "adopt a Bayesian approach and report the highest posterior density interval (HPDI), which is the narrowest connected interval with coverage probability of 1 - α". It's fine. I understand the Bayesian approach and result. But suppose that another method produces intervals with coverage probability of 1 - (α - ϵ) but length only L/2. Surely that would be preferable under most reasonable preferences. I still can't wrap myself completely around that. See Askanasi, Diebold, Schorfheide and Shin 2018 JTSA, who raise more questions than they answer, but still...

– Good use of CRPS and LPS for density forecast evaluation. See the attached (from a paper in progress) for some others and some sketches of relationships.

Karun:

(i) **[DONE]** Maybe you might want to mention the relationship of your paper to Liu (2020). I think Laura had a Dirichlet Process prior as well, but the model is different in your case?

(ii) **[PARTIALLY DONE]** It seems to me you only specify a group structure for the intercepts α, but the heterogeneous parameter β is treated as standard random effects. I think with only a bit more effort you can specify a group structure for the joint distribution of (α, β).

Frank Schorfheide:

[DONE] (for the cases of the subjective priors) allow for reallocation in the algorithm, change number of groups along the algorithm.

(check fuzzy c-means).

1 Introduction

With the increasing availability of panel data, many works have examined and demonstrated its central role in the empirical research throughout the social and business sciences. Analysis of panel data has various edges over that of pure cross-sectional or time-series data. The most important one is that the panel data provide researchers with a flexible way to model both heterogeneity among individuals, firms, regions, and countries and possible structural changes over time. Apart from the principal role in the model estimation, it is interesting and essential to study their relevance for forecasting. Among novel methods emerged recently, the latent group structure in the heterogeneity attracts wide attention. In this paper, we allow for grouped patterns of unobserved heterogeneity in the dynamics panel data models. We aim to evaluate whether this latent structure improves the predictive performance in an extensive collection of short time series.

In the dynamics panel data model, it is common to assume that each cross-sectional unit has unique intercept. This assumption introduces a large number of parameters that become a burden in estimation. In models that have as many parameters as individual units, fixed effects estimators are known to suffer from the "incidental parameters" problem (Neyman and Scott, 1948), which can bring about significant biases in estimates of common parameters. This problem becomes severe in short panels even if the number of units goes to infinity (Chamberlain 1980, Nickell 1981), and the fixed-effects themselves are often poorly estimated. The econometric model, its estimates, and forecasts are all intertwined. An unreliable estimate leads to concerns about the predictive power of panel data models as biased estimates affect not only the point forecast but also the set and density forecast, all of which are our main focus in the paper.

To address this issue[1], econometricians attempt to reduce the number of unknown parameters by dividing units into a finite number of groups. The premise of this idea is that units in the same group share the unit-specific parameters. Previous works include Lin and Ng (2012), Bonhomme and Manresa (2015), Ando and Bai (2016), Su, Shi, and Phillips (2016), Bester and Hansen (2016), Su, Wang, and Jin (2019) and Bonhomme, Lamadon, and Manresa (2019). Moreover, finite mixture model provide a well-known probabilistic approach to model-based clustering (McNicholas and Murphy 2010, Frühwirth-Schnatter 2011). With

[1]Another important strand of literature implements generalized method of moments (GMM) methods to eliminate bias, see Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998). Though successfully solved the "incidental parameters" problem, this set of methods doesn't allow for any latent group structure.

a finite number of groups, econometricians could avoid "incidental parameter" bias under several particular assumptions and derive consistent estimators for the common parameters.

However, the convenience of the group structure does not come without any cost. The number of groups is an unknown but fixed quantity, and the need to specify the number in advance is deemed one of the significant drawbacks of applying these methods in a clustering context. Many methods have been suggested to estimate the optimal number a posteriori from the data such as BIC (Keribin 2000, Bonhomme and Manresa 2015), marginal likelihoods (Frühwirth-Schnatter, 2004), or the integrated classification likelihood (Biernacki, Celeux, and Govaert, 2000). Bayesian approaches sometimes pursue a similar strategy, often adding the DIC to the list of model choice criteria, e.g., Celeux, Forbes, Robert, and Titterington (2006) and Kim and Wang (2019). If both N and T are large enough, the information criterion could select the true group structure. However, under a short time span, these criteria might fail to achieve their goal. As noted in Bonhomme and Manresa (2015), the choice of the number of groups is crucial to estimation and inference for model parameters. Misspecification in group number forces the algorithm to consider incorrect group membership. We will later show that it is the information criteria that substantially affects the performance of Grouped Fixed Effects (GFE) estimator proposed by Bonhomme and Manresa (2015).

To avoid using the information criterion, we treat the number of groups as an unknown parameter that is estimated jointly with the component-specific parameters from a Bayesian viewpoint. In general, it is determined throughout the whole sampling period, which makes our model attractive in comparison with existing panel data models. We follow the Bayesian approach proposed by Kim and Wang (2019), who is considered as the Bayesian version of Bonhomme and Manresa (2015). In particular, we don't truncate the number of components at a predetermined level but leave the largest possible number unrestricted, and directly estimate the optimal partition under the assumption that group membership remains constant over time. We implement the slice-sampling approach proposed by Walker (2007), through which the posterior sampling reduces to a finite-dimensional practicable procedure. This method leads to a simple Bayesian framework where a straightforward MCMC sampling procedure is applied to estimate the unknown number of components, determine cluster-relevant variables, and perform component-specific inference at the same time.

Armed with a proper Bayesian approach, we step further and examine whether such a grouped Bayesian estimator helps us make predictions for short panel data. Bonhomme and Manresa (2015) and its Bayesian counterpart Kim and Wang (2019) aim at estimating

the heterogeneous coefficients and understanding the finite-sample properties of the grouped estimator. None of them explore the potential merit of the grouped estimator in terms of forecasting. Our work will fill in this area and examine the performance of grouped random effects panel data estimator in various scenarios. Besides, our full Bayesian analysis allows us to conduct a density forecast, which gives us the richest insight regarding future prediction. To be precise, a density forecast provides a predictive distribution of future values for each unit that incorporates uncertainties of common/heterogeneous parameters, grouped heterogeneity, and future innovations.

The contributions of this paper are fourfold. First, closely following Kim and Wang (2019) and Liu (2020), we develop a posterior sampling algorithm that addresses the nonparametric estimation of latent grouped effects and proposes Bayesian Grouped Random Effects (BGRE) estimator. Instead of using the Finite Mixture model, which needs to specify the number of groups a priori, we use Dirichlet Process (DP) prior, in particular the stick-breaking prior, that allows for infinite potential groups and direct estimation of the number of groups along with parameters of interest. By using the DP prior, we circumvent the need to explicitly specify ahead of time how many clusters there are, although the concentration parameter in the DP prior still controls it implicitly. The entire posterior sampler is constructed following the blocked Gibbs sampler[2] proposed by Ishwaran and James (2001).

Second, we leverage the researcher's prior knowledge of the latent group structure to facilitate estimation and forecasting. Depending on the degree of expertise, we summarize and incorporate the information of subjective group structure in the prior distribution of random effects or membership probabilities. If the subjective prior on the group structure is more precise than the random guess, even with incorrect presumed number of groups, we show that including it in the prior improves the BGRE estimators' performance by guiding the group membership estimates.

Third, we explore the potential link between the proposed BGRE estimators and unsupervised machine learning method. Theoretically, we show that our block Gibbs sampler for the BGRE estimator is closely related to the *Kmeans* algorithm (MacQueen, 1967) under certain assumptions. In particular, both algorithms assign units to the closest centroid when forming the clusters and recalculate the means of the new cluster afterward. To compare the performance of clustering, we modify our algorithm to incorporate *Kmeans* and construct a

[2]Unlike the Pólya urn Gibbs sampler (Escobar and West, 1995), blocked Gibbs sampler approach avoids marginalizing over the prior and thus allows for direct sampling of the nonparametric posterior, leading to computational and inferential advantages.

two-step BGRE estimator where individuals are clustered in the first step using *Kmeans*, and the group-specific heterogeneity is estimated in the second step. In the simulation section, we document that our BGRE estimators dominate the two-step GRE estimator in terms of the performance of both clustering and forecasting. We also find that the two-step GRE estimators with *Kmeans* algorithm severely underestimate the number of groups under all data generating processes, whereas BGRE estimators deliver accurate estimates.

Last but not least, we examine the performance of BGRE estimators using various sets of simulated data and real data. The Monte Carlo study presents that grouped heterogeneity brings gains in estimating group structure and one-step ahead point, set, and density forecasting relative to commonly used predictors with different parametric priors on individual effects. In particular, our estimators outperform the rising star — Grouped Fixed Effects (GFE) estimator proposed by Bonhomme and Manresa (2015) in various settings of the data generating process. The better performance is primarily due to the accurate estimate of the group structure. Regarding other predictors, we show that failing to model group structure and to pool information across units severely deteriorates the results for both estimation and forecasting. Finally, we use our method to forecast the investment rate across a broad range of firms.

Our paper relates to several branches of the literature. Our work is closely related to Bonhomme and Manresa (2015), Bonhomme, Lamadon, and Manresa (2019) and Kim and Wang (2019). All of these three papers aim to estimate the unobserved heterogeneity in a linear dynamic panel data model and develop statistical inference methods. Bonhomme and Manresa (2015) estimates the parameters of the model using the GFE estimator that minimizes the least-squares criterion for all possible groupings of the cross-sectional units. They jointly estimate the individual types and the model's parameter given the number of groups and performance model selection afterward. One the other hand, Bonhomme, Lamadon, and Manresa (2019) modify this method and split the procedure into two steps with *Kmeans* clustering algorithm is used in the first step. From Bayesian's point of view, Kim and Wang (2019) proposes a full Bayesian estimator that simultaneously estimates the group structure and parameters. They are viewed as a Bayesian counterpart of Bonhomme and Manresa (2015) since they replicate the latter's empirical analysis. Unfortunately, none of these works examine the potential forecasting gain when considering the group structure.

The paper that most related to ours is Liu (2020). She considers a linear dynamic panel data model and implements the finite mixture model to estimate the underlying distribution of unit-specific intercepts. However, the assumptions of the underlying model in Liu (2020)

are different from ours. She assumes fully heterogeneous intercepts in a panel data model, which amounts to the standard random-effects model. The finite mixture model in her setting serves as a method to pool information across units. In our work, we specify a group-specific intercept in the population level, and the semiparametric method is the critical ingredient to deliver not only the estimates but also the group structure. Moreover, her main object of interest is to construct individual-specific density forecasts for a panel, while our work includes point, set, and density forecasts and, most important, view the performance of group clustering as an essential assessment on top of forecasts.

This paper also relates to the literature on nonparametric Bayesian approach in the group structure estimation problem. We model the unknown distribution of the heterogeneous coefficient (including grouped intercept and innovation variance) as the Dirichlet Process of Normals with potential infinite groups. The idea of sampling from the Dirichlet Process Model has been widely used by a number of authors including Escobar and West (1995), Neal (2000), Ishwaran and James (2001), Molitor, Papathomas, Jerrett, and Richardson (2010), Yau, Papaspiliopoulos, Roberts, and Holmes (2011), Hastie, Liverani, and Richardson (2015), Liverani, Hastie, Azizi, Papathomas, and Richardson (2015), Liu, Moon, and Schorfheide (2019), and Liu (2020). To make the infinite-dimensional problem operable, our blocked Gibbs sampler (which based on Ishwaran and James (2001)) relies on the slice sampling described by Walker (2007), a more efficient version was later proposed by Kalli, Griffin, and Walker (2011).

We proceed as follows. In section 2, we present the specification of a linear dynamic panel data model and discuss the construction and evaluation of point, set, and density forecasts. Section 3 provides details on nonparametric Bayesian priors and subjective group structure priors. It also documents the posterior Gibbs sampler and shows its connection to the *Kmeans* algorithm. We conduct various Monte Carlo experiments in section 4 to examine the performance of the proposed estimator in a controlled environment in the light of point, density, and set forecasts. We also examine the performance of a few variants of the BGRE estimator. In section 5, we conduct empirical analysis in which we forecast the investment rate across firms. Finally, we conclude in section 6. A description of the data sets, additional empirical results, and derivations are relegated to the appendix.

2 The modeling framework

2.1 Model

We consider a panel with observations for cross-sectional units $i = 1, \ldots, N$ in periods $t = 1, \ldots, T$. Given a panel data set $\{(y_{it}, x_{it})\}$, a simple linear dynamic panel data model with grouped patterns of heterogeneity takes the following form:

$$y_{it} = \alpha_{g_i t} + \rho y_{it-1} + \beta_i' x_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_{g_i}^2\right). \tag{2.1}$$

where x_{it} are a $p \times 1$ vector of exogenous variables, they are uncorrelated with ε_{it} but is allowed to be arbitrarily correlated with $\alpha_{g_i t}$. $\alpha_{g_i t}$ denote the time-varying group-specific heterogeneity. The subscript $g_i \in \{1, ..., K\}$ is the group membership variable with unknown and unconstrained K. y_{it-1} is the lagged outcome variable. ρ is the homogeneous AR(1) parameters that are common for all cross-sectional units, and β_i is a $p \times 1$ vector of unit-specific slope coefficients. ε_{it} is the idiosyncratic error term featured by zero mean and grouped heteroskedasticity $\sigma_{g_i}^2$, with cross-sectional homoskedasticity being a special case where $\sigma_{g_i}^2 = \sigma^2$. This setting leads to a heterogeneous panel with group pattern modeled through $\alpha_{g_i t}$ and $\sigma_{g_i}^2$.

By stacking all observations for unit i, we get an aggregated model:

$$y_i = \alpha_{g_i} + \rho y_{i,-1} + x_i \beta_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N\left(\mathbf{0}, \mathbf{\Sigma}_{g_i}\right). \tag{2.2}$$

where $y_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]'$, $y_{i,-1} = [y_{i0}, y_{i1}, \ldots, y_{iT-1}]'$, $T \times 1$, $x_i = [x_{i1}, x_{i2}, \ldots, x_{iT}]'$, $\alpha_{g_i} = [\alpha_{g_i 1}, \alpha_{g_i 2}, \ldots, \alpha_{g_i T}]'$, $\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}]'$, $\mathbf{\Sigma}_{g_i} = \sigma_{g_i}^2 \mathbf{I}_T$. To indicate the component from which each observation stems, we introduce a group membership variable $G = [g_1, \ldots, g_N]$ taking values in $\{1, \ldots, K\}^N$. Define a set of unit that belongs to group k: $C_k = \{i \in \{1, 2, ..., N\} | g_i = k\}$. Let $|C_k|$ denote the cardinality of the set C_k.

Following Sun (2005), Lin and Ng (2012) and Bonhomme and Manresa (2015), we assume that individual group membership does not vary over time. In addition, for any group $i \neq j$, we assume that they have different paths of random effects, e.g., $\alpha_i \neq \alpha_j$, and no single unit can simultaneously belong to these two groups: $C_i \bigcap C_j = \varnothing$.

The main goal of this paper is to estimate the grouped random effects α_{g_i}, common parameter ρ, hetergenous coeffecients β_i and group membership G using full sample and provide the point, set, and density forecasts of y_{it+h} for each unit i. Throughout this paper, we focus on the one-step ahead forecast where $h = 1$. For the multiple-step forecast, the

procedure can be extended by iterating y_{iT+h} in accordance with (2.1) given the estimate of parameters and realizations of data.

2.2 Estimation and Forecast Evaluation

2.2.1 Posterior Predictive Densities

Our goal is to generate one-step ahead forecasts of $y_{i,T+1}$ for $i = 1, ..., N$ conditional on the history of observations,

$$Y = [y_1, y_2, ..., y_N], y_i = [y_{i1}, y_{i2}, ..., y_{iT}]',$$
$$X = [x_1, x_2, ..., x_N], x_i = [x_{i1}, x_{i2}, ..., x_{iT}]'.$$

and newly available exogenous variables x_{iT+1} at $T + 1$. For illustration purpose, we drop both X and x_{iT+1} from notations but we always condition on these exogenous variables.

The posterior predictive distribution for unit i is given by

$$p(y_{iT+1}|Y) = \int p(y_{iT+1}|Y, \Theta)p(\Theta|Y)d\Theta, \tag{2.3}$$

where Θ is a vector of parameters $\Theta = \{\rho, \beta_i, \alpha_{g_i}, \Sigma_{g_i}, g_i\}$. This density is the posterior expectation of the following function,

$$p(y_{iT+1}|Y, \Theta) = \sum_{k=1}^{K} \mathbf{1}(g_i = k)p\left(y_{iT+1}|Y, \rho, \beta_i, \alpha_k, \Sigma_k\right), \tag{2.4}$$

which is invariant to relabeling the components of the mixture. Therefore, given M^* posterior draws, the density estimated from the MCMC draws is

$$\hat{p}(y_{iT+1}|Y) = \frac{1}{M^*} \sum_{j=1}^{M^*} \left(\sum_{k=1}^{K^{(j)}} \mathbf{1}(g_i = k)p\left(y_{iT+1}|Y, \rho^{(j)}, \beta_i^{(j)}, \alpha_k^{(j)}, \Sigma_k^{(j)}\right) \right). \tag{2.5}$$

Therefore, we can draw samples from $\hat{p}(y_{iT+1}|Y)$ by simulating (2.1) forward conditional on the posterior draws of Θ and observations.

2.2.2 Point Forecasts

We evaluate the point forecasts via the Root Mean Square Forecast Error (RMSFE) under the quadratic loss function averaged across units. Let $\hat{y}_{iT+1}$ represent the predicted value

conditional on the observed data up to period T, the loss function is written as

$$L\left(\widehat{y}_{1:N,T+1}, y_{1:N,T+1}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_{iT+1} - y_{iT+1}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{iT+1}^2, \tag{2.6}$$

where $y_{i,T+1}$ is the realization at $T+1$ and $\hat{\varepsilon}_{iT+1}$ denote the forecast error. Then, the formula for RMSFE is provided in the following equation,

$$RMSFE = \sqrt{L\left(\widehat{y}_{1:N,T+1}, y_{1:N,T+1}\right)}. \tag{2.7}$$

The optimal posterior forecast under quadratic loss function is obtain by minimizing the posterior risk,

$$\begin{aligned}\hat{y}_{1:N,T+1} &= \underset{\hat{y}\in\mathbb{R}^N}{\operatorname{argmin}} \int_{-\infty}^{\infty} L\left(\hat{y}, y_{1:N,T+1}\right) p(y_{1:N,T+1}|Y) dy_{1:N,T+1} \\ &= \underset{\hat{y}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{y} - y_{iT+h}\right)^2 |Y\right].\end{aligned} \tag{2.8}$$

This implies optimal posterior forecast is the posterior mean,

$$\hat{y}_{i,T+1} = \mathbb{E}\left(y_{iT+1}|Y\right), \text{ for } i = 1, \ldots, N. \tag{2.9}$$

Conditional on posterior draws of parameters $j = 1, 2, ..., M^* = M - M_{\text{burn-in}}$, the mean forecast can be approximated by Monte Carlo averaging,

$$\mathbb{E}\left(y_{iT+1}|Y\right) \approx \frac{1}{M^*}\sum_{j=1}^{M^*}\mathbb{E}\left(y_{iT+1}^{(j)}|Y, \rho^{(j)}, \beta_i^{(j)}, \alpha_{g_i}^{(j)}, \sigma_{g_i}^{2,(j)}\right). \tag{2.10}$$

For the one-step ahead forecast,

$$\mathbb{E}\left(y_{iT+1}^{(j)}|Y, \rho^{(j)}, \beta_i^{(j)}, \alpha_{g_i}^{(j)}, \sigma_{g_i}^{2,(j)}\right) = \alpha_{g_i}^{(j)} + \rho^{(j)} y_{iT} + \beta_i^{\prime(j)} x_{iT+1}, \tag{2.11}$$

and hence,

$$\mathbb{E}\left(y_{iT+1}|Y\right) \approx \bar{\alpha}_{g_i} + \bar{\rho} y_{iT} + \bar{\beta}_i^{\prime} x_{iT+1}. \tag{2.12}$$

where $\bar{\alpha}_{g_i}$, $\bar{\rho}$ and $\bar{\beta}_i^{\prime}$ are means of posterior draws.

2.2.3 Set Forecasts

We construct set forecasts CS_{iT+1} from the posterior predictive distribution of each unit. In particular, we adopt a Bayesian approach and report the highest posterior density interval (HPDI), which is the narrowest connected interval with coverage probability of $1-\alpha$. Put

differently, it requires that the probability of $y_{iT+h} \in CS_{i,T+1}$ conditional on having observed the history Y is at least $1-\alpha$, i.e.,

$$P(y_{iT+1} \in CS_{iT+1}) \geqslant 1-\alpha, \quad \text{for all } i, \tag{2.13}$$

and this interval is the shortest among all possible single connected candidate sets. Let δ^l be the lower bound and δ_u be the upper bound, then $CS_{iT+1} = \left[\delta_i^l, \delta_i^u\right]$.

The assessment of set forecasts in simulation studies and empirical applications is based on two metrics: (1) the cross-sectional coverage frequency,

$$Cov_{T+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left\{y_{iT+1} \in CS_{iT+1}\right\}, \tag{2.14}$$

and (2) the average length of the sets $C_{i,T+1}$,

$$AvgL_{T+1} = \frac{1}{N}\sum_{i=1}^{N}(\delta_i^u - \delta_i^l). \tag{2.15}$$

2.2.4 Density Forecasts

To compare the performance of density forecast for various estimators, we examine the continuous ranked probability score (CRPS) across units. The CRPS is frequently used to assess the respective accuracy of two probabilistic forecasting models. It is a quadratic measure of the difference between the forecast cumulative distribution function (CDF), $F_i^{T+1}(y)$, and the empirical CDF of the observation with the formula as follows,

$$\begin{aligned} CRPS_{T+1} &= \frac{1}{N}\sum_{i=1}^{N} CRPS(F_i^{T+1}, y_{iT+h}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_0^{\infty}\left(F_i^{T+1}(y) - \mathbb{I}\left\{y_{iT+h} \leqslant y\right\}\right)^2 dy, \end{aligned} \tag{2.16}$$

where y_{iT+h} is the realization at $T+1$.

In practice, the true forecast cumulative distribution function $F_i^{T+1}(y)$ or the PIT of y_{iT+1} is not available to econometricians. We approximate it via the empirical distribution function for each unit based on the posterior draws from the predictive density,

$$\hat{F}_i^{T+1}(x) = \frac{1}{M^*}\sum_{j=1}^{M^*} \mathbb{I}\left\{y_{iT+1}^{[j]} \leqslant x\right\}, \tag{2.17}$$

where $y_{iT+h}^{[j]}$ denotes the j-th element of the sorted sequence $\{y_{iT+h}^{(j)}\}_{j=1}^{M*}$. The merit of using the empirical distribution function is that we can express the integral as a Riemann sum as both terms inside the integral are step functions.

Moveover, we report another metric called the average log predictive scores (LPS) to assess the performance of the density forecast from the view of the probability distribution function (PDF). As suggested in Geweke and Amisano (2010), the LPS for a panel reads as,

$$LPS_{T+1} = \frac{1}{N}\sum_{i=1}^{N} \ln p\left(y_{iT+1}|Y\right), \tag{2.18}$$

where the log-predictive density can be approximated by,

$$\ln p\left(y_{iT+1}|Y\right) \approx \ln\left(\frac{1}{M*}\sum_{j=1}^{M*} p\left(y_{iT+1}|\mu_{iT+1}^{(j)}, \sigma_{g_i}^{2,(j)}\right)\right), \tag{2.19}$$
$$\mu_{iT+1}^{(j)} = \alpha_{g_i}^{(j)} + \rho^{(j)} y_{iT} + \beta_i^{'(j)} x_{iT+1}.$$

2.2.5 Estimation

To evaluate the statistical superiority of pooling within K clusters, we report the bias, standard deviation, average length of 95% credible set, and frequentist coverage of the posterior mean estimate of ρ across Monte Carlo repetitions. For the random effects α, we only present the average bias as it may not be of interest for most empirical analysis. All these metrics are averaged over 100 Monte Carlo repetitions.

To estimate the number of groups, we derive a point estimator from its posterior distribution, typically, the posterior mean, which is consistent with a quadratic loss function. In the empirical analysis, we also consider the posterior mode suggested by Malsiner-Walli, Frühwirth-Schnatter, and Grün (2016), which is equal to the most frequent number of non-empty components visited during MCMC sampling. These approaches constitute an automatic and straightforward strategy to estimate the unknown number of groups without using model selection criteria or marginal likelihoods.

2.3 Extension

(TBW) It seems to me you only specify a group structure for the intercepts α, but the heterogeneous parameter β is treated as standard random effects. I think with only a bit

more effort you can specify a group structure for the joint distribution of (α, β).

$$y_{it} = \theta'_{g_i} \breve{x}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma^2_{g_i}\right) \tag{2.20}$$

where $\breve{x}_{it} = [1 \; y_{it-1} \; x'_{it}]'$, and $\theta_{g_i} = [\alpha_{g_i} \; \rho_{g_i} \; \beta'_{g_i}]'$.

With a joint prior for θ_k is $N(\theta_k, \Sigma_k)$, we can modify our block Gibbs Sampler to draw α, ρ and β simultaneously. The derivation of the posterior distributions are presented in Appendix C.3. NEED SIMULATION.

3 Bayesian Estimation

In this section, we provide details in Bayesian analysis. In Section 3.1, we document the specification of the prior distribution for all parameters, including the auxiliary variable in the random coefficient model, and the subjective group prior if econometricians have prior knowledge on group structure. Section 3.2 present the posterior sampler and algorithm is shown in the Appendix C.1.1. Finally, we provide preliminary thoughts on the connection between our Bayesian method and unsupervised machine learning methods in Section 3.3.

3.1 Nonparametric Bayesian Prior

Two sets of specifications of prior distributions are considered in this section. In the first prior specification, we concentrate on random effects models and implement a full Bayesian analysis. In addition, we specify a hyperprior for the distribution of unobserved heterogeneity and then construct a joint posterior for the coefficients of this hyperprior as well as the actual unit-specific and common coefficients. In the second specification, econometricians could provide useful information on the latent group structure and incorporate it in the prior.

3.1.1 Random Effects Model

In this paper, we focus on the random coefficients model where heterogeneous parameters $\alpha_{g_i t}$ and σ_{g_i} are independent and are assumed to be independent of the initial value of each unit y_{i0}. The specification can be extended to correlated random coefficients model by modeling the joint distribution of heterogeneous parameters and initial values y_{i0}.

A typical choice in the nonparametric Bayesian literature is the Dirichlet Process (DP) prior or stick-breaking prior. With group probabilities π_k and parameter in prior: (mean,

variance) $= (\mu_\alpha, \Sigma_\alpha)$, a draw of $\alpha_{g_i t}$ from the DP prior could be viewed as a mixture of point mess with the probability mass function,

$$\alpha_{g_i t} \sim \sum_{k=1}^{K} \pi_k \delta_{\alpha_{kt}}, \text{ with } \alpha_{kt} \sim N(\mu_\alpha, \Sigma_\alpha), \tag{3.1}$$

where δ_x denotes the Dirac-delta function concentrated at x, each α_{kt} is drawn from a normal distribution and K is unknown. μ_α are set to the OLS estimate of α assuming $K = 1$ and Σ_α equals $200 \times \hat{\Sigma}_\alpha$ where $\hat{\Sigma}_\alpha$ is the standard deviation of the OLS estimator. In the same fashion, we can define the DP prior for grouped heteroskedasticity $\sigma^2_{g_i}$ given identical group probabilities π_k:

$$\sigma^2_{g_i} \sim \sum_{k=1}^{K} \pi_k \delta_{\sigma^2_k}, \text{ with } \sigma^2_k \sim IG\left(\frac{\nu_\sigma}{2}, \frac{\delta_\sigma}{2}\right), \nu_\sigma = 12, \text{ and } \delta_\sigma = 10, \tag{3.2}$$

where each component is drawn from inverse-Gamma distribution.

Put together, the posterior draws of grouped related coefficients can be characterized by a grouped triplet $\{\pi_k, \alpha_k, \sigma^2_k\}$ for $k = 1, 2, ..,$ and $\alpha_k = [\alpha_{k1}, \alpha_{k2}, ..., \alpha_{kT}]'$. Importantly, the distributions of both α_k and σ^2_k are discrete, because draws can only take the values in the set $\{(\alpha_k, \sigma^2_k) : k \in \mathbb{Z}^+\}$. This nonparametric nature makes the Dirichlet Process prior an ideal choice for clustering problems especially when the distinct number of clusters is unknown beforehand. The group parameters (α_k, σ^2_k) are assumed to follow the base distribution B_0 which is an independent (non-conjugate) Multivariate Normal-Inverse-Gamma (IMNIG) distribution.

On the other hand, the group probability is formalized through an infinite-dimensional stick-breaking prior governed by the concentration parameter a,

$$\pi_k = \xi_k \prod_{j<k} (1 - \xi_j) \text{ for } k > 1, \text{ and } \pi_1 = \xi_1, \tag{3.3}$$

where ξ_k, which are called stick lengths, are independent random variables drawn from the beta distribution $Beta(1, a)$. This construction can be viewed as a stick-breaking procedure, where at each step, we independently and randomly break the leftover of a stick of unit length and assign the length of this break to the current value of π_k. The smaller a is, the less of the stick will be left for subsequent values (on average), yielding more concentrated distributions.

The concentration parameter a specifies how strong this discretization is. As $a \to 0$, the realizations are all concentrated at a single value, while when $a \to \infty$, the realizations become continuous-valued from its based distribution. Escobar and West (1995) shows that

the number of estimated groups under a DP prior is sensitive to a, which indicates that a data-driven estimate should be more reasonable. To determine how discrete we want and how many groups are needed given the data, it is convenient to treat a as a parameter under the nonparametric Bayesian framework. Put differently, we can set up a relatively general hyperprior for $a \sim Gamma\,(0.4, 10)$, and update it based on the observations. This step generates a posterior estimate of a, which implicitly chooses the optimal K without re-estimating the models with different numbers of groups.

Finally, the prior distribution for the common parameter ρ is chosen to be a normal distribution,

$$\rho \sim N(0, \sigma_\rho^2) \text{ with } \sigma_\rho^2 = 100. \tag{3.4}$$

The prior of heterogeneous parameter β_i follows,

$$\beta_i \sim N(0, \Sigma_\beta) \text{ with } \Sigma_\beta = 100 \times \mathbf{I}_p. \tag{3.5}$$

To sum up, in the random coefficients model, we specify the Dirichlet Process priors for group random effects $\alpha_{g_i t}$ and heteroskedasticity $\sigma_{g_i}^2$, a stick-breaking process for group probabilities π_k, a hyperprior for the concentration parameter a and a normal prior for the common parameter ρ and heterogeneous parameter β_i.

3.1.2 Subjective Priors With Knowledge on Groups

Frequently, researchers could provide a group structure on all or at least part of the units based on personal expertise and the nature of individuals. For example, firms coming from the same industry may share a similar growth pattern; countries having the same level of development form comparable fiscal policies. Though this presumed group structure might be subjective or purely based on theoretical analysis, it is still valuable to integrate this group information as it guides estimation when it enters the algorithm via a prior.

Depending on the degree of knowledge, we leverage the information of prior group structure in two ways. If the true number of groups is unknown, but the researcher can at least partition N units into $\widetilde{K}$ groups, then we can impose group structure on the prior distribution of random effects α's. In the full Bayesian analysis, we adopt a common prior for α_i for all i. But now we need to differentiate units and assume that units in the same subjective group share identical prior mean and variance in the base distribution. We call this prior for α_i as *Subjective Group Member (SGM) Prior*. In practice, for those units to enjoy the same

group, we pool their observations and run a standard OLS regression to find the prior mean and variance of α_k. On the other hand, if there are units left unclassified, they are automatically assigned to a new group called "TBA". Parameter estimation is then straightforward: least squares estimation can be applied directly to this group.

If the number of groups is known a priori or a theoretical restriction is imposed on the number of groups, we introduce the prior for the membership probability $\omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{iK^p}]'$ for all unit $i = 1, .., N$, where K^p is the presumed number of groups. Namely, before estimating the group membership, the researcher assigns each unit to different groups with a set of subjective group-specific probabilities, and these probabilities will enter the algorithm through a prior distribution for ω_i. We name this prior for ω_i as *Subjective Group Probability (SGP) Prior*. In practice, one could provide a table (for example, Table 1) documenting the subjective group probability of a unit falling into a specific group.

Table 1: Example of prior group probability

	Group			
Unit	1	2	3	4
1	0.75	0.20	0.05	0
2	0.30	0.30	0.20	0.20
3	0	0	0.50	0.50
4	0	1	0	0
⋮	⋮	⋮	⋮	⋮

Building on the prior for the random effects model in Section 3.1.1, we allow for incorporating the researchers' prior knowledge while inheriting the feature of reallocating units and changing the number of groups along the MCMC sampling. These flexible features enable the block Gibbs sampler to automatically correct and update the imprecise subjective prior, especially when K^p doesn't match the true number of groups.

To incorporate these subjective group probabilities, it is important to choose a proper prior for ω. Dirichlet distribution is an applicable candidate among assorted densities since it is the conjugate prior of the multinomial distribution, which facilitates direct sampling and provides a natural channel to integrate subjective group probability.

To see this, we use the simplest case where the number of potential groups (K^*) in an iteration equals the presumed K^p. Let $\omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{iK^p}]$ be the vector of group-specific

probability for unit i, we set the prior density for ω_i as an unsymmetric Dirichlet distribution,

$$\omega_i \sim Dir(a_{i1}, a_{i2}, ..., a_{iK^p}), \tag{3.6}$$

where a_{ik} are concentration parameters and strictly positive. Conditional on ω_i, the group membership g_i is assumed to be drawn from a multinomial distribution,

$$g_i \sim Multinomial(\omega_i), \text{ i.e., } P(g_i = k|\omega_i) = \omega_{ik}, \text{ for } k = 1, \ldots, K^p. \tag{3.7}$$

It can be shown posterior probability of ω_i given g_i is also a Dirichlet distribution with modified parameters: $Dir(a_{i1} + \mathbf{1}(g_i = 1), a_{i2} + \mathbf{1}(g_i = 2), ..., a_{iK^p} + \mathbf{1}(g_i = K^p))$.

Another important property of Dirichlet distribution that enables itself to be the most suitable prior is that we can tie our prior probability directly with the expected value of ω_{ik},

$$E(\omega_{ik}) = \frac{a_{ik}}{\sum_{i=k}^{K^p} a_{ik}}. \tag{3.8}$$

To integrate the researcher's prior knowledge, one only need to deliberately choose a set of $\{a_{ik}\}$ such that the expected probability matches her subjective probability on groups.

Since the membership probabilities ω_{ik} are updated based on observations, and we allow for reallocating units and changing the number of groups along the MCMC sampling, a revision of our block Gibbs sampler is needed to adjust for such changes. The details of the new algorithm are presented in Appendix C.2. In practice, we can restrict $\sum_{i=k}^{K^p} a_{ik}$ to be 1 so that a_{ik} represents both the subjective group probability for unit i belonging to group k and the prior mean of ω_{ik}.

3.2 Posterior Sampling

Draws from the joint posterior distribution for the can be obtained by using blocked Gibbs sampling. We will subsequently describe the conditional distributions over which the Gibbs sampler iterates[3]. We focus on the time-varying grouped random effects model with grouped heteroskedasticity, which is the most sophisticated specification. Other specifications can be estimated by merely ignoring time effects in α's or shutting down the heteroskedasticity.

The blocked Gibbs sampler is based on Ishwaran and James (2001) and Walker (2007). Though the algorithm in Ishwaran and James (2001) has been widely used for sampling stick-breaking priors, it alone can't fulfill our need for estimating the number of groups

[3]To avoid being overwhelmed by formulas in the main text, we provide detailed derivations for the posterior sampler in the Appendix C.

without any predetermined level or upper bound since it requires a finite-dimensional prior and truncation. To avoid approximation and a predetermined number of groups K, we implement slice-sampling proposed by Walker (2007) and modify the framework of Ishwaran and James (2001) with additional posterior sampling steps.

To facilitate derivation, we stack observations and parameters,

$$\begin{aligned}
\text{Observations: } & Y = [y_1, y_2, ..., y_N], y_i = [y_{i1}, y_{i2}, ..., y_{iT}]', \\
\text{Covariates: } & X = [x_1, x_2, ..., x_N], x_i = [x_{i1}, x_{i2}, ..., x_{iT}]', \\
\text{Random effects: } & \alpha = [\alpha_1, \alpha_2, \ldots], \\
\text{Covariance matrices: } & \Sigma = [\sigma_1^2, \sigma_2^2, \ldots], \\
\text{Hetergeneous coefficients: } & \beta = [\beta_1, \ldots, \beta_N], \\
\text{Stick length: } & \Xi = [\xi_1, \xi_2, \ldots], \\
\text{Group membership: } & G = [g_1, \ldots, g_N], \\
\text{Auxiliary varaible: } & u = [u_1, u_2, ..., u_N], \\
\text{Hyper parameters: } & \phi = [\mu_\alpha, \Sigma_\alpha, \nu_\sigma, \delta_\sigma].
\end{aligned}$$

The posterior of unknown objects in the random effects model is,

$$\begin{aligned}
& p(\rho, \beta, \alpha, \Sigma, \Xi, a, G|Y, X) \\
\propto\ & p(Y|X, \rho, \beta, \alpha, \Sigma, G)p(\rho, \beta, \alpha, \Sigma, \Xi, a, G) \\
\propto\ & p(Y|X, \rho, \beta, \alpha, \Sigma, G)p(\alpha, \Sigma|\phi)p(\Xi|a)p(G|\Xi)p(\rho)p(\beta)p(a) \\
=\ & \prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2) \prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi) \prod_{j=1}^{\infty} p(\xi_j|a) \prod_{i=1}^{N} p(g_i|\Xi)p(\rho) \prod_{i=1}^{N} p(\beta_i)p(a) \\
=\ & \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2)p(g_i|\Xi)p(\beta_i)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)p(\xi_j|a)\right] p(\rho)p(a). \qquad (3.9)
\end{aligned}$$

Walker (2007) augments the posterior distribution with a set of auxiliary variables $u = [u_1, u_2, ..., u_N]$, which are i.i.d. standard uniform random variables, i.e, $u_i \sim U(0, 1)$. Then

the augmented posterior is written as,

$$\begin{aligned} & p(\rho, \beta, \alpha, \Sigma, \Xi, a, G, u|Y, X) \\ \propto & \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2)\mathbf{1}(u_i < \pi_{g_i})p(\beta_i)\right]\left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)p(\xi_j|a)\right]p(\rho)p(a) \\ = & \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2)p(u_i|\pi_{g_i})\pi_{g_i}p(\beta_i)\right]\left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)p(\xi_j|a)\right]p(\rho)p(a), \end{aligned} \tag{3.10}$$

where $\pi_{g_i} = p(g_i|\Xi)$, and $\mathbf{1}(\cdot)$ is the indicator function, which is equal to zero unless the specific condition is satisfied. The original posterior can be recovered by integrating out u_i for $i = 1, 2, ..., N$. As we don't limit the upper bound of the number of groups, it is impossible to sample from an infinite-dimensional posterior density. The merit of slice-sampling is that it reduces the dimensions and allows us to solve a manageable problem with finite dimensions, which we will see below.

With a set of auxiliary variables $u = [u_1, u_2, ..., u_N]$, we define the largest possible number of potential components as

$$K^* = \min_k \left\{\sum_{j=1}^{k} \pi_j > 1 - u^*\right\}, \tag{3.11}$$

where

$$u^* = \min_{1 \leqslant i \leqslant N} u_i. \tag{3.12}$$

Such specification ensures that for any group $k > K^*$ and any unit $i \in \{1, 2, ..., N\}$, we have $u_i > \pi_k$[4]. This crucial property limits the dimension of α and Σ to K^* as the densities of α_k and σ_k equal 0 for $k > K^*$ due to $\mathbf{1}(u_i < \pi_k) = 0$, which will be clear in the subsequent posterior derivation.

Next, we define the number of active groups

$$K^a = \max_{1 \leqslant i \leqslant N} g_i. \tag{3.13}$$

It can be shown that $K^a \leqslant K^*$[5].

Conditional posterior of α (grouped random effects):

$$p(\alpha|\rho, \beta, \Sigma, G, Y, X) \propto \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2)\mathbf{1}(u_i < \pi_{g_i})\right]\left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)\right] \tag{3.14}$$

[4]See proof in proposition A.1
[5]See proof in proposition A.1

Since we assume an independent normal conjugate prior for α_k, the posterior for α_k is also normal. If group k is empty, we draw α_k from its prior $N(\mu_\alpha, \Sigma_\alpha)$.

Conditional posterior of Σ (grouped variance): Under the cross-sectional independence, for $k \in \{1, 2, ..., K^*\}$,

$$p(\sigma_k^2|\rho, \beta, \alpha, G, Y, X) \propto \left[\prod_{i \in C_k} p(y_i|x_i, \rho, \beta_i, \alpha_k, \sigma_k^2)\right] p(\sigma_k^2|\phi), \tag{3.15}$$

where C_k is defined as a set of units that belongs to group k,

$$C_k = \{i \in \{1, 2, ..., N\} | g_i = k\}, \tag{3.16}$$

In essence, we pool all the units of group k to estimate the group-specific coefficients. The conjugate prior for $\sigma_k^2 \sim IG\left(\frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$ makes it possible to directly sample from the posterior distribution of σ_k^2 which is also an inverse Gamma distribution. If group k is empty, we draw σ_k^2 from its prior.

Conditional posterior of ρ (common AR(1) coefficient).

$$p(\rho|\beta, \alpha, \Sigma, G, Y, X) \propto \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2)\right] p(\rho) \tag{3.17}$$

Using a normal conjugate prior $\rho \sim N(\mu_\rho, \sigma_\rho^2)$, we could solve the following standard Bayesian linear regression to get the posterior density of the common coefficient ρ, for $i = 1, ..., N$ and $t = 1, .., T$,

$$y_{it} - \alpha_{g_i,t} - x_i\beta_i = y_{i,t-1}\rho + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma_{g_i}^2).$$

The independence of the ε_{it} allows us to pool observations across i and t. The posterior distribution of ρ is also normal.

Conditional posterior of β (heterogeneous coefficients).

$$p(\beta|\rho, \alpha, \Sigma, G, Y, X) \propto \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2) p(\beta_i)\right] \tag{3.18}$$

As ε_{it} is independent across units, we solve for β for each unit separately. We transform the model into a standard linear model with a known form of heteroskedasticity,

$$y_{it} - \alpha_{g_i t} - y_{it-1}\rho = \beta_i' x_{it} + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma_{g_i}^2).$$

Using a normal conjugate prior $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$, we can directly sample β_i from its posterior distribution.

Conditional posterior of Ξ (stick length):

$$p(\Xi|\rho,\beta,\alpha,\Sigma,a,G,u,Y,X) \propto \left[\prod_{i=1}^{N}\xi_{g_i}\prod_{l<g_i}(1-\xi_l)\right]\left[\prod_{j=1}^{\infty}p(\xi_j|a)\right] \tag{3.19}$$

The prior for the stick length ξ_k is assumed to be a Beta distribution, which shares the same functional form as the stick-breaking process. To see this, for $k = 1, 2, ..., K^a$, the posterior of ξ_k is of the form,

$$\begin{aligned} p(\xi_k|\rho,\beta,\alpha,\Sigma,a,G,u,Y,X) &\propto \left(\prod_{i\in C_k}\xi_k\right)(1-\xi_k)^{\sum_{j=1}^{N}\mathbf{1}(g_j>k)}(1-\xi_k)^{a-1} \\ &\propto \xi_k^{|C_k|}(1-\xi_k)^{a+\sum_{j=1}^{N}\mathbf{1}(g_j>k)-1}. \end{aligned} \tag{3.20}$$

Therefore, the posterior distribution of ξ_k is also a Beta distribution,

$$\xi_k|\rho,\beta,\alpha,\Sigma,a,G,u,Y,X \sim Beta\left(|C_k|+1, a+\sum_{j=1}^{N}\mathbf{1}(g_j>k)\right). \tag{3.21}$$

Conditional posterior of a (concentration parameter). Regarding the concentration parameter for the Dirichlet process, the standard posterior derivation fails due to the unrestricted number of components in the current sampler. Instead, closely following Escobar and West (1995), we implement the 2-step procedure proposed by introducing a latent variable η. According to their approach, we first draw a latent variable η from

$$\eta \sim Beta\,(a+1, N). \tag{3.22}$$

Conditional on η and K^a we sample a from a mixture of two Gamma distribution:

$$a|\eta,K^a \sim \begin{cases} Gamma\,(m+K^a, n-\log(\eta)) & \text{with prob. } \pi_a \\ Gamma\,(m+K^a-1, n-\log(\eta)) & \text{with prob. } 1-\pi_a \end{cases},$$

with the probability π_a defined by

$$\frac{\pi_a}{1-\pi_a} = \frac{m+K^a-1}{N[n-\log(\eta)]}. \tag{3.23}$$

Conditional posterior of u (auxiliary variable). Conditional on the group stick lengths' ξ_k and group membership G, it is straightforward to show that the posterior density of u_i is a uniform distribution defining on $(0, \pi_{g_i})$,

$$u_i|\Xi,G \sim Unif\,(0,\pi_{g_i}), \tag{3.24}$$

where $\pi_{g_i} = \xi_{g_i}\prod_{j<g_j}(1-\xi_j)$.

Conditional posterior of G (group membership). We derive the posterior distribution of g_i conditional on $G^{(i)}$, where $G^{(i)}$ is a set including all member indices except for g_i, i.e., $G^{(i)} = G \backslash g_i$. Hence, for $k = 1, 2, ..., K^*$,

$$p(g_i = k|\rho, \beta, \alpha, \Sigma, \Xi, a, G^{(i)}, u, Y, X) \propto p(y_i|\rho, \beta_i, \alpha_k, \sigma_k^2, Y, X)\mathbf{1}(u_i < \pi_{g_i}). \tag{3.25}$$

As per a discrete distribution, we normalize the point mass to get a valid distribution:

$$p(g_i = k|\rho, \beta, \alpha, \Sigma, \Xi, a, G^{(i)}, u, Y, X) = \frac{p(y_i|\rho, \beta_i, \alpha_k, \sigma_k^2, Y, X)\mathbf{1}(u_i < \pi_k)}{\sum_{j=1}^{K^*} p(y_i|\rho, \beta_i, \alpha_j, \sigma_j^2, Y, X)\mathbf{1}(u_i < \pi_j)}. \tag{3.26}$$

3.3 Potential Link to Unsupervised Learning

In section 3.2, we articulate the detailed steps of the block Gibbs sampling, which aims to partition N units into G groups and, at the same time, generate posterior draws of parameters. This Gibbs sampler and our BGRE estimator inevitably remind us of one of the most popular clustering algorithms in the area of unsupervised machine learning: Kmeans algorithm. Indeed, the Kmeans algorithm plays a central role in Bonhomme and Manresa (2015) and Bonhomme, Lamadon, and Manresa (2019) who estimate the grouped fixed-effects from the frequentists' point of view. In this section, we seek to illustrate the similarity and connection between our block Gibbs sampling and the Kmeans algorithm in the limit.

We start with the Kmeans clustering algorithm. Given a set of observations $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$, where each observation contains the dependent variables and covariates, (y_i, x_i'). Kmeans clustering aims to partition the N observations into K sets so as to minimize the within-cluster sum of squares,

$$\min_{\{C_k\}_{k=1}^K} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{z}_i - \boldsymbol{\mu}_k\|^2 \text{ where } \boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \boldsymbol{z}_i \tag{3.27}$$

The algorithm alternates between reassigning points to clusters and recomputing the means. For the assignment step, one computes the squared Euclidean distance from each point to each cluster mean, and then assign each observation to the cluster with the nearest mean. The update step of the algorithm recalculates centroid for observations assigned to each cluster and updates $\boldsymbol{\mu}_k$ for all k.

For our block Gibbs sampler, we assign unit i to group k conditional on the draws of

other parameters (Eq. (3.26)) with probability,

$$\begin{aligned}
&p(g_i = k|\rho, \beta, \alpha, \Sigma, G^{(i)}, Y, X) \\
&= \frac{p(y_i|\rho, \beta_i, \alpha_k, \sigma_k^2, Y, X)\mathbf{1}(u_i < \pi_k)}{\sum_{j=1}^{K^*} p(y_i|\rho, \beta_i, \alpha_j, \sigma_j^2, Y, X)\mathbf{1}(u_i < \pi_j)}. \\
&= \frac{c_{ik} \exp\left[-\frac{1}{2}(y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_k)'\Sigma_k^{-1}(y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_k)\right]}{\sum_{j=1}^{K^*} c_{ij} \exp\left[-\frac{1}{2}(y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_j)'\Sigma_j^{-1}(y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_j)\right]} \\
&= \frac{c_{ik} \exp\left[-\frac{1}{2}(\tilde{y}_i - \alpha_k)'\Sigma_k^{-1}(\tilde{y}_i - \alpha_k)\right]}{\sum_{j=1}^{K^*} c_{ij} \exp\left[-\frac{1}{2}(\tilde{y}_i - \alpha_j)'\Sigma_j^{-1}(\tilde{y}_i - \alpha_j)\right]},
\end{aligned} \tag{3.28}$$

where $c_{ik} = (2\pi)^{-\frac{T}{2}}\Sigma_k^{-\frac{1}{2}}\mathbf{1}(u_i < \pi_k)$, $\tilde{y}_i = y_i - \rho y_{-1,i} - x_i\beta_i$, and $y_{-1,i}$ are the lagged values of y_i. If we assume homoskedasticity, i.e., $\Sigma_k = \Sigma$ for all k, then in the limit as $\Sigma \to 0$, the value of $p(g_i = k|\rho, \beta, \alpha, \Sigma, G^{(i)}, Y, X)$ approaches zero for all k except for the one corresponding to the smallest weighted distance $(\tilde{y}_i - \alpha_k)'\Sigma_k^{-1}(\tilde{y}_i - \alpha_k)$. In this case, the step is akin to the assignment step of Kmeans but using a weighted Euclidean distance. Then conditional on newly estimated group membership, we update the group random effects α_k through Bayesian linear regression only using the units of group k. This step exactly recalculates the means of the new clusters, establishing the equivalence of the update step.

Having this similarity in mind, it is natural to include the Kmeans algorithm in our Monte Carlo experiment and explore its performance relative to our BGRE estimator in terms of accuracy of clustering. Notably, following Bonhomme, Lamadon, and Manresa (2019), we construct a 2-step GRE estimator equipped with K-mean algorithm in the first step. The performance of this 2-step estimator is assessed in the section 4.3.3.

4 Monte Carlo Simulation

In this section, we conducted Monte Carlo simulation experiments to examine the performance of various Grouped Random Effects (GRE) estimators under different data generating processes (DGPs) and prior assumptions. These DGPs differ in whether the random effects are time-invariant or time-varying and whether to introduce heterogeneity in the variance of innovations. Such designs allow us to examine not only how our approach performs under DGPs with particular features, but also the reliability of appropriately estimating the number of data clusters.

We consider a setting with sample size $N = 100$, and time span $T = 11$. The last

observation of each unit forms the hold-out sample for evaluation as we focus on one-step ahead forecasts. A similar framework can be applied to multiple-step-ahead forecasts by iterating from period to period. We set the true numbers of groups $K^0 = 4$. Given N and K^0, we partition the entire sample into K^0 balanced blocks with N/K_0 units in each block[6]. For each DGP, 100 datasets are generated, and we run the block Gibbs samplers for each data set with $M = 10,000$ iterations after a burn-in of 5,000 draws.

4.1 Data Generating Process

The Monte Carlo simulation is based on the dynamic panel data model in (2.1), in which we suppress the exogenous predictors x_{it} for simplicity. In short, we consider four linear dynamic DGPs in this section. DGP1 and DGP2 involve time-invariant random effects while time-varying random effects are allowed in DGP3. Moreover, DGP1 and DGP3 consider homoskedasticity, but DGP2 has heteroskedastic innovations. DGP4 is the standard panel data model without a group structure. Throughout these DGPs, the random effects α_k and idiosyncratic error ε_{it} are standard normal distributed, independent across i, k, and t, and mutually independent. ε_{it} is independent of all regressors. The data are simulated according to the following DGPs:

DGP1: Time-invariant grouped random effects, homoskedasticity **(Grp Ti-Homo)**. This DGP is the most naive panel data model with group pattern in the random effect.

$$y_{it} = \alpha_{g_i} + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\varepsilon_{it} \overset{iid}{\sim} N(0, 0.8)$ and $\alpha_k \overset{iid}{\sim} N(k, 0.5^2)$ for $k = 1, 2, ..., K^0$.

DGP2: Time-invariant grouped random effects, heteroskedasticity **(Grp Ti-Hetero)**. This DGP aims to incorporate heteroskedasticity, which leads to a slightly complicated process that is hard to estimate.

$$y_{it} = \alpha_{g_i} + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma^2_{g_i}\right)$ where $\sigma^2_k = 1.5\left(1 - \frac{k-1}{K^0}\right)^2$, and $\alpha_k \overset{iid}{\sim} N(k, 0.5^2)$ for $k = 1, 2, ..., K^0$.

DGP3: Time-varying grouped random effects, homoskedasticity **(Grp Tv-Homo)**. So far, we have focused on time-invariant models. But when estimated on real data, it's reasonable

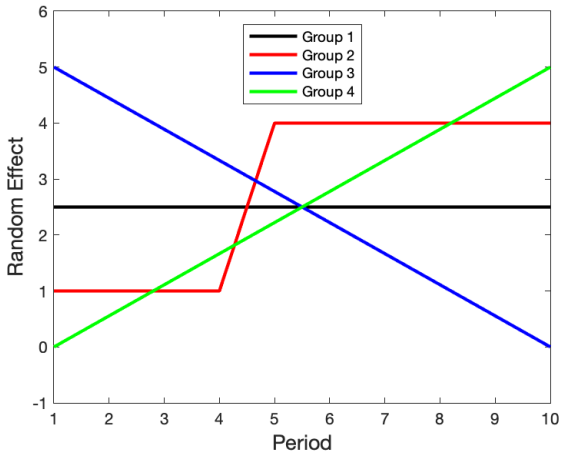
[6] If N/K_0 is not an integer, use $\lfloor N/K_0 \rfloor$ for group 1,2,..,$K_0 - 1$, the last group contains the residual units.

to believe the random effect could a have time pattern. Hence, we introduce various time-varying patterns of the random effects while keeping the assumption of homoskedasticity to avoid over-parameterization.

$$y_{it} = \alpha_{g_i} + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\varepsilon_{it} \overset{iid}{\sim} N(0,1)$ and $\alpha_{it} \overset{iid}{\sim} N\left(\underline{\alpha}_{g_i t}, 0.5^2\right)$ where $\underline{\alpha}_{g_i t}$ varies across periods and groups as depicted in Figure 1. To enrich the patterns of time-varying random effects, we construct 4 different paths. Group 1 has a constant prior mean for α_i. Group 2 and group 3 are equipped with monotonically increasing/decreasing prior means. The prior means for α_i in group 4 are constant but experience a structure change at $T = 5$.

Figure 1: Mean of Random Effects for GDP3, $K^0 = 4$



DGP4: Time-invariant random effects, homoskedasticity, no group structure **(Std Ti-Homo)**. This is the standard panel data model with unit-specific random effects and identical variance for the innovations.

$$y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\varepsilon_{it} \overset{iid}{\sim} N(0, 0.8)$ and $\alpha_i \overset{iid}{\sim} N(0, 0.5^2)$.

4.2 Simulation Results

We consider four types of Bayesian Grouped Random Effect (BGRE) estimators that differ on the assumptions made on the random effects (RE) and the variance of errors: (1) time-

invariant grouped RE with homoskedasticity (Ti-Homo); (2) time-invariant grouped RE with heteroskedasticity (Ti-Hetero); (3) time-varying grouped RE with homoskedasticity (Tv-Homo); (4) time-varying grouped RE with heteroskedasticity (Tv-Hetero)[7]. For instance, Ti-Homo estimator assumes the true model to have time-invariant grouped random effects, and the variance of error terms to be constant across units. Besides the results we will show below, we modify the DGPs and conduct more experiments using (a) larger variance of error terms σ^2, (b) shorter time span, and (c) different true number of groups K^0. These additional results are available in Appendix E.

Regarding alternative estimators, we consider the following Bayesian estimators that have different prior assumptions on the random effects α_i. (1) Bayesian pooled estimator (Pooled): α_i is treated as a common parameter as ρ does, this means all units share the same prior level of α_i; (2) flat-prior estimator (Flat): assume $p(\alpha_i) \propto 1$, this amounts to draw samples from a posterior whose mode is the MLE estimate. Given the estimate of common parameter, there is no pooling across units, α_i's are estimated only using their own history; (3) Parametric-prior estimator (Param): assume $\alpha_i \sim N(\mu, \pi^2)$, where a Normal-Inverse-Gamma hyperprior is further imposed on (μ, π^2)[8], this prior be thought of as a limit case of the DP prior when the concentration parameter $a \to 0$, so there is only one cluster, and (μ, π^2) are directly drawn from the base distribution.

4.2.1 Point Estimates

Table 2 shows the estimate comparison among alternative predictors. For the DGP1, Ti-Homo and Ti-Hetero estimator are the best in every aspect. This is as expected since they correctly model the time-invariant random effects. Among these two estimators, when we allow for group-level heteroskedasticity, the optimal number of groups decreases as Ti-Hetero underestimates the number of groups. The coverage probability, however, is not well-controlled, both of which are below the nominal coverage of 0.95. The Flat estimator also has good performance in terms of RMSE of ρ. Nonetheless, its coverage probability is relatively low: only 23% of credible sets successfully contain the true values. This is due to

[7]For the Tv-Homo and Tv-Hetero estimator, as we allow for time effects in α_i, we use the most recent α_{iT} to make one-step ahead prediciton. This is equivalent to assume the law of motion of α_{it} is a random walk. Modeling the trend of α_{it} would result in a more accurate forecast, but this is beyond the Scope of this paper.

[8]The Normal-Inverse-Gamma hyperprior for (μ, π^2) used in the Monte Carlo simulation is as follow: $\mu|\pi^2 \sim N(m, v\pi^2)$ with m equating to the pooled OLS estimator of α_i and $v = 1$; π^2 follows $IG(\nu_\pi/2, \delta_\pi/2)$ with $\nu_\pi = 6$ and $\delta_\pi = 4$.

the relatively large biases for α_i. The rest predictors are considerably worse. This implies that completely ignoring group structure (Pooled, Param) results in a significantly inferior estimate, so does wrongly modeling time-varying random effects (Tv-Homo, Tv-Hetero). Regarding the performance of clustering, Ti-Homo slightly underestimates the number of groups with an average K equals to 3.60 while the truth is 4.

For the case of DGP2, we keep time-invariant random effects while assuming heteroskedasticity. Tv-Homo and Tv-Hetero are still dominating. Tv-Hetero generates the best results with an accurate estimate of the number of groups as it correctly models heteroskedasticity which in turn improves the estimation efficiency. The Flat estimator closely follows them, and the rest are worse. Regarding DGP3, when time-varying random effects are introduced in the model, Tv-Homo and Tv-Hetero estimator yield the best performance and the estimated average Ks are close to the truth. But the biases are arguably low for these two estimators in sacrifice for small standard deviation and short credible intervals. It is worth noting that, unlike Ti-Homo in DGP1 and Ti-Hetero in DGP2, though correctly specified, the bias for α_i is still comparatively high. This is because, for simplicity, we don't model the law of motion for α_{it} and simply assume $\alpha_{iT+1} = \alpha_{iT}$, which results in large bias in α_i. As regards the DGP4 that doesn't have a group structure, the Flat estimator is the best since it doesn't pool cross-sectional information but estimate the unit-specific random effects. All of the BGRE estimators have almost identical performances with the estimated number of groups close or equal to 1. Since the Pooled and Param estimators assume no group structure, both have similar estimates as the BGRE estimators.

4.2.2 Point, Set and Density Forecast

Table 3 reports the predictive performance of a range of parametric forecasts. For the DGP1, the best forecasts are generated by the Ti-Homo estimator, as it is correctly specified in this environment. It has the smallest RMSFE, the shortest average length of the credible set, correct coverage probability, the largest LPS, and the smallest CRPS. Although allowing for heteroskedasticity along with the time-invariant random effects, Ti-Hetero generates a perfect point forecast as well as Ti-Homo. But Ti-Hetero introduces uncertainty revealed by a slightly wider credible set and worse density forecast. Moreover, estimators involving time-varying random effects (Tv-Homo and Tv-Hetero) worsen the forecast. Finally, incorrectly imposing no latent group pattern substantially deteriorates the predictive performance in all aspects.

Table 2: Monte Carlo Experiment: Point Estimates

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	Bias	Avg K
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.0198	0.0113	0.0120	0.0468	0.83	-0.0744	3.60
	Ti-Hetero	0.0202	0.0118	0.0120	0.0468	0.78	-0.0776	3.55
	Tv-Homo	0.2403	0.2387	0.0187	0.0712	0.06	-1.5255	1.82
	Tv-Hetero	0.2405	0.2389	0.0108	0.0689	0.07	-1.5301	1.89
	Pooled	0.2449	0.2447	0.0069	0.0268	0.00	-1.5588	1
	Flat	0.0369	-0.0344	0.0121	0.0469	0.23	0.2166	100
	Param	0.2711	0.2437	0.1148	0.4545	0.38	-1.5474	1
DGP 2 (Grp Ti He.)	Ti-Homo	0.0226	0.0097	0.0150	0.0583	0.85	-0.0681	3.74
	Ti-Hetero	0.0112	0.0036	0.0082	0.0321	0.95	-0.0261	3.98
	Tv-Homo	0.1924	0.1885	0.0234	0.0893	0.14	-1.2289	11.44
	Tv-Hetero	0.0965	0.0894	0.0255	0.0979	0.30	-0.5925	3.42
	Pooled	0.2318	0.2316	0.0079	0.0310	0.00	-1.4905	1
	Flat	0.0493	-0.0469	0.0146	0.0567	0.06	0.2984	100
	Param	0.2576	0.2303	0.1115	0.4407	0.38	-1.4792	1
DGP 3 (Grp Tv Ho.)	Ti-Homo	0.2726	0.2724	0.0119	0.0463	0.00	-2.1937	2.00
	Ti-Hetero	0.2741	0.2738	0.0121	0.0470	0.00	-2.2031	2.32
	Tv-Homo	0.0580	0.0525	0.0221	0.0860	0.33	-0.3679	3.89
	Tv-Hetero	0.0589	0.0534	0.0222	0.0863	0.33	-0.3743	3.85
	Pooled	0.1925	0.1923	0.0081	0.0314	0.00	-1.6381	1
	Flat	0.3230	0.3227	0.0126	0.0492	0.00	-2.5462	100
	Param	0.2172	0.1912	0.1021	0.4033	0.54	-1.6269	1
DGP 4 (Std Ti Ho.)	Ti-Homo	0.2177	0.2170	0.0164	0.0635	0.01	0.0038	1.03
	Ti-Hetero	0.2168	0.2159	0.0165	0.0644	0.01	0.0035	1.02
	Tv-Homo	0.2216	0.2210	0.0161	0.0627	0.00	0.0037	1.01
	Tv-Hetero	0.2204	0.2198	0.0164	0.0638	0.00	0.0038	1.16
	Pooled	0.2204	0.2198	0.0161	0.0628	0.00	0.0037	1
	Flat	0.1838	-0.1817	0.0277	0.1076	0.00	-0.0032	100
	Param	0.2321	0.2201	0.0714	0.2856	0.06	0.0154	1

For the DGP2, Ti-Hetero is expected to have an edge, and indeed, it dominates the remaining alternatives. Ti-Homo performs as great as Ti-Hetero in point forecast. This is because, from the previous section, we know that Ti-Homo generates accurate point estimates for both ρ and α_i. But Ti-Homo fails in the set forecast and density forecast, which illustrates the importance of modeling heteroskedasticity. Again, the rest estimators suffer apart from the Flat estimator.

In the DGP3, Ti-Homo and Ti-Hetero are doing badly by not capturing the time effects in α_{g_i}. Tv-Homo and Tv-Hetro are the best, beating the rest by a large margin, and equally accurate in this setup. The coverage probability for these two estimators is slightly lower than that of Ti-Homo and Ti-Hetero in part due to uncertainty introduced by more parameters of interest. Pooled, Flat, and Param estimator neglect both group structure and time-varying random effects, hence generating poor forecasts.

Regarding DGP4, all estimators beside Param have comparable forecasts as all of them deem no group structure in this environment (all estimated Ks are close to 1). However, Param suffers from high variance, and it always generates the widest credible interval and worse density forecast as in other DGP's.

4.3 Comparison with Variants of BGRE Estimator

This section conducts four sets of Monte Carlo simulation experiments aiming to examine the variants of BGRE estimator: (1) the GFE estimator proposed by Bonhomme and Manresa (2015), (2) a two-step GRE estimator with Kmeans, (3) the BGRE estimator with the subjective group prior, and (4) the BGRE estimator imposed with true K^0. The main text ignores part of the estimators we consider in the previous section and focuses on the correctly specified estimator for each DGP.

In addition to four DGPs specified in Section 4.1, we design three new DGPs that inherit the main features from DGP1, DGP2, and DGP 3, including balanced group structure and unit variance structure. However, instead of drawing α_{g_i} from the normal, we choose to use a constant α_k for each group. In this way, we could focus on the clustering result rather than repetitions to average out the randomness brought by random effects. We impose $K^0 = 4$ and use this number throughout the Gibbs sampling.

DGP5: Time-invariant grouped fixed effects, homoskedasticity.

$$y_{it} = \alpha_{g_i} + \rho y_{it-1} + \varepsilon_{it},$$

Table 3: Monte Carlo Experiment: Forecast

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.8117	-0.0073	0.8073	3.2036	0.95	-1.2134	0.4592
	Ti-Hetero	0.8122	-0.0064	0.8079	3.2268	0.95	-1.2158	0.4586
	Tv-Homo	0.8637	0.0171	0.8544	3.4461	0.95	-1.2754	0.4875
	Tv-Hetero	0.8638	0.0179	0.8544	3.4550	0.95	-1.2762	0.4877
	Pooled	0.9619	0.3974	0.8685	4.1093	0.97	-1.3905	0.5453
	Flat	0.8406	-0.0810	0.8326	3.2719	0.95	-1.2485	0.4753
	Param	0.9660	0.3994	0.8722	7.3886	1.00	-1.6369	0.6119
DGP 2 (Grp Ti He.)	Ti-Homo	1.0599	0.0123	1.0537	4.1027	0.93	-1.4746	0.5798
	Ti-Hetero	1.0428	0.0024	1.0365	3.7823	0.95	-1.2650	0.5416
	Tv-Homo	1.2662	0.0076	1.2551	3.7389	0.88	-1.6472	0.6700
	Tv-Hetero	1.0782	0.0025	1.0664	3.9051	0.95	-1.3074	0.5610
	Pooled	1.1839	0.3975	1.1072	4.8924	0.94	-1.5952	0.6516
	Flat	1.0768	-0.0814	1.0678	4.1956	0.93	-1.4962	0.5907
	Param	1.1883	0.3978	1.1119	7.8562	0.99	-1.7568	0.7096
DGP 3 (Grp Tv Ho.)	Ti-Homo	1.3740	0.3027	1.3360	5.2078	0.94	-1.7468	0.7850
	Ti-Hetero	1.3554	0.3067	1.3158	5.1512	0.95	-1.7230	0.7714
	Tv-Homo	1.0991	0.0127	1.0913	3.9240	0.93	-1.5167	0.6222
	Tv-Hetero	1.1062	0.0127	1.0985	3.9361	0.93	-1.5239	0.6262
	Pooled	1.8892	0.1200	1.8825	5.5759	0.85	-2.1509	1.1049
	Flat	1.2375	0.4145	1.1612	5.3199	0.97	-1.6427	0.7025
	Param	1.8927	0.1205	1.8860	8.2385	0.98	-2.0679	1.0809
DGP 4 (Std Ti Ho.)	Ti-Homo	0.8476	-0.0217	0.8429	3.3834	0.95	-1.2571	0.4785
	Ti-Hetero	0.8476	-0.0217	0.8429	3.3833	0.95	-1.2570	0.4784
	Tv-Homo	0.8517	0.0018	0.8426	3.3992	0.95	-1.2620	0.4811
	Tv-Hetero	0.8523	0.0016	0.8433	3.3987	0.95	-1.2627	0.4812
	Pooled	0.8472	-0.0218	0.8425	3.3850	0.95	-1.2567	0.4783
	Flat	0.8517	-0.0251	0.8470	3.2130	0.94	-1.2626	0.4821
	Param	0.8684	-0.0104	0.8426	8.5561	1.00	-1.6526	0.5961

with $\rho = 0.7$, $\alpha_k = k$ for $k = 1, 2, ..., K^0$ and $\varepsilon_{it} \overset{iid}{\sim} N(0, 1)$.

DGP6: Time-invariant grouped fixed effects, heteroskedasticity.

$$y_{it} = \alpha_{g_i} + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\alpha_k = k$ for $k = 1, 2, ..., K^0$ and $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma^2_{g_i}\right)$ with $\sigma^2_k = 1.5\left(1 - \frac{k-1}{K^0}\right)^2$.

DGP7: Time-varying grouped fixed effects, homoskedasticity.

$$y_{it} = \alpha_{g_i t} + \rho y_{it-1} + \varepsilon_{it},$$

with $\rho = 0.7$, $\varepsilon_{it} \overset{iid}{\sim} N(0, 1)$ and $\alpha_{g_i t} = \underline{\alpha}_{g_i t}$ where $\underline{\alpha}_{g_i t}$ varies across periods and groups as depicted in Figure 1.

4.3.1 GFE Estimator

In this experiment, we compare our BGRE estimator with the GFE estimator proposed by Bonhomme and Manresa (2015). In particular, we assess the performance of point forecast in y and the accuracy of coefficient estimates (group random effects α and common parameter ρ).

To compare the performance of estimators, we use the default numerical setting[9] in Bonhomme and Manresa (2015). It is worth noticing that Bonhomme and Manresa (2015) relies on information criteria to ex post select the optimal number of groups. Hence we consider at most 10 groups and estimate the number of groups K in accordance with the following Akaike information criterion (AIC)[10]:

$$AIC(k) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \hat{\rho}^{(k)} y_{it-1} - \hat{\beta}_i'^{(k)} x_{it-1} - \widehat{\alpha}^{(k)}_{\widehat{g}_i t}\right)^2 + 2\widehat{\sigma}^{2(k)}\frac{k(T+N-k)}{NT},$$

where $\widehat{\sigma}^2$ is a consistent estimate of the variance of ε_{it}:

$$\widehat{\sigma}^{2(k)} = \frac{1}{NT - K_{\max}T - N - (p+1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \hat{\rho}^{(k)} y_{it-1} - \hat{\beta}_i'^{(k)} x_{it-1} - \widehat{\alpha}^{(k)}_{\widehat{g}_i t}\right)^2.$$

[9] The default settings are as follow: (1) Number of groups = 4; Number of covariates = 1; Standard errors: 0 (no standard errors). (2) For algorithm 0, Number of simulations = 100; (3) For algorithm 1, Number of simulations = 10, Number of neighbors = 10 , Number of steps = 10.

[10] We also tried the alternative choice $\hat{\sigma}^2 \frac{kT+N+p+1}{NT} \ln(NT)$ for the penalty. This corresponds to the default BIC used in Bonhomme and Manresa (2015). We found that, in this case, BIC selected the smallest possible number of groups for all DGPs, i.e., no group structure, whereas the truth is $K^0 = 4$. Moreover, other forms of BIC could always select the largest K as well. Due to the inaccurate estimate of group structure and substantially poor performance, we don't show the result with the default BIC.

The results are shown in the Table 4. As Bonhomme and Manresa (2015) proposes two algorithms[11], we present the results for four versions of the GFE estimator. The first two estimators equip with the Iterative and Variable Neighborhood Search algorithm, respectively. We impose the true number of groups K^0 for the other two estimators, i.e., we don't perform model selection but choose $\hat{K} = K^0 = 4$ directly.

For the DGP5 and DGP6, Ti-Homo and Ti-Hetero estimator outperform the GFE estimators in all aspects. The GFE estimator's poor performance is mainly due to the incorrect estimate of the number of groups. This also emphasizes that an inaccurate estimate of group structure would deteriorate both estimates and point forecasts. It is worthy noting that, even imposing the true number of groups, GFE estimators (GFE^0_{a0} and GFE^0_{a1}) are still straggling and worse than their counterparts with $\hat{K}$ selected by the information criterion. GFE^0_{a0} and GFE^0_{a1} generate relatively high bias for both α_i and ρ. In the case of DGP7, Tv-Homo and Tv-Hetero perform only marginally worse than GFE^0_{a0} and GFE^0_{a1} estimators imposed with the true K^0. This is because they overestimate the number of groups in some posterior draws, and hence, on average, the posterior mean forecast and estimate are slightly off.

Moreover, the accuracy of the GFE estimator is profoundly affected by the choice of the information criteria. We implement several information criteria proposed by Bai and Ng (2002) in this Monte Carlo experiment and find that there is no single criterion that consistently selects the correct number of groups nor close to the truth. As the GFE estimator is designed for the time-varying model, we finally select the AIC mentioned above, which chooses the true model in time-varying DGPs This deliberately selected AIC is another reason for the great performance of the GFE estimator in DGP7. Once we switch to other forms of AIC or BIC, these superior results don't hold. These facts also emphasize the importance of not relying on ex-post model selection and the superiority of our Bayesian estimators.

4.3.2 GRE Estimator with Subjective Group Prior

In this section, we explore whether subjective group structure improves the accuracy of forecast and group clustering. We conduct two Monte Carlo experiments corresponding to SGM Prior and SGP Prior defined in Section 3.1.2.

SGM Prior: Given the true group structure, we assume a researcher split *all* the units into $\widetilde{K}$ groups with 60% of them being correctly classified. Notice that $\widetilde{K}$ can be either

[11] These two algorithms could generate different estimates as shown in the case of DGP6 below.

Table 4: Monte Carlo Experiment: BGRE vs GFE

		Point Forecast			$\hat{\rho}$	$\hat{\alpha}$	Group
		RMSFE	Error	Std	Bias	Error	Avg K
DGP 5 (Grp Ti Ho.)	Ti-Homo	0.7806	0.0261	0.7801	0.0214	-0.1359	4.63
	Ti-Hetero	0.7829	0.0254	0.7825	0.0212	-0.1353	4.81
	Tv-Homo	0.8062	0.0904	0.8011	0.3261	-2.1184	1
	Tv-Hetero	0.7995	0.0883	0.7946	0.3262	-2.1183	1.52
	GFE_{a1}	0.7882	0.0845	0.7837	0.2891	-1.8767	2
	GFE^0_{a1}	0.8421	0.0496	0.8406	0.0463	-0.2935	4
	GFE_{a0}	0.7882	0.0845	0.7837	0.2891	-1.8767	2
	GFE^0_{a0}	0.8421	0.0496	0.8406	0.0463	-0.2935	4
	Flat	0.7881	-0.0487	0.7866	-0.0258	0.1752	1
DGP 6 (Grp Ti He.)	Ti-Homo	1.1399	0.2578	1.1104	0.0100	-0.0475	4.59
	Ti-Hetero	1.0971	0.2525	1.0677	0.0060	-0.0189	4.38
	Tv-Homo	1.2850	0.3136	1.2462	0.2692	-1.7265	9.62
	Tv-Hetero	1.1356	0.2989	1.0955	0.0970	-0.6102	3.53
	GFE_{a1}	1.1786	0.3098	1.1371	0.1580	-1.0065	2
	GFE^0_{a1}	1.3201	0.3119	1.2827	0.1792	-1.1441	4
	GFE_{a0}	1.1786	0.3098	1.1371	0.1580	-1.0065	2
	GFE^0_{a0}	1.2408	0.3117	1.2010	0.1774	-1.1324	4
	Flat	1.1209	0.1540	1.1103	-0.0489	0.3402	1
DGP 7 (Grp Tv Ho.)	Ti-Homo	1.3620	0.4483	1.2862	0.2668	-2.0886	2
	Ti-Hetero	1.3625	0.4477	1.2869	0.2668	-2.0884	2
	Tv-Homo	0.9908	0.1201	0.9835	0.0453	-0.2998	4.02
	Tv-Hetero	0.9953	0.1155	0.9886	0.0519	-0.3478	4.02
	GFE_{a1}	0.9728	0.1146	0.9661	-0.0071	0.0673	4
	GFE^0_{a1}	0.9728	0.1146	0.9661	-0.0071	0.0673	4
	GFE_{a0}	0.9728	0.1146	0.9661	-0.0071	0.0673	4
	GFE^0_{a0}	0.9728	0.1146	0.9661	-0.0071	0.0673	4
	Flat	1.2750	0.5588	1.1460	0.3183	-2.4505	1

Note: The meanings of notation in the second column are as follow: GFE = Bonhomme and Manresa (2015)'s estimator; GFE^0 = GFE estimator with the true K_0; a1 = algorithm 1: Variable Neighborhood Search; a0 = algorithm 0: Iterative Search.

larger or smaller than the true K^0 and in the following experiment, we set $\widetilde{K} = 7 > K^0 = 4$. The remaining 40% of units are randomly assigned to the wrong groups.

SGP Prior: we consider five scenarios, each of which differs in the structure of subjective prior probability and hence the prior group probability π. The exact specification is characterized in Table 5. The first three scenarios set the preset number of groups as the truth K^0, whereas subjective group probabilities are different in levels. In the scenario 1, the researcher is pretty confident about her decision and assigns 100% to the right group for each unit, which amounts to knowing the true membership. However, she is entirely uninformed and cluster units with even probability for each group (i.e., $\omega_{ik} = 1/K^0$ for $\forall i, k$) in the scenario 3. Scenario 2 is an intermediate case where one is less confident in her knowledge and correctly assigns a unit to its group with the prior probability of 70% (other groups equally split the remaining 30%). For the scenario 4 and 5, the number of groups is different from the truth. We assume the researcher divides all units into $K \neq K^0$ even groups with the prior probability of 100%[12].

Table 5: Simulation Design: Subjective Group Probability

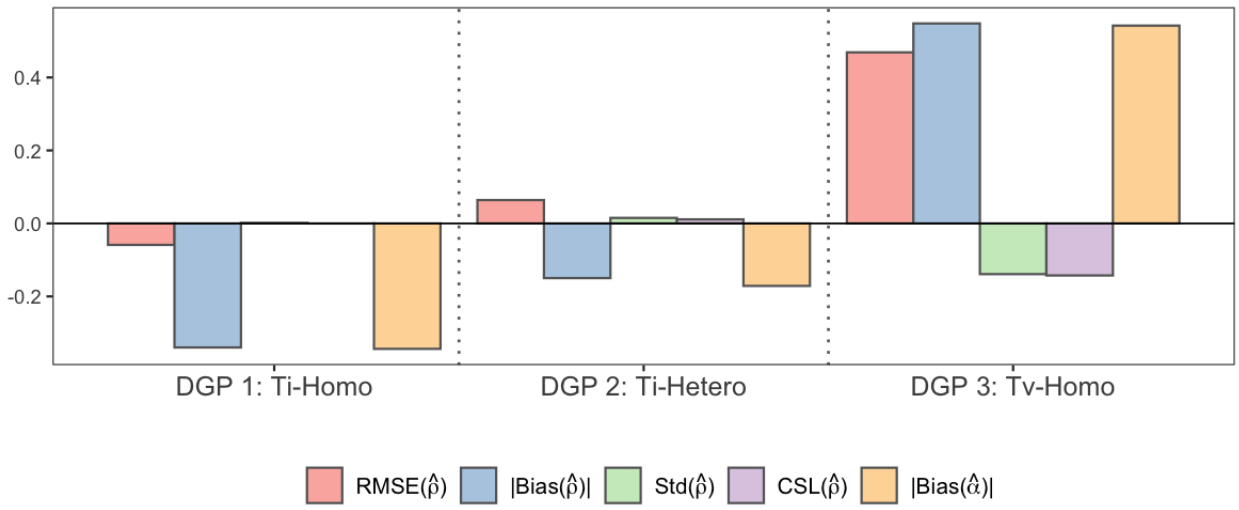
Scenarios	# of Groups	Descriptions
1	K^0	very confident, assign 100 % to the correct group
2	K^0	less confident, assign 70 % to the correct group
3	K^0	uninformed, evenly assign $1/K^0 \times 100$ % to the each group
4	$K^0 - 1$	very confident, assign 100 % to a particular group (might be wrong)
5	$K^0 + 1$	very confident, assign 100 % to a particular group (might be wrong)

The simulation results[13] for the SGM prior are depicted in the Figure 2 and 3. Each bar represents the relative performance of the GRE estimator using SGM prior against the performance of the benchmark BGRE estimator. Above zero indicates the SGM prior is less favorable while GRE estimator using SGM prior is better than the benchmark when its bars show negative values. Moreover, the main models are the correctly specified for each DGP,

[12]The last two scenarios aim to evaluate the performance of SGP prior when the number of groups is wrong. Instead of randomly assigning a unit to each groups with a set of probability, we assume the econometrician is confident on her prior and set 100% for a particular group. In particular, we assign the first N/K units into group 1, the next N/K units into group 2, and so on. We also run other designs for scenario 4 and 5 with different prior probabilities. The results show that, as long as a certain amount of units are correctly clustered into groups, the performances of SPG-RE estimators are slightly better than those of the BGRE estimator.

[13]The full results are presented in the Appendix E.5.

Figure 2: Monte Carlo Experiment: Estimates, SGM Prior

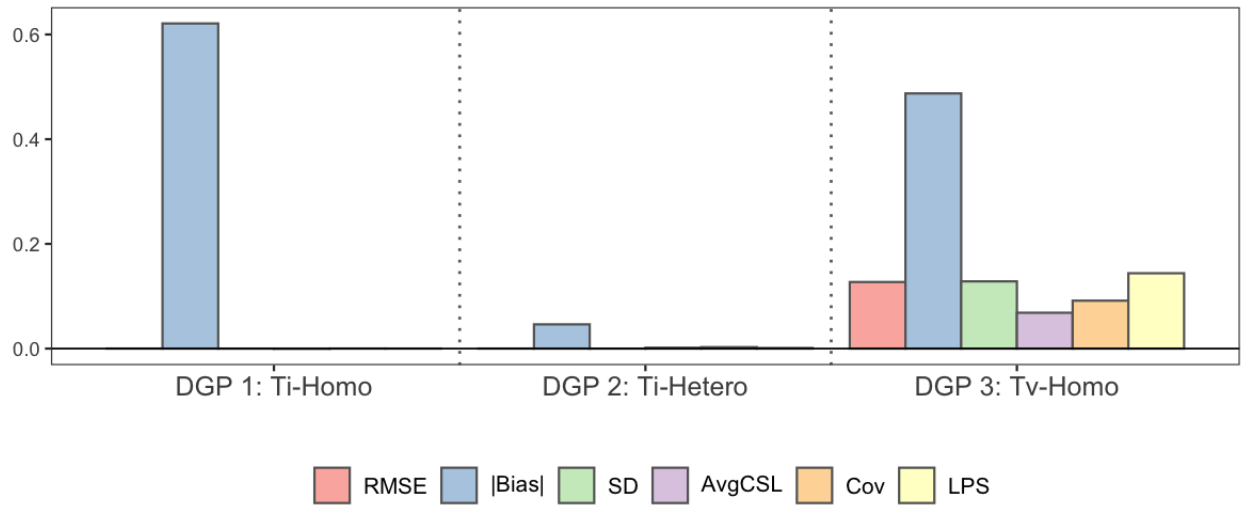


i.e., Ti-Homo for DGP 1, Ti-Hetero for DGP 2, and Tv-Homo for DGP 3, since they have the best performance in terms of both estimates and forecast.

Regarding the DGP1 and DGP2, from the perspective of the estimate, SGM prior decreases the bias for ρ and α_i. This is because SGM prior reveals partial true group structure through the prior of α_i. However, it has little effect on RMSFE of one-step ahead forecast and the others. These interesting results suggest that if we focus on the one-step forecast only, even incorporating partially correct group structure in the prior, it is still hard to outperform the BGRE estimator significantly. The spike of forecast error for the point forecast in Figure 3 is not as bad as it looks like. Rather, both BGRE estimators with and without SGM prior produces relatively great forecasts in terms of forecast error. But, the errors are so close to 0 that any small increase in the error makes the bar surge rapidly. In the DGP3, SGM prior fails to improve the estimate and forecast as most bars lie above zero. This is in part because the OLS estimate of the prior mean for time-varying α_i introduces a great amount of bias and hence deteriorates the performance. This also suggests that if there are time effects in α_i, SGM prior is not the optimal choice. Instead, SGP prior addresses this issue successfully, as illustrated in the next experiment below.

We also conduct the simulation experiment for the SGP Prior, where we examine the impact of subjective group probability prior on the performance of estimate and forecast under DGP3 (time-varying random effects model). To adopt the SGP prior, we revise the

Figure 3: Monte Carlo Experiment: Forecast, SGM Prior

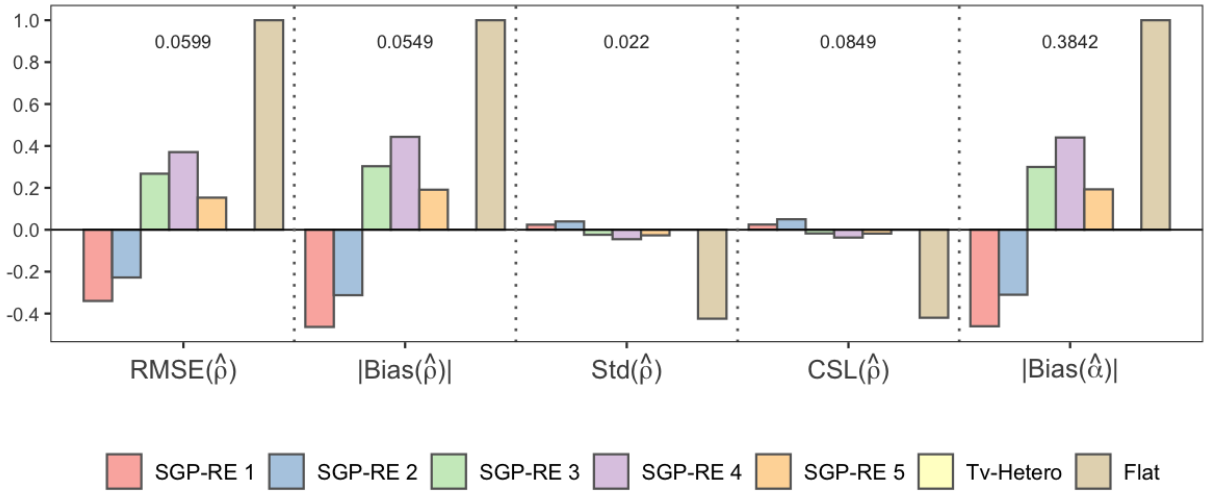


BGRE estimator and construct the new Bayesian estimator under the assumptions: (1) time-varying random effects and (2) heteroskedasticity, which corresponds to the most general case of a panel data model. We name it as Subjective Group Probability Random Effect (SGP-RE) estimator. Given different specifications of SGP prior, we report the performance of five SGP-RE estimators relative to the Tv-Hetero estimator (benchmark model).

Figure 4 and 5 depict the relative performance of various SGP-RE estimators against the Tv-Hetero estimator[14]. Remember that the prior knowledge is most accurate in scenario 1, where the researcher is equivalent to know the true group structure. In this regard, a clear gain in estimate emerges as the RMSE for ρ, bias for ρ and α_i generated by SGP-RE1 (100% confidence) decreases by more than 35% relative to the Tv-Hetero estimator. As we move from scenario 1 to the other scenarios, the prior information becomes less accurate. SGP-RE2 (70% confidence) beats the benchmark with moderate improvement on the RMSE ($\sim$22%) and the bias ($\sim$32%) while SGP-RE3 (uninformed) underperforms the benchmark by a huge margin. The poor performance of SGP-RE3 is not surprising. Though it correctly specifies $K = K^0$, the uninformative prior forces the algorithm to consider other incorrect groups with a large chance ($= 1 - \frac{1}{K^0}$), and hence deteriorates the performance of both estimates and forecasts. In terms of the one-step ahead forecast, SGP-RE1 leads the rest by scoring the lowest values for each metrics (and the highest LPS), closely followed by SGP-RE2. SGP-RE3 is suffering in terms of point and set forecast. As for scenario 4 and 5,

[14]The full results are presented in the Appendix E.5.

Figure 4: Monte Carlo Experiment: Estimates, SGP Prior



Note: the value of each relative metric is capped by 1 to enhance the readability. The number in each subpanel represents the (original) value of metric for the benchmark model (Tv-Hetero).

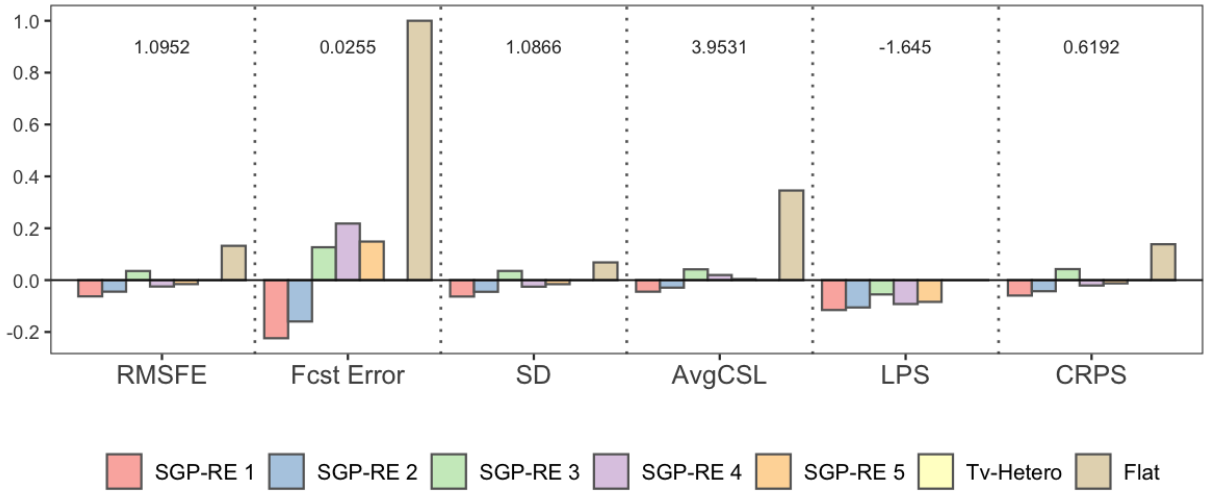
despite the flexibility of allowing for changing a and K^p along MCMC sampling, the incorrect specifications of the group number deteriorate the estimation, both of which fail to deliver reliable estimates for ρ and α. Nonetheless, such prior structures help point and density forecasts. Both estimators beat the benchmark and generate comparable LPS to SGP-RE1 and SGP-RE2. This valuable improvement mainly results from the fact the algorithm could exploit the prior knowledge on group structure that successfully partitions merely a fraction of units and adapt the number of groups accordingly.

In short, regarding the overall performance regarding SGP prior, the best case is that the researcher has a relative confident prior and knows the true number of groups. In this case, the SGP-RE estimator dominates the Tv-Hetero estimator from every angle. However, in practice, we rarely come up with such a precise prior due to the incomplete understanding of the population of the data. Instead, we might be less confident on our knowledge or even specify more/fewer groups than the truth. Under this circumstance, the SGP-RE estimator could still deliver a better density forecast because of the great exploration of the prior information and the adaptive scheme featured by our Bayesian method.

4.3.3 Two-Step GRE Estimator

In this section, we compare our BGRE estimator with a two-step GRE estimator, where units are clustered into groups in the first step using *Kmeans* algorithm, and the model

Figure 5: Monte Carlo Experiment: Forecast, SGP Prior



Note: the value of each relative metric is capped by 1 to enhance the readability. The number in each subpanel represents the (original) value of metric for the benchmark model (Tv-Hetero).

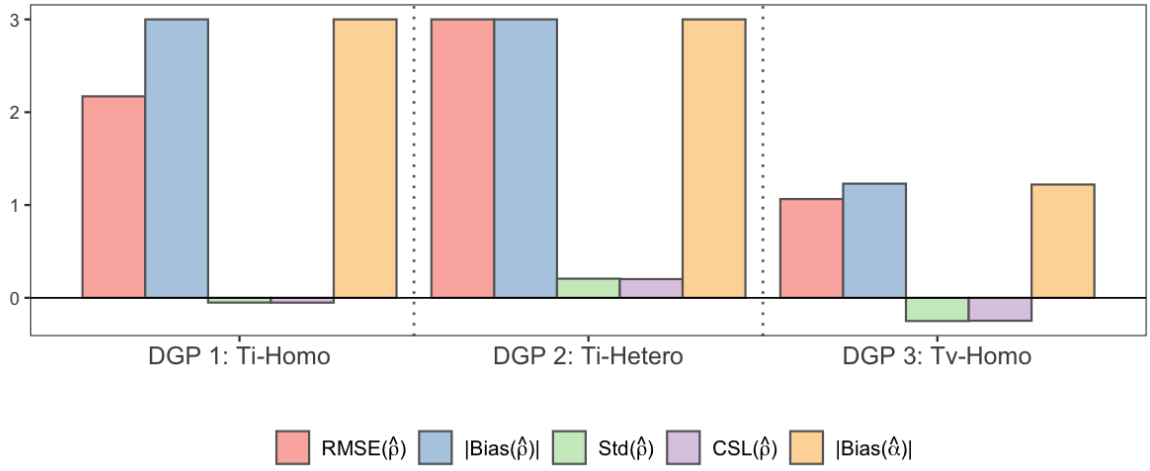
is then estimated in the second step with group-specific heterogeneity. Unlike Bonhomme, Lamadon, and Manresa (2019), we implement the Bayesian framework in the second step to echo other Bayesian estimators presented in the previous section. This two-step procedure allows us to examine the clustering accuracy of Kmeans relative to our full Bayesian estimate as two-step GRE estimators can be viewed as a special case of the GRE estimator with group membership determined by Kmeans. The optimal number of clusters in Kmeans is selected by the average silhouette method.

To avoid cluttering the tables in the main text, we depict the selected results[15] for estimates and forecast in Figure 6 and 7. Each bar represents the performance of the two-Step GRE estimators against the performance of the original GRE estimator. Above zero indicates the 2-step estimator underperforms the benchmark while a 2-step estimator is better when its bars show negative values. The main models are correctly specified for each DGP, i.e., Ti-Homo for DGP 1, Ti-Hetero for DGP 2, and Tv-Homo for DGP 3.

Figure 6 presents the point estimates for each DGP. We document the root of mean squared forecast error, absolute bias, standard deviation, and the average length of the 95% credible set for the common parameter ρ, while the metric for the random effects is the absolute bias. According to these measures, the two-Step GRE estimators perform worse than the Bayesian GRE estimator as they introduce much higher bias in the estimate of ρ (hence larger RMSE for ρ) and α_i. It is worth noting that the estimator equipped with

[15]The full results are presented in the Appendix E.4.

Figure 6: Monte Carlo Experiment: Point Estimates, Two-Step GRE with Kmeans



Note: the value for each relative metric is capped by 3 to enhance the readability.

Kmeans doesn't affect the standard deviation and the average length of 95% credible set of ρ.

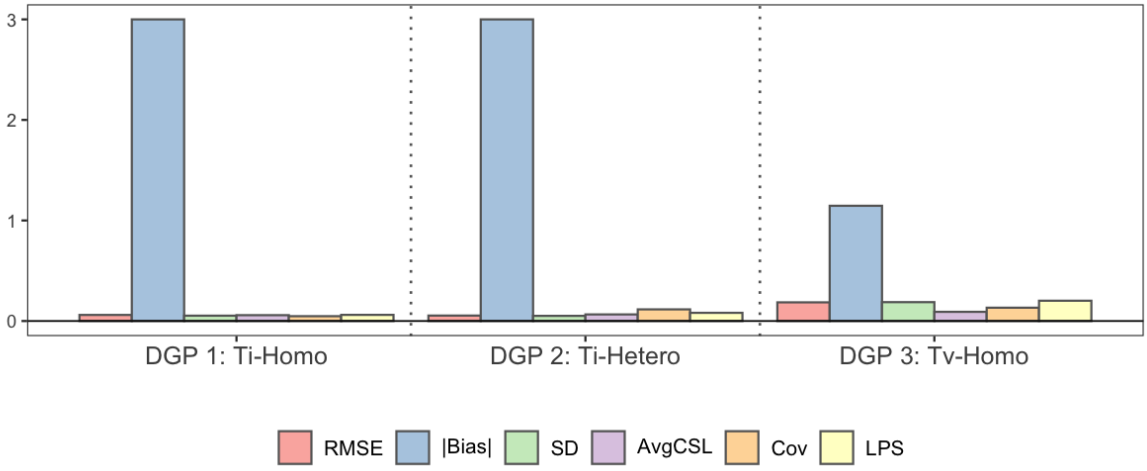
The inferior performance of the 2-step GRE estimator is due to the inaccurate estimate of group structure. Table 6 reports the estimated number of groups from the two-step GRE estimators with Kmeans and the BGRE estimator. Regarding the performance of clustering, the Kmeans algorithm severely underestimates the number of groups as it prefers much less groups, while the true number is 4. Meanwhile, our BGRE approach accurately estimates the number of groups, though slightly underestimated in the DGP 1.

Table 6: Number of groups, Kmeans vs. BGRE

	DGP 1	DGP 2	DGP 3
Kmeans	2.20	2.20	2.26
BGRE	3.60	3.98	3.85

Figure 7 shows the point, set, and density forecast for each DGP. As Kmeans fails to estimate the group structure, none of the 2-step GRE estimators outperform the GRE estimators. Namely, Kmeans clustering doesn't help make a more accurate forecast, and instead, it generates a much higher forecast bias and standard deviation.

Figure 7: Monte Carlo Experiment: Forecast, Two-Step GRE with Kmeans



Note: the value for each relative metric is capped by 3 to enhance the readability.

4.3.4 Fixed K Estimator: Imposing the true number of groups

As shown in the previous subsection, the Bayesian GRE estimator works reasonably well in finite samples to determine the number of groups. In this subsection, we assume that the number of groups is known and focus on clustering. We present a table of the accuracy of clustering, where each row shows the faction of units that are correctly assigned to the true group. As an orthodox clustering algorithm, the results for Kmeans are also included as the benchmark. To avoid cluttering the tables in the main text, we don't present the results for suboptimal estimators. To be more precise, for the DGP involving time-invariant random effects, we only document the result for Ti-Homo and Ti-Hetero since other estimators are arguably worse in clustering, based on the simulation presented in the previous section. The same rule applies for time-varying DGPs.

Table 7 shows the accuracy of clustering for each estimator. Overall, the accuracy is high for Kmeans and correctly specified estimators in each DGP, while our BGRE estimators are slightly dominated by the Kmeans algorithm. The reasons are straightforward. Our BGRE estimators simultaneously estimate parameters and group units while Kmeans merely performs clustering. The additional estimation steps in our block Gibbs sampler depend on priors and parametric assumptions that could affect the clustering. On the other hand, the Kmeans algorithm forms clusters through spatial relationships between units, free of any assumption. Such differences yield the discrepancies in accuracy between Kmeans and BGRE estimators. But they are acceptable as the discrepancies are within 10% most of the time. Comparing the performance of Kmeans in the two-step GRE estimator (Table 6),

imposing the correct number of groups indeed improves the clustering ability of Kmeans. Nevertheless, it is uncommon to know the truth in practice.

Table 7: Monte Carlo Experiment: Accuracy of Clustering, Fixed K^0

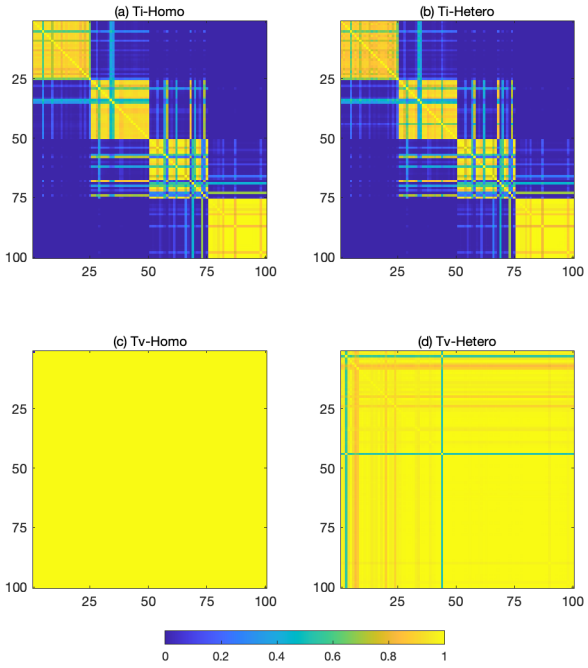
		Group 1	**Group 2**	**Group 3**	**Group 4**
DGP 5 (Grp Ti Ho.)	Ti-Homo	86.87%	83.62%	60.42%	91.64%
	Ti-Hetero	73.59%	66.17%	56.47%	95.82%
	Kmeans	88.00%	96.00%	68.00%	100.00%
DGP 6 (Grp Ti He.)	Ti-Homo	78.17%	78.66%	66.58%	99.14%
	Ti-Hetero	89.63%	84.22%	85.18%	99.98%
	Kmeans	84.00%	100.00%	88.00%	100.00%
DGP 7 (Grp Tv Ho.)	Tv-Homo	99.33%	68.99%	93.40%	84.61%
	Tv-Hetero	98.92%	71.53%	93.41%	75.87%
	Kmeans	96.00%	88.00%	100.00%	88.00%

Next, we visualize the clusters to provide a clear view of the performance of clustering. We construct a posterior similarity matrix, a matrix containing the posterior probabilities of observations i and j being in the same cluster (estimated empirically from the MCMC draws). This design avoids the problem of reassigning group members to give posterior draw and show a clear group structure.

Figure 8, 9 and 10 present the similarity matrices for the simulation using DGP5, DGP6 and DGP7, respectively. The colors depict the degree of similarity. Ideally, a perfect estimator should reveal four light yellow squares in the heatmap, leaving the remaining area in dark blue. As DGP5 implements fixed-effects and assumes homoskedasticity, Ti-Homo and Ti-Hetero estimator reveal a clear partition that matches the design of DGP5. Though a few units are incorrectly clustered, four yellow squares on the diagonal indicate that the posterior partition is reliable. However, Tv-Homo and Tv-Hetero estimators deliver inferior estimates and present one major group instead of four.

Turning to DGP6, the best partition is generated by the Ti-Hetero estimator, which is correctly specified under this DGP. Even though the data density of group 2 heavily overlaps with the one of group 1 and group 3, due to the relatively small mean and large variance in α_i, Ti-Hetero estimator succeeds in delivering a clear group pattern that clearly distinguishes these three groups. The Ti-Homo estimator also has an excellent performance with ignorance of heteroskedasticity, but it generates much more vague boundaries between groups 1, 2, and

Figure 8: Heatmap for Similarity Matrix, DGP5, fix K^0



3. The Tv-Homo and Tv-Hetero results are incredibly messed, none of which depicts the correct partition.

As for DGP7, Tv-Homo and Tv-Hetero are the best, which is expected in this DGP. We see a clear four-group pattern from the similarity matrix in panel (c) and (d). A few yellow and light blue stripes in the off-diagonal block suggest Tv-Homo and Tv-Hetero estimators wrongly allocate a few units in posterior draws, especially for the units in group 2 and 4. Indeed, the paths of random effects in these two groups share great similarities. As depicted in Figure 1, the red line (group 2) can be roughly viewed as the step function approximation of the green line (group 4). Ti-Home and Ti-Hetero struggle as they ignore the time effect in α_i by construction.

Figure 9: Heatmap for Similarity Matrix, DGP6, fix K^0

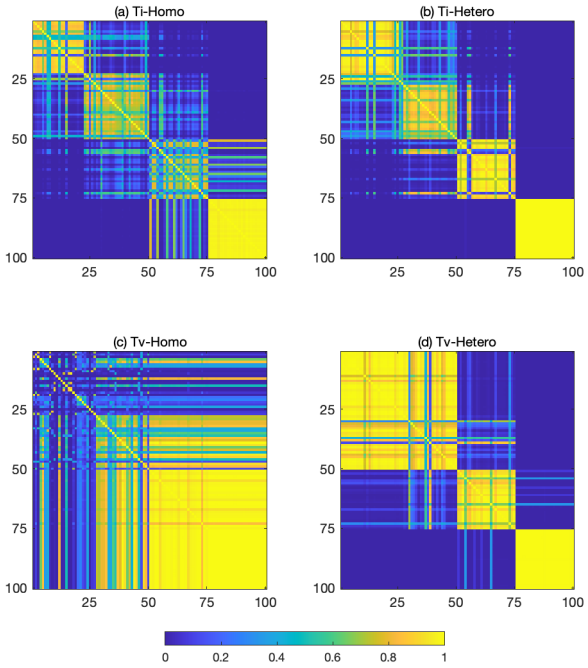
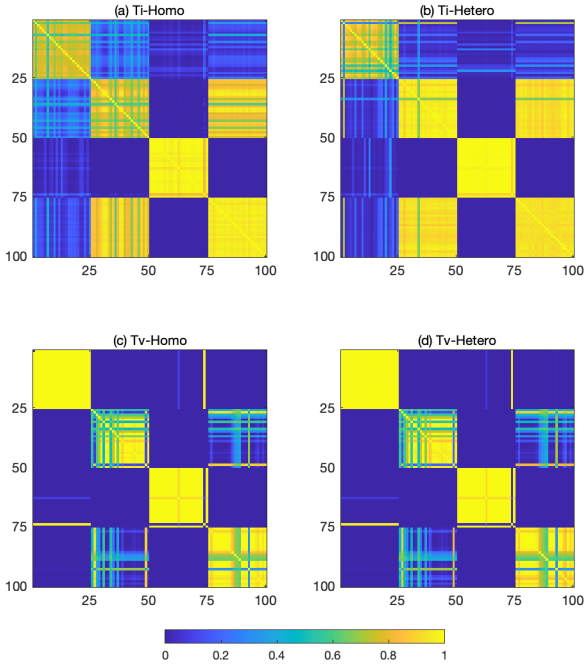


Figure 10: Heatmap for Similarity Matrix, DGP7, fix K^0



5 Empirical Application

In this section, we illustrate the use of Bayesian Grouped Random Effects estimators in a cross-firm study. We revisit the investment regression and use a different version of the dynamic grouped panel model to forecast the investment rate[16] for a panel of firms in all industries. Instead of using the traditional Tobin's Q-type investment regression, we implement a new scheme proposed by Gala, Gomes, and Liu (2019), who directly estimates the corporate investment rate without Tobin's Q. Again, our main focus is the one-step ahead point, set and density forecast. Due to space limitations, we only report forecast results for the most recent year in the main text. Summary statistics, and additional details of implementation are stored in Appendix F.

5.1 Model Specification

We consider a general model with grouped latent heterogeneity in α_i. Following Hsiao and Tahmiscioglu (1997) and Gala, Gomes, and Liu (2019), the investment equations are specified as,

$$\left(\frac{I}{K}\right)_{it} = \alpha_{g_i t} + \rho \left(\frac{I}{K}\right)_{it-1} + \beta_{i1} \left(\frac{CF}{K}\right)_{it-1} + \beta_{i2} \ln K_{it-1} + \beta_{i3} \ln \left(\frac{Y}{K}\right)_{it-1} + \varepsilon_{it}, \quad (5.1)$$

where capital stock, K_{it} is defined as net property, plant and equipment; I_{it} is capital investment; CF_{it}, is a liquidity variable defined as cash flow minus dividends; Y_t is the end-of-year sales; ε_{it} are the normally distributed error terms. The subscript i denotes companies, and t denotes time. Unlike the commonly specified investment equation using Tobin's Q, the additional terms, including the natural logarithm of lagged capital and sales-to-capital ratio, are based on the regression proposed by Gala, Gomes, and Liu (2019). The lagged values of the investment rate are included as explanatory variables to avoid endogeneity problems.

As we focus on forecasting, we can relax a few assumptions to achieve better predictive performance. These assumptions include time-invariant random effects α_{g_i}, homoskedasticity in σ_i and homogeneous coefficients for all dependent variables ($\beta_{i\cdot} = \beta_i$). Table 8 summarizes the estimators and their properties we consider in this section. The implementation of time-invariant RE and homoskedasticity is similar to the one in the previous section, i.e., construct four versions of BGRE estimator: Ti-Homo, Ti-Hetero, Tv-Homo, and Tv-Hetero.

[16] The investment rate for a firm in a particular year is defined as the fraction of capital expenditures in property, plant, and equipment in terms of the beginning-of-year capital stock.

Despite the fact that the homogeneous slopes have been frequently rejected in empirical studies of estimates and inference, such an assumption could provide potential improvement in forecasts.

Table 8: Summary of Estimators in the Empirical Analysis

		Time-nvariant α_i	Homogeneity	Group Structure
Homogenous Coef.	Ti-Homo	**X**	**X**	**X**
	Ti-Hetero	**X**		**X**
	Tv-Homo		**X**	**X**
	Tv-Hetero			**X**
	Flat	**X**	**X**	
	Pooled	**X**	**X**	
	Param	**X**	**X**	
Heterogenous Coef.	Ti-Homo	**X**	**X**	**X**
	Ti-Hetero	**X**		**X**
	Tv-Homo		**X**	**X**
	Tv-Hetero			**X**
	Flat	**X**	**X**	

5.2 Data

The individual company data are obtain from COMPUSTAT Annual database. To account for potential structural breaks and the advanced speed of capital accumulation in the recent decades, our sample is composed of a balanced panel of firms for the years 2000 to 2019, that includes firms from all industries with no missing value in accounting data.

We keep only firm-years that have non-missing information required to construct the primary variables of interest, namely: capital stock K, investment I, and sales revenues Y. The further details of constructing the sample can be found in the Appendix F. The final sample comprises 337 firms and the observations for each firm is 20.

To examine the performance of various estimators with limited observations, we choose to use a rolling window of 15 years. In this sense, we create five balanced panels which end in years 2014, ..., 2018 ($t = T$), respectively. The observations in the next year ($t = T + 1$) are reserved for pseudo-out-of-sample forecast evaluation. For illustrative purposes, we will present the results for the year 2019 in the remainder of this section. The full results are presents in Appendix F.

Table 9: Empirical Application: Forecast Performance

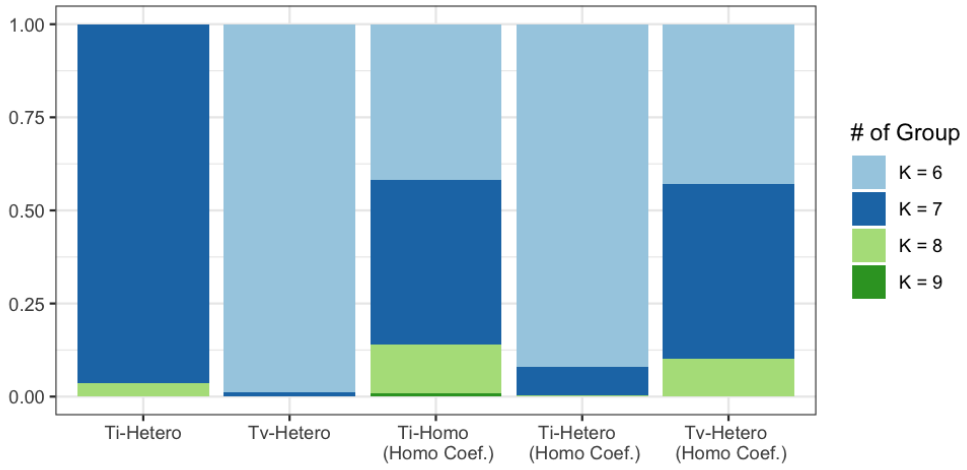
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Point		Set		Density	
		RMSFE	Avg K	Cov	Length	LPS	CRPS
Homogenous Coef.	Ti-Homo	0.1108	2	0.9614	0.5908	0.7039	0.0605
	Ti-Hetero	0.0822	7.86	0.9021	0.4012	1.3724	0.0464
	Tv-Homo	0.1177	1	0.9525	0.5875	0.6671	0.0634
	Tv-Hetero	0.0812	6.75	0.8813	0.3867	1.2981	0.0454
	Pooled	0.1150	1	0.9555	0.5966	0.6746	0.0627
	Flat	0.1100	1	0.9703	0.6041	0.6935	0.0604
	Param	0.2043	1	1.0000	6.8211	-1.2722	0.3554
Heterogenous Coef.	Ti-Homo	0.1144	7.48	0.8724	0.2837	1.1904	0.0485
	Ti-Hetero	0.1152	6.68	0.8724	0.2841	1.1883	0.0485
	Tv-Homo	0.1070	1	0.9703	0.6344	0.6879	0.0615
	Tv-Hetero	0.1101	7.46	0.8338	0.2741	1.1013	0.0486
	Flat	0.1164	1	0.9733	0.6753	0.6205	0.0650

5.3 Results

We begin the empirical analysis by comparing the performance of point, set, and density forecast for the last panel (in-sample periord: 2005 - 2018). We aim to forecast the investment rate in 2019. We consider all the model specifications depicted in the Table 8 and their performance is presented in the Table 9. Throughout the analysis, the Flat estimator serves as the benchmark as it essentially assumes individual effects. In Table 9, the third column shows the RMSFE for the one-step ahead forecast. For the panel considered in the table, we first notice that the best model is the Tv-Hetero (time-varying random effects, heteroskedasticity) in homogeneous coefficients specification. It outperforms the benchmark – Flat estimator by 25%. Ti-Hetero also delivers accurate point forecast, which suggests time effects provide merely marginal improvement. Under heterogeneous coefficients specification, for the BGRE estimators, though all of them beat the Flat estimator, their RMSFEs are relatively larger. This may arise from the fact that heteroskedasticity alone can capture a great amount of individual effects, thus imposing heterogeneous coefficients in β_i may overfit the model and lead to poor forecasts.

The fourth column documents the average number of latent groups in α_i. Most of our BGRE estimators deem a group structure with more than six underlying components. And

Figure 11: Empirical Application: Distributions of Group Number

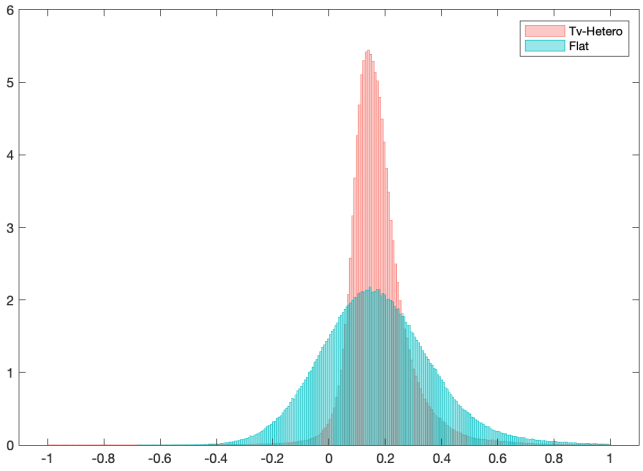


as we will discuss later, this rich group structure is the cornerstone of the accurate and flexible density forecast. In Figure 11, we present the posterior distribution of the number of groups for those BGRE estimators that have more than two groups. Regardless of the predictive performance, most estimators agree on the number of groups, with the posterior mode ranging from 6 to 7. On the other hand, the SIC code, which categorizes companies into the industries by their business activities, suggests that there are ten different industries in our sample. This indicates that our block Gibbs sampling algorithm reshuffles the default group structure and optimally pools firms from several sectors.

The fifth and sixth columns present the average coverage rate with the nominal coverage probability of 95% and the average length of the 95% credible set. In general, the coverage rates generated by the homogeneous coefficients specifications are substantially larger than the sets obtained from the models with heterogeneous coefficients and are closer to the nominal coverage probability of 95%. However, the homogeneous setting doesn't improve the average length of the credible set. Indeed, the decrease in the average length is evident for the heterogeneous coefficient models, and it becomes even more pronounced once we impose heteroskedasticity. This is because the sizeable cross-sectional variation in the posterior predictive distributions leaves plenty of room for heterogeneous and heteroskedastic models to shorten the credible set while maintaining the coverage rate in a reasonable range.

The last two columns in Table 9 depict the performance of the density forecast. Consistent with the point forecast, the Ti-Hetero and Tv-Hetero models under homogeneous coeffi-

Figure 12: Empirical Application: Posterior Predictive Density for all industries



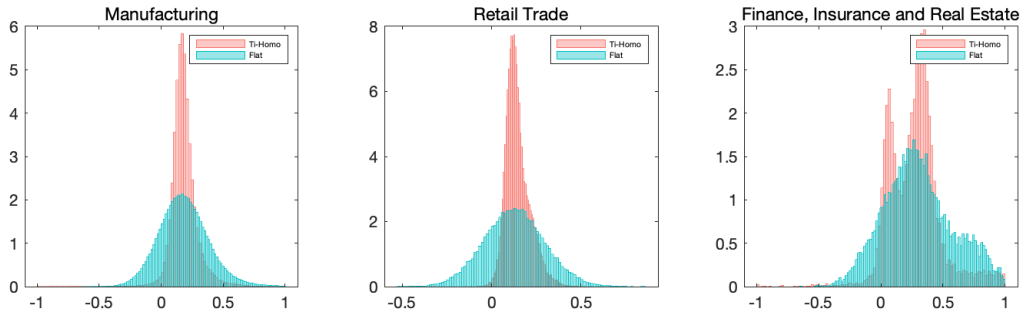
cients specification have comparable performance and dominate the rest with larger LPS and smaller CRPS. The Ti-Hetero has the largest LPS while Tv-Hetero scores the lowest CRPS. These facts emphasize that incorporating homogeneous coefficients and heteroskedasticity is crucial for density forecasting while time effects are not important.

The results for estimation and forecast in other years are presented in Appendix F. In short, the result for 2019 is representative, as most conclusions discussed above also apply for other years. Although no single estimator consistently dominates the rest across the years, at least one of our BGRE estimators always offers the best performance and beats the standard panel data models.

To further investigate the posterior predictive density, we plot the densities of the investment rate generated by the Tv-Hetero and Flat estimator under homogeneous specification in Figure 12. Both posterior predictive densities have similar posterior means while imposing the grouped random effects remarkably sharpens the density around the mean. The reason is that Tv-Hetero estimator leverages latent group structure and pools the information of firms that share great similarities while the Flat estimator treats individual firm separately and makes a prediction based on limited observations.

Figure 13 further aggregates the predictive density over industries. Comparing Tv-Hetero and Flat estimator across industries, several observations stand out. First, Tv-Hetero predictive densities tend to be more concentrated in each industry. This is in line with Figure 12.

Figure 13: Empirical Application: Posterior Predictive Density for selected industries



Second, there is substantial heterogeneity in density forecasts across sectors. While pooling the forecast for all firms yielding a well-behaved bell shape, the posterior predictive densities for the individual industries are in various shapes. This would pose difficulties for the standard estimator to forecast the future and call for a flexible model specification. Third, the Flat estimator is not flexible enough to portray the potential non-normal predictive distribution due to the restrictive normality assumption. In this case, our BGRE estimator, especially the Tv-Hetero estimator, manages to depict the skewed trimodal distribution for the Retail Trade division and bimodality in the Finance, Insurance and Real Estate divison via combining different groups.

6 Conclusion

This paper studies the estimation and prediction of a dynamic panel data model with latent grouped random effects. We adopt a nonparametric Bayesian approach to identify coefficients and group membership in the random effects simultaneously. This approach avoids the severe issue introduced by the ex-post model selection and allows us to incorporate any forms of prior knowledge on group structure. In Monte Carlo experiments, we show that the BGRE estimators have the edge over standard Bayesian estimators. Regarding clustering, the BGRE estimators generate comparable performance with the *Kmeans* algorithm. Our empirical application to investment rates across firms reveals that the estimated latent group structure provides a great amount of flexibility and facilitates point, set, and density forecasts.

The present work raises interesting issues for further research. First, it may be appealing

to consider group structures in the AR(1) parameters and heterogeneous coefficients. This would allow us further to reduce the complexity of a panel data model and may improve predictive performance. Second, more clever attempts could be made to incorporate the subjective prior group structure. Our proposed method summarizes prior information in the prior of random effects or group probability, both of which can be further improved to overcome their limitation. Third, our method can be extended to nonstationary panels, where panel units and co-integrating relationships may possess latent group structures. Four, the assumption that an individual cannot change its group identity during the whole sampling period can be relaxed in the next step, leading to an even more flexible specification.

References

Ando, T., and J. Bai (2016): "Panel data models with grouped factor structure under unknown group membership," *Journal of Applied Econometrics*, 31(1), 163–191.

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

Arellano, M., and S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The Review of Economic Studies*, 58(2), 277–297.

Arellano, M., and O. Bover (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 68(1), 29–51.

Bai, J., and S. Ng (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70(1), 191–221.

Bester, C. A., and C. B. Hansen (2016): "Grouped effects estimators in fixed effects models," *Journal of Econometrics*, 190(1), 197–208.

Biernacki, C., G. Celeux, and G. Govaert (2000): "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, 22(7), 719–725.

Blundell, R., and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87(1), 115–143.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "Discretizing unobserved heterogeneity," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-16).

Bonhomme, S., and E. Manresa (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83(3), 1147–1184.

Celeux, G., F. Forbes, C. P. Robert, and Titterington (2006): "Deviance information criteria for missing data models," *Bayesian Analysis*, 1(4), 651–673.

Chamberlain, G. (1980): "Analysis of covariance with qualitative data," *The Review of Economic Studies*, 47(1), 225–238.

Escobar, M. D., and M. West (1995): "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90(430), 577–588.

Frühwirth-Schnatter, S. (2004): "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques," *The Econometrics Journal*, 7(1), 143–167.

——— (2011): "Panel data analysis: a survey on model-based clustering of time series," *Advances in Data Analysis and Classification*, 5(4), 251–280.

Gala, V. D., J. F. Gomes, and T. Liu (2019): "Investment without q," *Journal of Monetary Economics*.

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

Hastie, D. I., S. Liverani, and S. Richardson (2015): "Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations," *Statistics and Computing*, 25(5), 1023–1037.

Hsiao, C., and A. K. Tahmiscioglu (1997): "A panel analysis of liquidity constraints and firm investment," *Journal of the American Statistical Association*, 92(438), 455–465.

Ishwaran, H., and L. F. James (2001): "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96(453), 161–173.

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

Keribin, C. (2000): "Consistent estimation of the order of mixture models," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 49–66.

Kim, J., and L. Wang (2019): "Hidden group patterns in democracy developments: Bayesian inference for grouped heterogeneity," *Journal of Applied Econometrics*, 34(6), 1016–1028.

Lin, C.-C., and S. Ng (2012): "Estimation of panel data models with parameter heterogeneity when group membership is unknown," *Journal of Econometric Methods*, 1(1), 42–55.

Liu, L. (2020): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," .

Liu, L., H. R. Moon, and F. Schorfheide (2019): "Forecasting with a panel tobit model," Discussion paper, National Bureau of Economic Research.

Liverani, S., D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson (2015): "PReMiuM: An R package for profile regression mixture models using Dirichlet processes," *Journal of Statistical Software*, 64(7), 1.

MacQueen, J. (1967): "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297. Oakland, CA, USA.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016): "Model-based clustering based on sparse finite Gaussian mixtures," *Statistics and Computing*, 26(1-2), 303–324.

McNicholas, P. D., and T. B. Murphy (2010): "Model-based clustering of longitudinal data," *Canadian Journal of Statistics*, 38(1), 153–168.

Molitor, J., M. Papathomas, M. Jerrett, and S. Richardson (2010): "Bayesian profile regression with an application to the National Survey of Children's Health," *Biostatistics*, 11(3), 484–498.

Neal, R. M. (2000): "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

Neyman, J., and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica*, pp. 1–32.

Nickell, S. (1981): "Biases in dynamic models with fixed effects," *Econometrica: Journal of the Econometric Society*, pp. 1417–1426.

Papaspiliopoulos, O., and G. O. Roberts (2008): "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95(1), 169–186.

Su, L., Z. Shi, and P. C. Phillips (2016): "Identifying latent structures in panel data," *Econometrica*, 84(6), 2215–2264.

Su, L., X. Wang, and S. Jin (2019): "Sieve estimation of time-varying panel data models with latent structures," *Journal of Business & Economic Statistics*, 37(2), 334–349.

Sun, Y. (2005): "Estimation and inference in panel structure models," *Available at SSRN 794884*.

Walker, S. G. (2007): "Sampling the Dirichlet mixture model with slices," *Communications in Statistics—Simulation and Computation®*, 36(1), 45–54.

Yau, C., O. Papaspiliopoulos, G. O. Roberts, and C. Holmes (2011): "Bayesian non-parametric hidden Markov models with applications in genomics," *Journal of the Royal Statistical Society*, 73(1), 37–57.

Supplemental Appendix to "Forecasting with Bayesian Grouped Random Effects in Panel Data"

Boyuan Zhang

A Proof

Proposition A.1 *Suppose that we have a model with posterior as given in the section 3.2. Given the definition of the number of potential component K^* (eq.(3.11)), the minimum of auxiliary variables u^* (eq.(3.12)) and the number of active group K (eq.(??)), we have*

(i) $u_i > \pi_k$ *for* $\forall i = 1, 2, ..., n$ *and* $\forall k > K^*$;

(ii) $K < K^*$.

Proof:

(i) By definition, $u^* = \min_{1 \leqslant i \leqslant N} u_i$ for $i = 1, 2, ..., n$, then,

$$u_i \geqslant u^* > 1 - \sum_{j=1}^{K^*} \pi_j = \sum_{j=K^*}^{\infty} \pi_j \geqslant \pi_k, \forall k > K^*.$$

(ii) Let i' be an unit i such that $g_{i'} = K$. According to the posterior of G, the group K exists if $u_{i'} < \pi_K$, otherwise $p(g_i = K|\cdot) = 0$. Then by definition,

$$u^* \leqslant u_{i'} < \pi_K \Rightarrow 1 - u^* > 1 - \pi_K = \sum_{j=1}^{K-1} \pi_j.$$

Since K^* is the smallest number $s.t.$ $1 - u^* < \sum_{j=1}^{K^*} \pi_j$, then $K \leqslant K^*$.

B Data Set

The individual company raw annual data are obtained from the COMPUSTAT database. We constructed the sample using the data from the year 1999 to 2019. The reason to not use the data back to the 1970s is to avoid potential structure breaks in the variable of interest and to reflect the advanced speed of capital accumulation in recent decades. The primary variables of interest are:

- K = Capital stock: net property, plant, and equipment. [PPENT]
- I = Investment: capital expenditures in property, plant, and equipment. [CAPX]
- Y = Sales: net sales revenues. [SALE]
- CF = Cash Flow: income after taxes and interest plus depreciation minus dividends. [EBITDA - TXT - XINT - DVT]

Additional variables used in the alternative model specification:

- Q = Tobin's Q: define as (E+B-INV) / K - 1.
- E = Market value of equity: the sum of common equity and preferred equity. [PRCC$_f$*CSHO + PSTK]
- B = Book value of debt: the sum of short-term and long-term debt. [DLC + DLTT]
- INV = Market value of inventories. [INVT].

The variable names and formula in the bracket are corresponding items in COMPUSTAT. We process the raw data according to the following guidance:

1. Observations where capital stock and sales are either zero or negative are eliminated.
2. Firms that have missing values in the primary variables of interest during 1999-2019 are excluded.
3. We eliminate any firm-year observation if the firm involved in merger and acquisition.
4. Each firm must have valid annual observations from the year 1999 to 2019.

The final sample comprises 337 firms and the observations on each firm is 20. The summary statistics are reported in Table A-1.

Table A-1: Descriptive Statistics for the Variables of Interest

	Min	25%	Median	Mean	75%	Max	StD	Skew.	Kurt.
I/K	0.03	0.11	0.16	0.17	0.21	0.53	0.09	1.41	2.53
CF/K	-1.13	0.12	0.26	0.38	0.51	2.55	0.48	1.55	5.94
Y/K	-1.53	0.54	1.35	1.19	1.95	4.19	1.17	-0.23	-0.09
N/K	-8.63	-5.36	-4.19	-4.56	-3.46	-1.77	1.52	-0.74	-0.12
log(K)	-0.37	5.16	6.77	6.60	8.32	9.82	2.26	-0.63	0.21
q	-0.55	0.83	2.96	7.37	8.32	90.06	12.92	4.00	19.63

Notes: The descriptive statistics are computed across N and T dimension of the panel.

C Posterior Distributions and Algorithms

C.1 Random Effects Model

Below, I present the conditional posterior distribution for the time-varying random effects model with heteroskedasticity, which is the most complicated scenarios. For other models, such as its time-invariant counterparts and homoskedastic model, adjustment can be easily made by eliminating time effects and heteroskedasticity.

In the following derivation and algorithm, we adopt the slice sampler (Walker, 2007) that avoids approximation in Ishwaran and James (2001). With a set of uniformly distributed i.i.d. auxiliary variables $u = [u_1, u_2, ..., u_N]$, we define the largest possible number of potential components as

$$K^* = \min_k \left\{ \sum_{j=1}^{k} \pi_j > 1 - u^* \right\}, \tag{A.1}$$

where

$$u^* = \min_{1 \leqslant i \leqslant N} u_i. \tag{A.2}$$

Such specification ensures that for any group $k > K^*$ and any unit $i \in \{1, 2, ..., N\}$, we have $u_i > \pi_k$[17]. This crucial property limits the dimension of α and Σ to K^* as the densities of α_k and σ_k equal 0 for $k > K^*$ due to $\mathbf{1}(u_i < \pi_k) = 0$, which will be clear in the subsequent posterior derivation.

[17] See proof in proposition A.1

Next, we first define the number of active groups

$$K^a = \max_{1 \leqslant i \leqslant N} g_i. \tag{A.3}$$

It can be shown that $K^a \leqslant K^*$[18].

Conditional posterior of α (grouped random effects).

$$p(\alpha|\rho, \beta, \Sigma, \Xi, a, G, u, Y, X) \propto \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \Sigma_{g_i}) \mathbf{1}(u_i < \pi_{g_i})\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)\right],$$

For $k \in \{1, 2, ..., K^a\}$, define a set of unit that belongs to group k,

$$C_k = \{i \in \{1, 2, ..., N\} | g_i = k\}, \tag{A.4}$$

then the posterior density for α_k read as

$$\begin{aligned}
& p(\alpha_k|\rho, \beta, \Sigma, \Xi, a, G, u, Y, X) \\
\propto & \left[\prod_{i \in C_k} p(y_i|x_i, \rho, \beta_i, \alpha_k, \sigma_k^2)\right] p(\alpha_k|\phi) \\
\propto & \exp\left[-\sum_{i \in C_k} (y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_k)' \Sigma_k^{-1} (y_i - \rho y_{-1,i} - x_i\beta_i - \alpha_k)\right] \exp\left[-(\alpha_k - \mu_\alpha)' \Sigma_\alpha^{-1} (\alpha_k - \mu_\alpha)\right] \\
\propto & \exp\left[-(\alpha_k - \bar{\mu}_{\alpha_k})' \bar{\Sigma}_\alpha^{-1} (\alpha_k - \bar{\mu}_{\alpha_k})\right],
\end{aligned}$$

where $y_{-1,i}$ are lagged values for y_i. Assuming an independent normal conjugate prior for α_k, the posterior for α_k is given by

$$\alpha_k|\rho, \beta, \Sigma, \Xi, a, G, u, Y, X \sim N\left(\bar{\mu}_{\alpha_k}, \bar{\Sigma}_{\alpha_k}\right). \tag{A.5}$$

[18] See proof in proposition A.1

where

$$\bar{\Sigma}_{\alpha_k} = \left(\Sigma_\alpha^{-1} + \sum_{i \in C_k} \Sigma_i^{-1} \right)^{-1},$$
$$\bar{\mu}_{\alpha_k} = \bar{\Sigma}_{\alpha_k} \left[\Sigma_\alpha^{-1} \mu_\alpha + \sum_{i \in C_k} \Sigma_i^{-1} \widetilde{y}_i \right],$$
$$\widetilde{y}_i = y_i - \rho y_{-1,i} - x_i \beta_i.$$

If group k is empty, we draw α_k from its prior $N(\mu_\alpha, \Sigma_\alpha)$.

Conditional posterior of Σ (grouped variance). Under the cross-sectional independence, for $k = 1, 2, ..., K^a$,

$$p(\sigma_k^2 | \rho, \beta, \alpha, \Xi, a, G, u, Y, X) \propto \left[\prod_{i \in C_k} p(y_i | x_i, \rho, \beta_i, \alpha_k, \sigma_k^2) \right] p(\sigma_k^2 | \phi)$$

Assuming a inverse-gamma prior $\sigma_k^2 \sim IG\left(\frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$, the posterior distribution of σ_k^2 is

$$\begin{aligned}
& p(\sigma_k^2 | \rho, \beta, \alpha, G, u, Y, X) \\
\propto & \prod_{i \in C_k} \left[(\sigma_k^2)^{-\frac{T}{2}} \exp\left(-\frac{\sum_{t=1}^T (y_{it} - \rho y_{it-1} - \beta' x_{it} - \alpha_{kt})^2}{2\sigma_k^2} \right) \right] \left(\frac{1}{\sigma_k^2} \right)^{\frac{v_\sigma}{2}+1} \exp\left(-\frac{\delta_\sigma}{2\sigma_k^2} \right) \\
= & \left(\frac{1}{\sigma_k^2} \right)^{\frac{v_\sigma + T|C_k|}{2}+1} \exp\left(-\frac{\delta_\sigma + \sum_{i \in C_k} \sum_{t=1}^T (y_{it} - \rho y_{it-1} - \beta' x_{it} - \alpha_{kt})^2}{2\sigma_k^2} \right).
\end{aligned}$$

This implies

$$\sigma_k^2 | \rho, \beta, \alpha, \Xi, a, G, u, Y, X \sim IG\left(\frac{\bar{v}_{\sigma,k}}{2}, \frac{\bar{\delta}_{\sigma,k}}{2} \right), \tag{A.6}$$

where

$$\begin{aligned}
\bar{v}_{\sigma,k} &= v_\sigma + T|C_k|, \\
\bar{\delta}_{\sigma,kt} &= \delta_\sigma + \sum_{i \in C_k} \sum_{t=1}^{T} (\widetilde{y}_{it} - \alpha_{kt})^2, \\
|C_k| &= \text{occurrence of } g_i = k, \\
\widetilde{y}_{it} &= y_{it} - \rho y_{it-1} - \beta' x_{it}.
\end{aligned}$$

If group k is empty, we draw σ_k^2 from its prior $IG\left(\frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$.

Conditional posterior of ρ (common coefficient). Using a normal conjugate prior $\rho \sim N(\mu_\rho, \Sigma_\rho)$, we could solve standard Bayesian linear regression to get the posterior density of the common coefficient ρ,

$$\begin{aligned}
p(\rho|\beta, \alpha, \Sigma, \Xi, a, G, u, Y, X) &\propto \left[\prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \Sigma_{g_i})\right] p(\rho) \\
&\propto \exp\left[-\sum_{i=1}^{N} (y_i - \alpha_{g_i} - x_i\beta_i - \rho y_{-1,i})' \Sigma_{g_i}^{-1} (y_i - \alpha_{g_i} - x_i\beta_i - \rho y_{-1,i})\right] \exp\left[-(\rho - \mu_\rho)' \Sigma_\rho^{-1} (\rho - \mu_\rho)\right].
\end{aligned}$$

This implies

$$\rho|\beta, \alpha, \Sigma, \Xi, a, G, u, Y, X \sim \mathcal{N}\left(\bar{\mu}_\rho, \bar{\Sigma}_\rho\right), \tag{A.7}$$

where

$$\begin{aligned}
\bar{\Sigma}_\rho &= \left(\Sigma_\rho^{-1} + \sum_{i=1}^{N} y_{-1,i}' \Sigma_{g_i}^{-1} y_{-1,i}\right)^{-1}, \\
\bar{\mu}_\rho &= \bar{\Sigma}_\rho \left[\Sigma_\rho^{-1} \mu_\rho + \sum_{i=1}^{N} y_{-1,i}' \Sigma_{g_i}^{-1} \hat{y}_i\right], \\
\hat{y}_i &= y_i - \alpha_{g_i} - x_i\beta_i.
\end{aligned}$$

Conditional posterior of β (heterogeneous coefficients). As ε_{it} is independent across units, we solve for β for each unit separately. We transform the model into a standard linear

model with a known form of heteroskedasticity,

$$y_{it} - \alpha_{g_i t} - \rho y_{it-1} = \beta_i' x_{it} + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma_{g_i}^2).$$

Using a normal conjugate prior $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$, for the unit i, the posterior distribution is written as,

$$\begin{aligned} & p(\beta_i|\rho, \alpha, \Sigma, \Xi, a, G, u, Y, X) \\ \propto\ & p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma_{g_i}^2) p(\beta_i) \\ \propto\ & \exp\left[-\frac{\sum_{t=1}^T (y_{it} - \alpha_{g_i} - \rho y_{it} - x_{it}'\beta_i)^2}{2\sigma_{g_i}^2} \right] \exp\left[-(\beta_i - \mu_\beta)' \Sigma_\beta^{-1} (\beta_i - \mu_\beta) \right]. \end{aligned}$$

This implies

$$\beta_i|\rho, \alpha, \Sigma, \Xi, a, G, u, Y, X \sim \mathcal{N}\left(\bar{\mu}_{\beta_i}, \bar{\Sigma}_{\beta_i} \right), \tag{A.8}$$

where

$$\begin{aligned} \bar{\Sigma}_{\beta_i} &= \left(\Sigma_\rho^{-1} + \sigma_{g_i}^{-2} \sum_{t=1}^T x_{it} x_{it}' \right)^{-1} \\ \bar{\mu}_{\beta_i} &= \bar{\Sigma}_\rho \left[\Sigma_\rho^{-1} \mu_\rho + \sigma_{g_i}^{-2} \sum_{t=1}^T x_{it} \hat{y}_{it} \right] \\ \hat{y}_{it} &= y_{it} - \alpha_{g_i} - \rho y_{it-1} \end{aligned}$$

Conditional posterior of Ξ (stick length).

$$\begin{aligned} p(\Xi|\rho, \beta, \alpha, \Sigma, a, G, u, Y, X) \propto & \left[\prod_{i=1}^N p(u_i|\pi_{g_i}) \pi_{g_i} \right] \left[\prod_{j=1}^\infty p(\xi_j|a) \right] \\ \propto & \left[\prod_{i=1}^N p(u_i|\pi_{g_i}) \xi_{g_i} \prod_{l<g_i} (1 - \xi_l) \right] \left[\prod_{j=1}^\infty p(\xi_j|a) \right] \end{aligned}$$

For $k = 1, 2, ..., K^a$,

$$p(\xi_k|\rho, \beta, \alpha, \Sigma, a, G, u, Y, X) \propto \left(\prod_{i \in C_k} \xi_k\right)(1-\xi_k)^{\sum_{j=1}^{N} \mathbf{1}(g_j > k)}(1-\xi_k)^{a-1}$$
$$\propto \xi_k^{|C_k|}(1-\xi_k)^{a+\sum_{j=1}^{N} \mathbf{1}(g_j > k)-1}.$$

Therefore, posterior distribution of ξ_k is

$$\xi_k|\rho, \beta, \alpha, \Sigma, a, G, u, Y, X \sim Beta\left(|C_k| + 1, a + \sum_{j=1}^{N} \mathbf{1}(g_j > k)\right). \tag{A.9}$$

Give $\Xi = [\xi_1, \xi_2, ..., \xi_{K^a}]$, update $\pi_1, \pi_2, ..., \pi_{K^a}$,

$$\pi_k|G, \Xi \sim \begin{cases} \xi_1, & k = 1 \\ \xi_k \prod_{j<k}(1-\xi_j), & k = 2, \ldots, K^a \end{cases}. \tag{A.10}$$

Conditional posterior of a (concentration parameter). Regarding the DP concentration parameter, the standard posterior derivation doesn't work due to the unrestricted number of components in the current sampler. Instead, we implement the 2-step procedure proposed by Escobar and West (1995) (p.8-9). Following their approach, we first draw a latent variable η from

$$\eta \sim Beta\ (a+1, N). \tag{A.11}$$

Then, conditional on η and K^a, we assume sample a from a mixture of two Gamma distribution:

$$p(a|\eta, K^a) = \pi_a Gamma\ (m + K^a, n - \log(\eta)) + (1-\pi_a) Gamma\ (m + K^a - 1, n - \log(\eta)), \tag{A.12}$$

with the weights π_a defined by

$$\frac{\pi_a}{1-\pi_a} = \frac{m + K^a - 1}{N[n - \log(\eta)]}.$$

Conditional posterior of u (auxiliary variable). Conditional on the group "stick lengths" ξ_k and group member indices G, it is straightforward to show that the posterior density of u_i is a uniform distribution ranging define on $(0, \pi_{g_i})$, that is

$$u_i|\Xi, G \sim \mathit{Unif}\,(0, \pi_{g_i}), \tag{A.13}$$

where $\pi_{g_i} = \xi_{g_i} \prod_{j<g_j}(1-\xi_j)$. Moreover, it is worth noting that the values for K^* and u^* need to be updated according to (A.1) and (A.2) after this step.

Conditional posterior of G (group membership). We derive the posterior distribution of g_i consider on $G^{(i)}$, where $G^{(i)}$ is a set including all member indices except for g_i, i.e., $G^{(i)} = G \backslash g_i$. Hence, for $k = 1, 2, ..., K^*$,

$$p(g_i = k|\rho, \beta, \alpha, \Sigma, \Xi, a, G^{(i)}, u, Y, X) \propto p(y_i|\rho, \beta_i, \alpha_k, \sigma_k^2, Y, X)\mathbf{1}(u_i < \pi_{g_i}).$$

As per a discrete distribution, we normalize the point mass to get a valid distribution:

$$p(g_i = k|\rho, \beta, \alpha, \Sigma, \Xi, a, G^{(i)}, u, Y, X) = \frac{p(y_i|\rho, \beta_i, \alpha_k, \sigma_k^2, Y, X)\mathbf{1}(u_i < \pi_k)}{\sum_{j=1}^{K^*} p(y_i|\rho, \beta_i, \alpha_j, \sigma_j^2, Y, X)\mathbf{1}(u_i < \pi_j)}. \tag{A.14}$$

C.1.1 Blocked Gibbs Sampler and Algorithm

Initialization:

(i) Preset the initial number of active groups K_0^a. As derived by Antoniak (1974), the expected number of unique groups is $E\,[K|a] \approx a \log\left(\frac{a+N}{a}\right)$. We set K_0^a to its expected value with concentration parameter a replaced by prior mean.

(ii) In ignorance of group heterogeneity ($K = 1$) and heteroskedasticity, run OLS to get $\hat{\alpha}_{OLS}$, $\hat{\rho}_{OLS}$, $\hat{\beta}_{iOLS}$ and $Cov(\hat{\alpha}_{OLS})$. These OLS estimators serve as the mean and covariance matrix in the related priors.

(iii) Generate K_0^* random sample from the distribution $N(\hat{\alpha}_{OLS}, Cov(\hat{\alpha}_{OLS}))$.

(iv) Initialize group membership G by sampling from (A.14) ignoring $\mathbf{1}(u_i < \pi_{g_i})$. Remove empty groups.

For each iteration $s = 1, 2, .., N_{sim}$

(i) Number of active groups:

$$K^a = \max_{1 \leqslant i \leqslant N} g_i^{(s-1)}.$$

(ii) Group "stick length": for $k = 1, 2, ..., K^a$, draw ξ_k from a Beta distribution in (A.9):

$$\xi_k | \rho^{(s-1)}, \beta^{(s-1)}, \alpha^{(s-1)}, \Sigma^{(s)}, a^{(s-1)}, G^{(s-1)}, u^{(s-1)}, Y, X \sim Beta\left(|C_k| + 1, a + \sum_{j=1}^{N} \mathbf{1}(g_j > k) \right),$$

and calculate group probability in accordance to (A.10).

(iii) Group heterogeneity: for $k = 1, 2, ..., K^a$, draw $\alpha_k^{(s)}$ from a normal distribution in (A.5):

$$\alpha_k | \rho^{(s-1)}, \beta^{(s-1)}, \Sigma^{(s-1)}, a^{(s-1)}, G^{(s-1)}, u^{(s-1)}, Y, X \sim N\left(\bar{\mu}_{\alpha_k}, \bar{\Sigma}_{\alpha_k}\right).$$

(iv) Group heteroskedasticity: for $k = 1, 2, ..., K^a$ and $t = 1, 2, ..., T$, draw $\sigma_k^{2(s)}$ from an inverse Gamma distribution in (A.6):

$$\sigma_k^2 | \rho^{(s-1)}, \beta^{(s-1)}, \alpha^{(s)}, G^{(s-1)}, u^{(s-1)}, Y, X \sim IG\left(\frac{\bar{v}_{\sigma,k}}{2}, \frac{\bar{\delta}_{\sigma,k}}{2} \right).$$

(v) Label switching[19]: after each iteration an additional random permutation step is added to the MCMC scheme which randomly permutes the current labeling of the components. Random permutation ensures that the sampler explores all K! modes of the full posterior distribution and avoids that the sampler is trapped around a single posterior mode. Following Liu (2020)[20], we update $\left\{ \alpha_k^{(s)}, \sigma_k^{2(s)}, \pi_k^{(s)}, g_i^{(s-1)} \right\}$ by three Metropolis-Hastings label-switching moves developed by Papaspiliopoulos and Roberts (2008) (step (a) and (b)) and Hastie, Liverani, and Richardson (2015) (step (c)). All these label switching moves aim to improve numerical convergence.

(a) Randomly select two nonempty groups i and j, swap group labels $g_i^{(s-1)}$ and $g_j^{(s-1)}$

[19] Without this step, the one-at-a-time updates of the allocations mean that clusters rarely switch labels, and consequentially the ordering will be largely determined by the (perhaps random) initialization of the sampler.

[20] See Algorithm C.4 in the appendix

for all units in these groups, accept new label with probability:

$$\min\left(1, \frac{\pi_i^{N_j}\pi_j^{N_i}}{\pi_i^{N_i}\pi_j^{N_j}}\right) = \min\left(1, (\pi_i/\pi_j)^{N_j - N_i}\right),$$

where N_i, N_j are the number of units in the group i and j respectively.

(b) Randomly select two adjacent groups l and $l+1$ such that $\{l, l+1\} \subset \{1, 2, ..., K^a\}$, swap group label $g_l^{(s-1)}$ **and** "stick length" $\xi_l^{(s)}$, accept new label and stick length with probability:

$$\min\left(1, \frac{\tilde{p}_l^{N_{l+1}}\tilde{p}_{l+1}^{N_l}}{\pi_l^{N_l}\pi_{l+1}^{N_{l+1}}}\right),$$

where $\tilde{p}_i$ and $\tilde{p}_j$ are new group probabilities derived with new $\xi_l^{(s)}$ and $\xi_{l+1}^{(s)}$.

(c) Randomly select two adjacent groups k and $k+1$ such that $\{k, k+1\} \subset \{1, 2, ..., K^a\}$, swap group label $g_i^{(s-1)}$, "stick length" $\xi_k^{(s)}$ **and update** group-specific parameter $\{\alpha_k^{(s)}, \sigma_k^{2(s)}\}$, accept new new label and stick length with probability

$$\min\left\{1, \left(R_1/\widetilde{R}\right)^{N_{k+1}}\left(R_2/\widetilde{R}\right)^{N_k}\right\},$$

where

$$\begin{aligned} R_1 &= \frac{1 + a + N_{k+1} + \sum_{l>k+1} N_l}{a + N_{k+1} + \sum_{l>k+1} N_l}, \\ R_2 &= \frac{a + N_k + \sum_{l>k+1} N_l}{1 + a + N_k + \sum_{l>k+1} N_l}, \\ \widetilde{R} &= \frac{\pi_{k+1}R_1 + \pi_k R_2}{\pi_k + \pi_{k+1}}. \end{aligned}$$

The new group probability is defined as $p'_k = \pi_{k+1}R_1/\widetilde{R}$ and $p'_{k+1} = \pi_k R_2/\widetilde{R}$.

Additionally, we update the "stick lengths"[21] for group k and $k+1$ such that

$$\xi'_k = \frac{p'_k}{\prod_{l<c}(1-\xi_l)},$$
$$\xi'_{k+1} = \frac{p'_{k+1}}{(1-\xi'_k)\prod_{l<c}(1-\xi_l)}.$$

(vi) Auxiliary variables: for $i = 1, 2, ..., N$, draw u_i from an uniform distribution in (A.13):

$$u_i|\Xi^{(s)}, G^{(s)} \sim \mathrm{U}(0, p^{(s)}_{g_i}).$$

Then calculate u^* according to (A.2).

(vii) DP concentration parameter:

(a) Draw latent variable η from a Beta distribution in (A.11):

$$\eta \sim Beta(a+1, K^a)$$

(b) Draw concentration parameter a from a mixture of Gamma distribution in (A.12):

$$a|\eta, K^a \sim \begin{cases} Gamma\left(m + K^a, n - \log(\eta)\right) & \text{with prob. } \pi_a \\ Gamma\left(m + K^a - 1, n - \log(\eta)\right) & \text{with prob. } 1-\pi_a \end{cases},$$

and π_a is defined as

$$\frac{\pi_a}{1-\pi_a} = \frac{m + K^a - 1}{N(n - \log(\eta))}.$$

(viii) Potential groups: start with $\tilde{K} = K^a$,

(a) Group probabilities:

(1) if $\sum_{j=1}^{\tilde{K}} \pi_j^{(s)} > 1 - u^*$, set $K^* = \tilde{K}$ and stop

(2) otherwise, let $\tilde{K} = \tilde{K}+1$, draw $\xi_{\tilde{K}} \sim Beta\left(1, \alpha^{(s)}\right)$, update $\pi_{\tilde{K}} = \xi_{\tilde{K}} \prod_{j<\tilde{K}} (1-\xi_j)$ and go to step (1)

(b) Group parameters: for $k = K+1, \cdots, K^*$, draw $\alpha_k^{(s)}$ and $\sigma_k^{2(s)}$ from their prior distributions.

[21] This particular choices of ξ'_k and ξ'_{k+1} ensure the group probabilities that are changed are those associated with the the group k and $k+1$, and the rest are unchanged. Moreover, it can be shown that $(1-\xi'_k)(1-\xi'_{k+1}) = (1-\xi_k)(1-\xi_{k+1})$. See more details in the appendices of Hastie, Liverani, and Richardson (2015).

(ix) Common AR(1) parameter: draw ρ from a normal distribution in (A.7):

$$\rho|\beta^{(s-1)}, \alpha^{(s)}, \Sigma^{(s)}, a^{(s)}, G^{(s-1)}, u^{(s)}, Y, X \sim N\left(\bar{\mu}_{\rho}, \bar{\Sigma}_{\rho}\right).$$

(x) Heterogeneous parameter: draw β_i from a normal distribution in (A.8):

$$\beta_i|\rho^{(s)}, \alpha^{(s)}, \Sigma^{(s)}, a^{(s)}, G^{(s-1)}, u^{(s)}, Y, X \sim N\left(\bar{\mu}_{\beta_i}, \bar{\Sigma}_{\beta_i}\right).$$

(xi) Group membership: for $i = 1, 2, ..., N$ and $k = 1, 2, ..., K^*$, draw g_i from a multinomial distribution in (A.14):

$$p(g_i = k|\rho^{(s)}, \beta^{(s)}, \alpha^{(s)}, \Sigma^{(s)}, \xi^{(s)}, a^{(s)}, G^{(i)}, u^{(s)}, Y, X) = \frac{p(y_i|\rho^{(s)}, \beta_i^{(s)}, \alpha_k^{(s)}, \Sigma_k^{(s)})\mathbf{1}(u_i^{(s)} < \pi_k)}{\sum_{j=1}^{K^*} p(y_i|\rho^{(s)}, \beta_i^{(s)}, \alpha_j^{(s)}, \Sigma_j^{(s)})\mathbf{1}(u_j^{(s)} < \pi_{g_j})}$$

C.2 Random Effects Model with Subjective Group Probability Prior

This algorithm is designed for the random effect model where econometricians have prior knowledge on the group structure and presume the number of groups K^p. Building on the algorithm for the random effect model in Section C.1, we allow for incorporating the researchers' prior knowledge while inheriting the feature of reallocating units and changing the number of groups along the MCMC sampling.

We use the same notation as in Section 3.2,

$$\begin{aligned}
\text{Observations: } & Y = [y_1, y_2, ..., y_N], y_i = [y_{i1}, y_{i2}, ..., y_{iT}]', \\
\text{Covariates: } & X = [x_1, x_2, ..., x_N], x_i = [x_{i1}, x_{i2}, ..., x_{iT}]', \\
\text{Random effects: } & \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K], \\
\text{Covariance matrices: } & \Sigma = [\Sigma_1, \Sigma_2, \ldots, \Sigma_K], \\
\text{Group membership: } & G = [g_1, \ldots, g_N], \\
\text{Stick length: } & \Xi = [\xi_1, \xi_2, \ldots], \\
\text{Group probability: } & \pi = [\pi_1, \ldots, \pi_N], \\
\text{Membership probability: } & \omega = [\omega_1, \ldots, \omega_N], \omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{iK}]', \\
\text{Auxiliary varaible: } & u = [u_1, u_2, ..., u_N], \\
\text{Hyper parameters: } & \phi = [\mu_\alpha, \Sigma_\alpha, \nu_\sigma, \delta_\sigma].
\end{aligned}$$

Notice that we define two sets of probabilities, π and ω. In practice, they have distinct roles in the algorithm. The group membership π captures the groups' probability based on the entire sample, and, most importantly, determines the upper bound of the auxiliary variable u_{g_i} that has direct effect on the potential number of groups K^*. On the other hand, the membership probability ω_i represents the probabilities of a unit i belonging to each of K groups, through which the researcher's prior knowledge enters the algorithm.

As regards the choices of prior, we adopt the independent Multivariate Normal-Inverse-Gamma prior Dirichlet Process priors for group random effects $\alpha_{g_i t}$ and heteroskedasticity $\sigma^2_{g_i}$, a normal prior for the common parameter ρ, an unsymmetric Dirichlet prior for the membership probability ω with concentration parameters chosen by the econometrician, a multinomial prior for Group membership g_i, a Beta prior for the stick length ξ, and a mixture Gamma prior for the concentration parameter a

The posterior of unknown objects in this random effects model is:

$$
\begin{aligned}
& p(\rho, \beta, \alpha, \Sigma, G, \omega | Y, X) \\
\propto\ & p(Y|X, \rho, \beta, \alpha, \Sigma, G, \omega) p(\rho, \beta, \alpha, \Sigma, G, \pi) \\
\propto\ & p(Y|X, \rho, \beta, \alpha, \Sigma, G) p(\alpha, \Sigma|\phi) p(G|p) p(\rho) p(\beta) p(\omega|a) \\
=\ & \prod_{i=1}^{N} p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma^2_{g_i}) \prod_{j=1}^{K} p(\alpha_j, \sigma^2_j|\phi) \prod_{i=1}^{N} p(g_i|\omega_i) \prod_{i=1}^{N} p(\omega_i|a_i) \prod_{i=1}^{N} p(\beta_i) p(\rho) \\
=\ & \prod_{i=1}^{N} \left[p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma^2_{g_i}) p(g_i|\omega_i) p(\omega_i|a_i) p(\beta_i) \right] \prod_{j=1}^{K} p(\alpha_j, \sigma^2_j|\phi) p(\rho).
\end{aligned}
$$

To allow for automatically adjustment for the number of groups, we introduce a set of auxiliary variables $u = [u_1, u_2, ..., u_N]$ proposed by Walker (2007) and rewrite the posterior above as,

$$
\begin{aligned}
& p(\rho, \beta, \alpha, \Sigma, \Xi, a, G, u, \omega, \pi | Y, X) \\
\propto\ & \prod_{i=1}^{N} \left[p(y_i|x_i, \rho, \beta_i, \alpha_{g_i}, \sigma^2_{g_i}) p(g_i|\omega_i) p(\omega_i|a_i) \mathbf{1}(u_i < \pi_{g_i}) p(\beta_i) \right] \prod_{j=1}^{K} p(\alpha_j, \sigma^2_j|\phi) p(\rho).
\end{aligned}
$$

The number of potential groups K^* and the number of active groups K^a are defined in the equation (A.1) and (A.3).

Conditional posterior of α (grouped random effects). Identical to (A.5).

Conditional posterior of Σ (grouped variance). Identical to (A.6).

Conditional posterior of ρ (common coefficient). Identical to (A.7).

Conditional posterior of β (heterogeneous coefficients). Identical to (A.8).

Conditional posterior of Ξ (stick length). Identical to (A.9). Then generate π in accordance to (A.10).

Conditional posterior of a (concentration parameter). Identical to (A.12).

Conditional posterior of u (auxiliary variable). Identical to (A.13).

Conditional posterior of ω (membership probability). Sampling from the posterior of π can be implemented as follows. As we adopt Dirichlet prior for π and Multinomial prior

for g_i, for $i = 1, ..., N$, the posterior is written as,

$$\begin{aligned} p(\omega_i|\rho, \beta, \alpha, \Sigma, G, Y, X) &\propto p(g_i|\omega_i)p(\omega_i|a_i) \\ &\propto \left(\omega_{i1}^{\mathbf{1}(g_i=1)} \dots \omega_{iK^p}^{\mathbf{1}(g_i=K^p)}\right) \times \left(\omega_{i1}^{a_{i1}-1} \dots \omega_{ik}^{a_{iK^p}-1}\right) \\ &= \omega_{i1}^{a_{i1}+\mathbf{1}(g_i=1)-1} \dots \omega_{iK^p}^{a_{i1}+\mathbf{1}(g_i=K^p)-1}. \end{aligned} \tag{A.15}$$

This implies

$$\omega_i|\rho, \beta, \alpha, \Sigma, G, Y, X \sim Dir\left(a_{i1} + \mathbf{1}(g_i = 1), \dots, a_{iK^p} + \mathbf{1}(g_i = K^p)\right). \tag{A.16}$$

It is worth noting that, during MCMC sampling, we allow for more/fewer groups than the researcher expects, i.e., the potential number of groups $K^{*(s)}$ could be larger or smaller than K^p in some iteration s. In such circumstances, we modify the Dirichlet posterior distribution in (A.15) to account for such changes. Notably, we present the posterior in three cases.

Case 1: $K^{*(s)} = K^p$. The posterior distribution in (A.15) is still valid.

Case 2: $K^{*(s)} > K^p$. We have to address new groups. For the additional groups such that $k^\dagger > K^p$, we have, for $\forall i$,

$$\mathbf{1}(g_i^{(s-1)} = k^\dagger) = 0 \text{ if } k^\dagger > K^{a(s-1)} \text{ and } a_{ik^\dagger} = 0.$$

where $K^{a(s-1)} = \max_{1 \leqslant i \leqslant N} g_i^{(s)}$ denotes the number of active groups in the previous iteration.

To ensure a non-negative posterior probability $\omega_{ik^\dagger}$ for the new groups, we assume that, for some ϵ[22],

$$\sum_{k=k^\dagger}^{K^{*(s)}} \tilde{a}_{ik^\dagger} = \epsilon \text{ and } \tilde{a}_{im} = \tilde{a}_{in}, \forall m, n \geqslant k^\dagger.$$

and adjust the prior membership probability a_{ik} if $k < k^\dagger$ by multiplying $1-\epsilon$. This step artificially assigns nonzero probabilities to news groups and forms a set of new hyperparameters $\tilde{a}_{ik}$ such that $\sum_{k=1}^{K^{*(s-1)}} \tilde{a}_{ik} = 1$. Then we draw ω_i from the posterior density,

$$\omega_i|\rho, \beta, \alpha, \Sigma, G, Y, X \sim Dir\left(\tilde{a}_{i1} + \mathbf{1}(g_i = 1), \dots, \tilde{a}_{iK^{*(s)}} + \mathbf{1}(g_i = K^{*(s)})\right). \tag{A.17}$$

[22]In the simulation, we set $\epsilon = 0.3$ so that a unit has a chance of 30% to be assigned to the new groups a priori.

Case 3: $K^{*(s)} < K^p$. We have few groups than the researcher assumes and a unit i might be assigned to a group that is no longer considered in the current iteration (i.e., $K^{*(s)} < g_i^{(s-1)}$). In this case, we need to select and renormalize a subset of a_{ik} since some groups are dismissed. In this regard, for a unit i, we select $K^{*(s)}$ most frequent non-empty groups among the groups visited in the previous iteration $s-1$. If there are not enough candidates, we add back non-selected groups in the first $K^{*(s)}$ out of the K^p groups. Then we normalize the selected a_{ik}[23] and get $\hat{a}_{ik}$. Finally we draw ω_i from the posterior density,

$$\omega_i|\rho,\beta,\alpha,\Sigma,G,Y,X \sim Dir\left(\hat{a}_{i1}+\mathbf{1}(g_i=1),\ldots,\hat{a}_{iK^{*(s)}}+\mathbf{1}(g_i=K^{*(s)})\right). \tag{A.18}$$

Conditional posterior of G (group membership). We derive the posterior distribution of g_i consider on $G^{(i)}$, where $G^{(i)}$ is a set including all member indices except for g_i, i.e., $G^{(i)} = G \backslash g_i$. Hence, for $k = 1, 2, ..., K^*$,

$$p(g_i=k|\rho,\beta,\alpha,\Sigma,G^{(i)},\pi,\omega,Y,X) \propto p(y_i|\rho,\beta_i,\alpha_k,\Sigma_k,Y,X)\omega_{ik}\mathbf{1}(u_i<\pi_{g_i}).$$

As per a discrete distribution, we normalize the point mass to get a valid distribution,

$$p(g_i=k|\rho,\beta_i,\alpha,\Sigma,G^{(i)},\pi,\omega,Y,X) = \frac{p(y_i|\rho,\beta_i,\alpha_k,\Sigma_k,Y,X)\omega_{ik}\mathbf{1}(u_i<\pi_k)}{\sum_{j=1}^{K} p(y_i|\rho,\beta,\alpha_j,\Sigma_j,Y,X)\omega_{ij}\mathbf{1}(u_i<\pi_j)}. \tag{A.19}$$

C.3 Random Effects Model with Group Structures in α, ρ and β

In this subsection, I present the conditional posterior distribution for the time-invariant random effects model with group structures in α, ρ and β. The model is,

$$y_{it} = \theta_{g_i}' \breve{x}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_{g_i}^2\right),$$

where $\breve{x}_{it} = [1\ y_{it-1}\ x_{it}']'$, and $\theta_{g_i} = [\alpha_{g_i}\ \rho_{g_i}\ \beta_{g_i}']'$.

[23] In practice, if all selected a_{ik} are zero for some i, we simply assume $a_{ik} = 1/K^{*(s)}$.

We use the same notation as in Section 3.2 aside from the coefficients θ,

$$\begin{aligned}
\text{Observations: } Y &= [y_1, y_2, ..., y_N], y_i = [y_{i1}, y_{i2}, ..., y_{iT}]', \\
\text{Covariates: } X &= [x_1, x_2, ..., x_N], x_i = [x_{i1}, x_{i2}, ..., x_{iT}]', \\
\text{Stacked coefficients: } \theta &= [\theta_1, \theta_2, \ldots], \\
\text{Covariance matrices: } \Sigma &= [\sigma_1^2, \sigma_2^2, \ldots], \\
\text{Stick length: } \Xi &= [\xi_1, \xi_2, \ldots], \\
\text{Group membership: } G &= [g_1, \ldots, g_N], \\
\text{Auxiliary varaible: } u &= [u_1, u_2, ..., u_N], \\
\text{Hyper parameters: } \phi &= [\mu_\theta, \Sigma_\theta, \nu_\sigma, \delta_\sigma].
\end{aligned}$$

The posterior of unknown objects in the random effects model is,

$$\begin{aligned}
&p(\theta, \Sigma, \Xi, a, G, u | Y, X) \\
&\propto \left[\prod_{i=1}^{N} p(y_i | x_i, \theta_{g_i}, \sigma_{g_i}^2) \mathbf{1}(u_i < \pi_{g_i})\right] \left[\prod_{j=1}^{\infty} p(\theta_j, \sigma_j^2 | \phi) p(\xi_j | a)\right] p(\rho) p(a). \qquad \text{(A.20)}
\end{aligned}$$

The number of potential groups K^* and the number of active groups K^a are defined in the equation (A.1) and (A.3).

Conditional posterior of θ (grouped random effects).

$$p(\theta | \Sigma, \Xi, a, G, u, Y, X) \propto \left[\prod_{i=1}^{N} p(y_i | x_i, \theta_{g_i}, \Sigma_{g_i}) \mathbf{1}(u_i < \pi_{g_i})\right] \left[\prod_{j=1}^{\infty} p(\theta_j, \sigma_j^2 | \phi)\right]$$

For $k \in \{1, 2, ..., K^a\}$, define a set of unit that belongs to group k,

$$C_k = \{i \in \{1, 2, ..., N\} | g_i = k\}, \qquad \text{(A.21)}$$

then the posterior density for θ_k read as

$$\begin{aligned}
& p(\theta_k|\Sigma, \Xi, a, G, u, Y, X) \\
\propto & \left[\prod_{i \in C_k} p(y_i|x_i, \theta_k, \sigma_k^2)\right] p(\theta_k|\phi) \\
\propto & \exp\left[-\sum_{i \in C_k} \left(y_i - \breve{x}_i\theta_k\right)' \Sigma_k^{-1} \left(y_i - \breve{x}_i\theta_k\right)\right] \exp\left[-\left(\theta_k - \mu_\theta\right)' \Sigma_\theta^{-1} \left(\theta_k - \mu_\theta\right)\right] \\
\propto & \exp\left[-\left(\theta_k - \bar{\mu}_{\theta_k}\right)' \bar{\Sigma}_\theta^{-1} \left(\theta_k - \bar{\mu}_{\theta_k}\right)\right].
\end{aligned}$$

Assuming an independent normal conjugate prior for θ_k, the posterior for θ_k is given by

$$\theta_k|\Sigma, \Xi, a, G, u, Y, X \sim N\left(\bar{\mu}_{\theta_k}, \bar{\Sigma}_{\theta_k}\right), \tag{A.22}$$

where

$$\begin{aligned}
\bar{\Sigma}_{\theta_k} &= \left(\Sigma_\theta^{-1} + \sigma_k^{-2} \sum_{i \in C_k} \breve{x}_i \breve{x}_i'\right)^{-1}, \\
\bar{\mu}_{\theta_k} &= \bar{\Sigma}_{\theta_k} \left[\Sigma_\theta^{-1} \mu_\theta + \sigma_k^{-2} \sum_{i \in C_k} \breve{x}_i y_i\right].
\end{aligned}$$

If group k is empty, we draw θ_k from its prior $N\left(\mu_\theta, \Sigma_\theta\right)$.

Conditional posterior of Σ (grouped variance). Under the cross-sectional independence, for $k = 1, 2, ..., K^a$,

$$p(\sigma_k^2|\theta, \Xi, a, G, u, Y, X) \propto \left[\prod_{i \in C_k} p(y_i|x_i, \rho, \beta_i, \alpha_k, \sigma_k^2)\right] p(\sigma_k^2|\phi)$$

Assuming a inverse-gamma prior $\sigma_k^2 \sim IG\left(\frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$, the posterior distribution of σ_k^2 is

$$\begin{aligned}
& p(\sigma_k^2|\rho, \beta, \alpha, G, u, Y, X) \\
\propto & \prod_{i \in C_k} \left[(\sigma_k^2)^{-\frac{T}{2}} \exp\left(-\frac{\sum_{t=1}^T (y_{it} - \theta_k' \breve{x}_{it})^2}{2\sigma_k^2}\right)\right] \left(\frac{1}{\sigma_k^2}\right)^{\frac{v_\sigma}{2}+1} \exp\left(-\frac{\delta_\sigma}{2\sigma_k^2}\right) \\
= & \left(\frac{1}{\sigma_k^2}\right)^{\frac{v_\sigma + T|C_k|}{2}+1} \exp\left(-\frac{\delta_\sigma + \sum_{i \in C_k} \sum_{t=1}^T (y_{it} - \theta_k' \breve{x}_{it})^2}{2\sigma_k^2}\right).
\end{aligned}$$

This implies

$$\sigma_k^2|\rho,\beta,\alpha,\Xi,a,G,u,Y,X \sim IG\left(\frac{\bar{v}_{\sigma,k}}{2},\frac{\bar{\delta}_{\sigma,k}}{2}\right), \tag{A.23}$$

where

$$\begin{aligned}
\bar{v}_{\sigma,k} &= v_\sigma + T|C_k|,\\
\bar{\delta}_{\sigma,kt} &= \delta_\sigma + \sum_{i\in C_k}\sum_{t=1}^{T}(y_{it}-\theta_k'\breve{x}_{it})^2,\\
|C_k| &= \text{occurrence of } g_i = k.
\end{aligned}$$

If group k is empty, we draw σ_k^2 from its prior $IG\left(\frac{v_\sigma}{2},\frac{\delta_\sigma}{2}\right)$.

Conditional posterior of Ξ (stick length). Identical to (A.9).

Conditional posterior of a (concentration parameter). Identical to (A.12).

Conditional posterior of u (auxiliary variable). Identical to (A.13).

Conditional posterior of G (group membership). Identical to (A.14).

D Convergence Diagnostic

To assess convergence, we assess the trace plot, cumulative mean, and auto-correlation of posterior draws for different coefficients. In particular, the data generating process used here is DGP7, where we assume time-varying grouped random effects and homoskedasticity. We evaluate the most complicated BGRE estimator: Tv-Hetero (time-varying α_i, heteroskedasticity), and report the convergence diagnostics for $\alpha_{5,1}$, σ_{10}^2 and ρ[24].

[24]Due to time effects and heteroskedasticity, we randomly present one of the α for unit $i = 5$ and in period $t = 1$, and the variance of error term σ^2 for unit $i = 10$.

Figure A-1: Convergence Diagnostics, $\alpha_{5,1}$ $(i = 5,\ t = 1)$

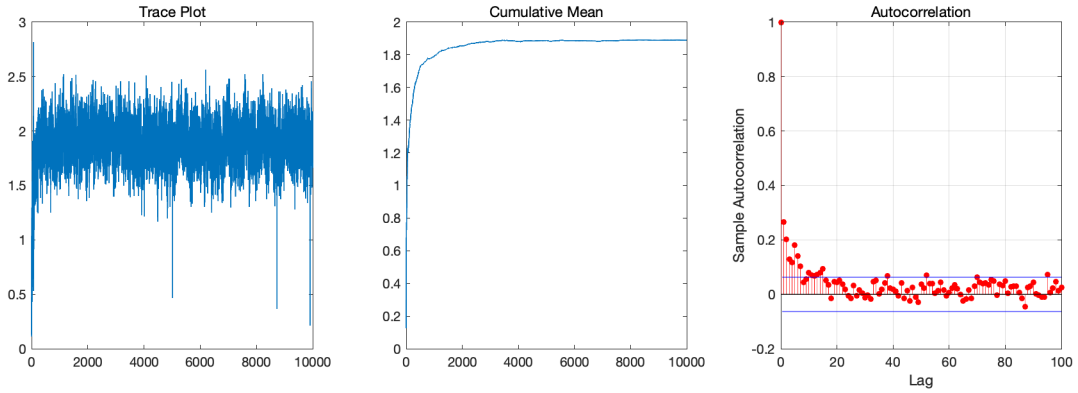


Figure A-2: Convergence Diagnostics, σ_{10}, $(i = 10)$

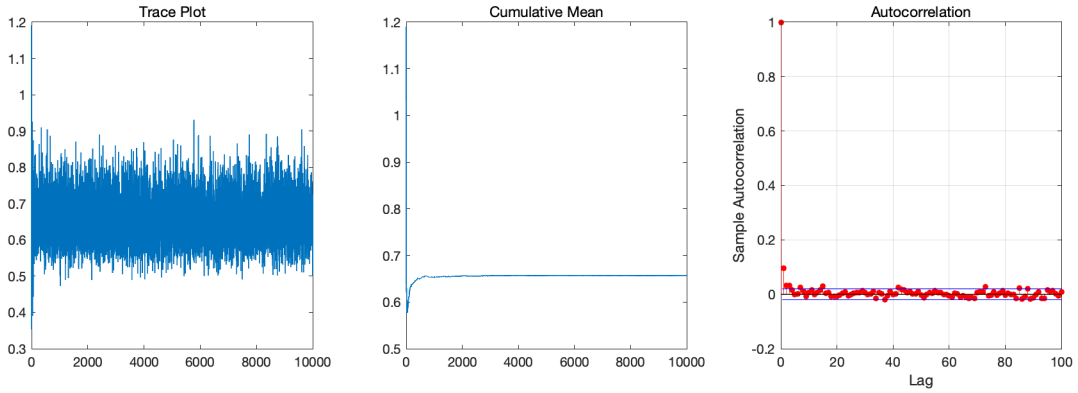
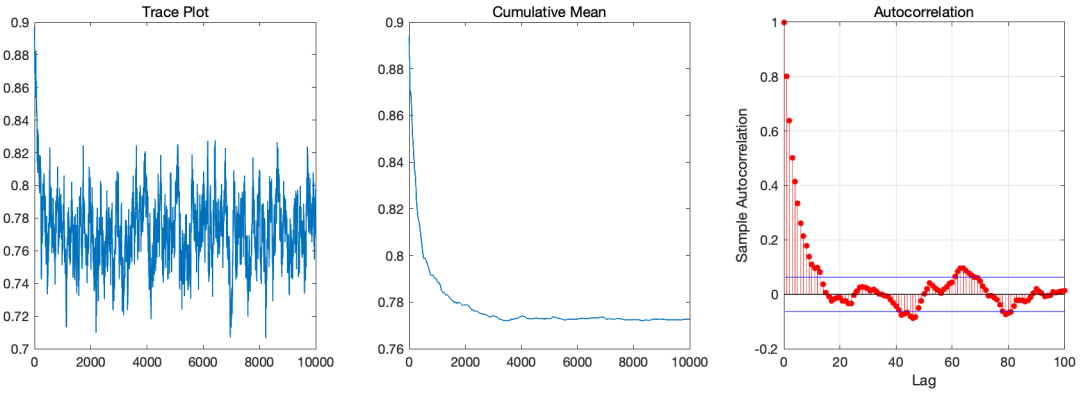


Figure A-3: Convergence Diagnostics, ρ



E Additional Simulation Results

E.1 Main MC Simulation: Larger variance

In this section, we present the additional simulation results of DGP1, DGP2 DGP3 and DGP4 with larger variance with $\sigma^2 = 1.2^2$. The anther settings remain the same: $N = 100$, $T = 10$ and the true number of groups is $K^0 = 4$.

Table A-2 and A-3 shows the estimate and forecast comparison among alternative predictors. For DGP1 and DGP2, the results are similar to those of smaller variance in the main text: the best models are Ti-Homo and Ti-Hetero, respectively. However, in the DGP3, the Tv-Homo and Tv-Hetero estimators, which are expected to stand out since they correctly model the time effects, don't offer the best performance. The potential reason is that the estimation becomes substantially difficult in the presence of both time-varying random effects and much noisier error terms, making it hard to accurately determine group structure. Regarding the DGP4, Ti-Homo and Ti-Hetero deliver outstanding performance relative to other alternative estimators. As there is no group structure in this DGP, the Flat estimator should be the best, which indeed generate accurate estimates and forecast but Ti-Homo and Ti-Hetero can still stand out. This is mainly because Ti-Homo and Ti-Hetero optimally partition similar units into several groups, which averages out the noisy error terms and, hence, scores great performance. These exciting results also suggest that, though no group structure in the sample, our BGRE estimators have the edge over other estimators who either pool all information (Pooled) or treat each unit separately (Flat).

Table A-2: Monte Carlo Experiment: Point Estimates, Larger σ^2

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.0336	0.0227	0.0169	0.0658	0.70	-0.1446	3.08
	Ti-Hetero	0.0342	0.0241	0.0165	0.0643	0.63	-0.1548	2.96
	Tv-Homo	0.2386	0.2363	0.0181	0.0686	0.05	-1.5126	1.30
	Tv-Hetero	0.2436	0.2419	0.0155	0.0597	0.05	-1.5488	1.41
	Pooled	0.2217	0.2215	0.0088	0.0342	0	-1.4113	1
	Flat	0.0665	-0.0642	0.0164	0.0639	0.03	0.4058	100
	Param	0.2493	0.2205	0.1119	0.4429	0.51	-1.4000	1
DGP 2 (Grp Ti He.)	Ti-Homo	0.0146	0.0058	0.0104	0.0404	0.94	-0.0405	4.19
	Ti-Hetero	0.0084	0.0033	0.0062	0.0242	0.91	-0.0231	3.96
	Tv-Homo	0.2231	0.2215	0.0193	0.0744	0.02	-1.4385	10.59
	Tv-Hetero	0.1639	0.1611	0.0218	0.0827	0.09	-1.0595	3.08
	Pooled	0.2495	0.2494	0.0063	0.0245	0	-1.6040	1
	Flat	0.0262	-0.0237	0.0105	0.0409	0.32	0.1506	100
	Param	0.2743	0.2480	0.1135	0.4488	0.29	-1.5926	1
DGP 3 (Grp Tv Ho.)	Ti-Homo	0.2074	0.2064	0.0172	0.0670	0.02	0.0052	1.03
	Ti-Hetero	0.2063	0.2054	0.0169	0.0659	0.01	0.0050	1.03
	Tv-Homo	0.2102	0.2095	0.0168	0.0653	0	0.0052	1.03
	Tv-Hetero	0.2113	0.2106	0.0169	0.0658	0	0.0051	1.17
	Pooled	0.2102	0.2096	0.0166	0.0646	0	0.0050	1
	Flat	0.1870	-0.1849	0.0277	0.1080	0	-0.0047	100
	Param	0.2162	0.2098	0.0505	0.2012	0.01	0.0168	1
DGP 4 (Std Ti Ho.)	Ti-Homo	0.0145	0.0063	0.0097	0.0376	0.88	-0.0439	4.45
	Ti-Hetero	0.0148	0.0069	0.0098	0.0384	0.89	-0.0481	4.23
	Tv-Homo	0.3096	0.3093	0.0137	0.0530	0	-1.9922	1.99
	Tv-Hetero	0.3108	0.3105	0.0135	0.0522	0	-1.9998	2.04
	Pooled	0.2529	0.2528	0.0062	0.0240	0	-1.6281	1
	Flat	0.0256	-0.0230	0.0100	0.0388	0.37	0.1481	100
	Param	0.2775	0.2515	0.1161	0.4593	0.26	-1.6167	1

Table A-3: Monte Carlo Experiment: Forecast, Larger σ^2

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
DGP 1 (Grp Ti Ho.)	Ti-Homo	1.2281	0.0012	1.2210	4.8430	0.95	-1.6271	0.6939
	Ti-Hetero	1.2297	0.0041	1.2224	4.8338	0.95	-1.6305	0.6952
	Tv-Homo	1.2863	0.0180	1.2724	5.1561	0.95	-1.6746	0.7266
	Tv-Hetero	1.2902	0.0193	1.2760	5.1626	0.95	-1.6776	0.7285
	Pooled	1.3385	0.3476	1.2832	5.4955	0.96	-1.7161	0.7567
	Flat	1.2637	-0.1433	1.2494	4.8660	0.95	-1.6564	0.7147
	Param	1.3415	0.3499	1.2856	8.2549	1	-1.8465	0.7971
DGP 2 (Grp Ti He.)	Ti-Homo	0.6989	0.0055	0.6949	2.7235	0.93	-1.0583	0.3819
	Ti-Hetero	0.6926	0.0024	0.6884	2.5818	0.96	-0.8640	0.3600
	Tv-Homo	0.8502	0.0101	0.8425	2.5678	0.89	-1.2358	0.4496
	Tv-Hetero	0.7313	0.0080	0.7230	2.7180	0.95	-0.9323	0.3812
	Pooled	0.8699	0.4283	0.7507	3.7531	0.95	-1.2926	0.4832
	Flat	0.7172	-0.0414	0.7120	2.8231	0.93	-1.0899	0.3935
	Param	0.8751	0.4284	0.7567	7.1947	1	-1.5801	0.5658
DGP 3 (Grp Tv Ho.)	Ti-Homo	1.2684	-0.0325	1.2614	5.0355	0.95	-1.6600	0.7162
	Ti-Hetero	1.2686	-0.0326	1.2616	5.0343	0.95	-1.6603	0.7164
	Tv-Homo	1.2750	0.0022	1.2616	5.0568	0.95	-1.6653	0.7200
	Tv-Hetero	1.2749	0.0026	1.2614	5.0580	0.95	-1.6653	0.7197
	Pooled	1.2678	-0.0328	1.2608	5.0383	0.95	-1.6596	0.7158
	Flat	1.2781	-0.0377	1.2711	4.7935	0.94	-1.6690	0.7235
	Param	1.2835	-0.0214	1.2608	9.2910	1	-1.8663	0.7878
DGP 4 (Std Ti Ho.)	Ti-Homo	0.6435	-0.0112	0.6400	2.5472	0.95	-0.9807	0.3636
	Ti-Hetero	0.6436	-0.0103	0.6401	2.6020	0.95	-0.9837	0.3639
	Tv-Homo	0.7069	0.0247	0.6988	2.8327	0.95	-1.0760	0.3993
	Tv-Hetero	0.7072	0.0247	0.6991	2.8513	0.95	-1.0771	0.3994
	Pooled	0.8200	0.4198	0.7011	3.5977	0.97	-1.2356	0.4660
	Flat	0.6720	-0.0580	0.6663	2.6339	0.95	-1.0246	0.3800
	Param	0.8243	0.4203	0.7059	7.0829	1	-1.5529	0.5467

E.2 Main MC Simulation: Shorter Time Periods

Here, we show the additional simulation results of DGP1, DGP2 and DGP4 with small period, i.e, $T = 5$. The rest settings remain the same: $N = 100$, $\sigma^2 = 0.8^2$ and the true number of groups is $K^0 = 4$.

Table A-4: Monte Carlo Experiment: Point Estimates, Smaller T

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.0379	0.0276	0.0198	0.0766	0.67	-0.1414	3.16
	Ti-Hetero	0.0387	0.029	0.0198	0.0772	0.66	-0.1488	3.09
	Tv-Homo	0.3577	0.3566	0.0199	0.0777	0.02	-1.8223	1.37
	Tv-Hetero	0.3654	0.3646	0.0195	0.0760	0.01	-1.8584	1.63
	Pooled	0.2789	0.2785	0.0120	0.0467	0.01	-1.4050	1
	Flat	0.0591	-0.0554	0.0190	0.0738	0.16	0.2782	100
	Param	0.3146	0.2783	0.1385	0.5503	0.43	-1.4006	1
DGP 2 (Grp Ti He.)	Ti-Homo	0.0523	0.0380	0.0245	0.0952	0.66	-0.189	2.91
	Ti-Hetero	0.0230	0.0126	0.0147	0.0574	0.86	-0.0639	3.40
	Tv-Homo	0.3099	0.3052	0.0288	0.1115	0.05	-1.5558	5.65
	Tv-Hetero	0.2099	0.2018	0.0377	0.1455	0.17	-1.0403	2.87
	Pooled	0.2613	0.2609	0.0138	0.0537	0	-1.3253	1
	Flat	0.0833	-0.0798	0.0232	0.0902	0.04	0.4030	100
	Param	0.2965	0.2602	0.1365	0.5417	0.51	-1.3210	1
DGP 4 (Std Ti Ho.)	Ti-Homo	0.2345	0.2329	0.0269	0.1051	0	0.0018	1
	Ti-Hetero	0.2344	0.2328	0.0269	0.105	0	0.0017	1.01
	Tv-Homo	0.2357	0.2342	0.0270	0.1051	0	0.0016	1
	Tv-Hetero	0.2363	0.2347	0.0272	0.1062	0	0.0018	1.24
	Pooled	0.2344	0.2328	0.0270	0.1051	0	0.0017	1
	Flat	0.3601	-0.3571	0.0459	0.1790	0	-0.0029	100
	Param	0.2516	0.2323	0.0941	0.3756	0.19	0.0066	1

Table A-5: Monte Carlo Experiment: Forecast, Smaller T

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.8491	0.0553	0.8407	3.3362	0.95	-1.2561	0.4798
	Ti-Hetero	0.8505	0.0585	0.8418	3.3743	0.95	-1.2605	0.4810
	Tv-Homo	0.9310	0.1221	0.9114	3.6884	0.95	-1.3514	0.5275
	Tv-Hetero	0.9352	0.1248	0.9156	3.6947	0.95	-1.3549	0.5296
	Pooled	1.0817	0.6143	0.8796	4.2897	0.96	-1.4992	0.6152
	Flat	0.8790	-0.1242	0.8653	3.4044	0.95	-1.2946	0.4980
	Param	1.0852	0.6144	0.8836	7.5513	1	-1.6938	0.6666
DGP 2 (Grp Ti He.)	Ti-Homo	1.1264	0.0956	1.1131	4.2430	0.93	-1.5308	0.6138
	Ti-Hetero	1.0782	0.0397	1.0698	3.9245	0.95	-1.3183	0.5621
	Tv-Homo	1.2929	0.1517	1.2713	4.0156	0.89	-1.6690	0.6880
	Tv-Hetero	1.1323	0.1113	1.1129	4.1041	0.94	-1.3952	0.5948
	Pooled	1.2862	0.6034	1.1257	5.0186	0.93	-1.6744	0.7084
	Flat	1.1320	-0.166	1.1129	4.3232	0.93	-1.5477	0.6210
	Param	1.2889	0.6022	1.1294	7.9859	0.99	-1.8040	0.7577
DGP 4 (Std Ti Ho.)	Ti-Homo	0.8550	-0.0081	0.8500	3.4368	0.96	-1.2665	0.4841
	Ti-Hetero	0.8550	-0.0082	0.8500	3.4362	0.96	-1.2668	0.4841
	Tv-Homo	0.8607	-0.026	0.8501	3.4509	0.95	-1.2733	0.4875
	Tv-Hetero	0.8620	-0.0259	0.8514	3.4513	0.95	-1.2745	0.4883
	Pooled	0.8550	-0.0082	0.8500	3.4362	0.95	-1.2665	0.4841
	Flat	0.8907	0.0036	0.886	3.2123	0.92	-1.3147	0.5052
	Param	0.8671	-0.0032	0.8498	8.5456	1	-1.6545	0.5970

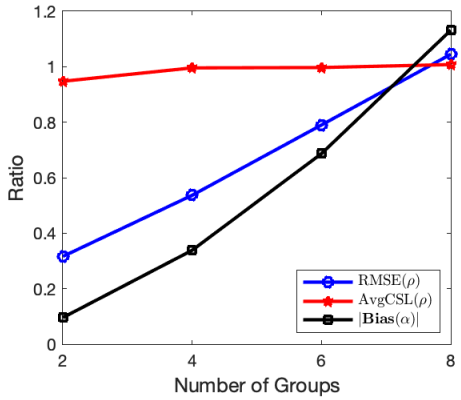
E.3 Main MC Simulation: Different K^0

In this section, we present the simulation results of DGP1, DGP2 and DGP3 with different number of groups. In particular, we consider $K^0 \in \{2, 4, 6, 8\}$. The rest settings remain the same: $N = 100$, $T = 10$, and $\sigma^2 = 0.8^2$.

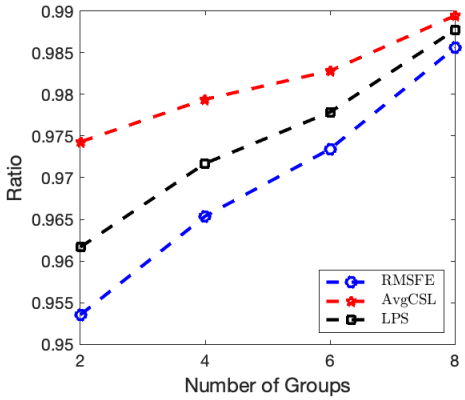
Figure A-4 presents the relative performance of the BGRE estimators against the flat estimator under different K^0. In particular, we show the results of the correctly specified estimators for each DGP, i.e., Ti-Homo estimator for DGP 1, Ti-Hetero estimator for DGP 2, and Tv-Homo estimator for DGP 3. For the DGP 1, the accuracy of the estimates and the predictive power of the BGRE estimator gradually vanish as K^0 increases. At $K^0 = 8$, the BGRE estimator still marginally dominates the flat estimator in all aspects besides the bias of α. Moving to the DGP 2, the BGRE estimator offers better performance than the flat estimator for all K^0. This suggests that the BGRE estimator successfully captures heterogeneity in variance and sophisticated group patterns, even with eight different clusters. Regarding the DPG 3, where we introduce time variation in α, the BGRE estimator outperforms the benchmark model in terms of forecasting. Moreover, we see RMSFE, the average length of the credible set, and LPS are all trending down. This suggests the more remarkable improvement in the predictive power of the BGRE estimator, as the true model becomes more sophisticated. It is also noteworthy that, while the average length of the credible set for ρ is relatively large, the BGRE estimator generates a much lower RMSE of ρ and absolute bias of α than the flat estimator.

Figure A-4: Monte Carlo Experiment: BGRE Estimator, Different K^0

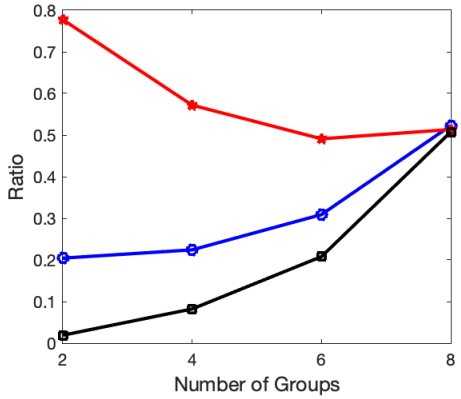
(a) DGP1, Estimates



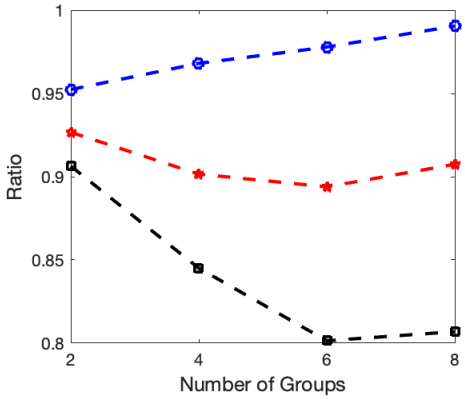
(b) DGP1, Forecasts



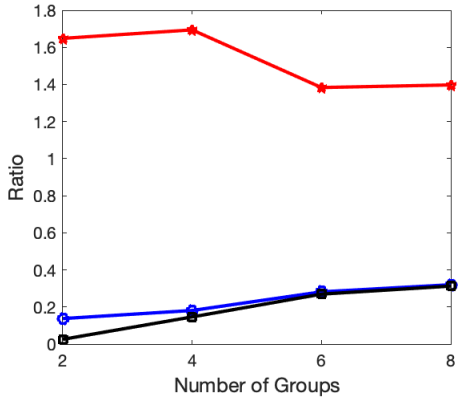
(c) DGP2, Estimates



(d) DGP2, Forecasts



(e) DGP3, Estimates



(f) DGP3, Forecasts

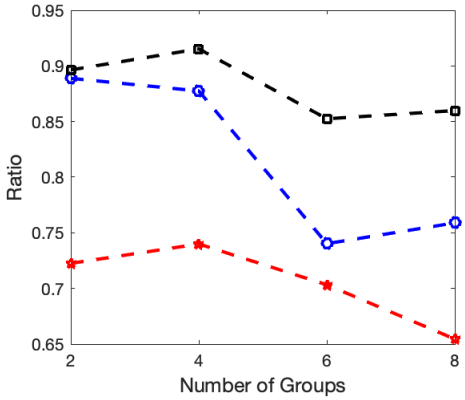


Table A-6: Monte Carlo Experiment: Point Estimates, Different K^0, DGP1 (Grp Ti Ho.)

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
	Ti-Homo	0.0257	0.0073	0.0172	0.0667	0.89	-0.0275	1.78
	Ti-Hetero	0.0265	0.0085	0.017	0.0658	0.89	-0.0296	1.77
	Tv-Homo	0.1254	0.1160	0.0200	0.0776	0.39	-0.4375	1.30
$K^0 = 2$	Tv-Hetero	0.1337	0.1240	0.0215	0.0815	0.39	-0.4593	1.33
	Pooled	0.1512	0.1481	0.0104	0.0404	0.18	-0.5410	1
	Flat	0.0813	-0.0791	0.0182	0.0704	0.01	0.2873	100
	Param	0.2616	0.1482	0.1886	0.7409	0.93	-0.5365	1
	Ti-Homo	0.0198	0.0111	0.0120	0.0465	0.77	-0.0735	3.63
	Ti-Hetero	0.0200	0.0115	0.0120	0.0466	0.81	-0.0758	3.58
	Tv-Homo	0.2388	0.2371	0.0194	0.0736	0.06	-1.5154	1.87
$K^0 = 4$	Tv-Hetero	0.2401	0.2385	0.0177	0.0680	0.06	-1.5275	1.93
	Pooled	0.2449	0.2448	0.0069	0.0267	0	-1.5591	1
	Flat	0.0369	-0.0345	0.0121	0.0467	0.20	0.2173	100
	Param	0.2695	0.2438	0.1100	0.4322	0.30	-1.5545	1
	Ti-Homo	0.0158	0.0114	0.0086	0.0331	0.70	-0.1094	5.21
	Ti-Hetero	0.0163	0.0120	0.0086	0.0334	0.73	-0.1152	5.12
	Tv-Homo	0.3086	0.3083	0.0129	0.0497	0	-2.8593	2.18
$K^0 = 6$	Tv-Hetero	0.3078	0.3075	0.013	0.0502	0	-2.8522	2.22
	Pooled	0.2708	0.2708	0.0055	0.0214	0	-2.5079	1
	Flat	0.0200	-0.0172	0.0086	0.0332	0.48	0.1593	100
	Param	0.2809	0.2703	0.0749	0.2944	0.02	-2.5033	1
	Ti-Homo	0.0140	0.0114	0.0067	0.0261	0.63	-0.1421	6.37
	Ti-Hetero	0.0154	0.0128	0.0070	0.0269	0.60	-0.1592	6.19
	Tv-Homo	0.3236	0.3234	0.0105	0.0406	0	-3.8671	2.38
$K^0 = 8$	Tv-Hetero	0.3247	0.3246	0.0103	0.0396	0	-3.8807	2.41
	Pooled	0.2797	0.2796	0.0050	0.0192	0	-3.3408	1
	Flat	0.0134	-0.0105	0.0067	0.0259	0.64	0.1256	100
	Param	0.2854	0.2794	0.0576	0.2262	0.02	-3.3362	1

Table A-7: Monte Carlo Experiment: Forecast, Different K^0, DGP1 (Grp Ti Ho.)

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
$K^0 = 2$	Ti-Homo	0.8043	-0.0166	0.7997	3.1591	0.95	-1.2040	0.4546
	Ti-Hetero	0.8045	-0.0158	0.8000	3.1677	0.95	-1.2049	0.4549
	Tv-Homo	0.8281	0.002	0.8194	3.2726	0.95	-1.2334	0.4681
	Tv-Hetero	0.8305	0.0025	0.8219	3.2871	0.95	-1.2364	0.4694
	Pooled	0.8565	0.1246	0.8364	3.4496	0.95	-1.2673	0.4844
	Flat	0.8435	-0.1001	0.8335	3.2424	0.95	-1.2520	0.4771
	Param	0.8660	0.1286	0.8426	7.0083	1	-1.5565	0.5588
$K^0 = 4$	Ti-Homo	0.8114	-0.0075	0.8072	3.1933	0.95	-1.2131	0.4586
	Ti-Hetero	0.8124	-0.0069	0.8081	3.2155	0.95	-1.2159	0.4594
	Tv-Homo	0.8638	0.0174	0.8545	3.4290	0.95	-1.2755	0.4877
	Tv-Hetero	0.8648	0.0177	0.8555	3.4423	0.95	-1.2774	0.4884
	Pooled	0.9619	0.3975	0.8685	4.0958	0.97	-1.3905	0.5456
	Flat	0.8405	-0.0811	0.8325	3.2604	0.95	-1.2484	0.4753
	Param	0.9708	0.3987	0.8770	7.3543	1	-1.6365	0.6132
$K^0 = 6$	Ti-Homo	0.8172	-0.0009	0.8130	3.2113	0.95	-1.2196	0.4619
	Ti-Hetero	0.8181	0.0006	0.8138	3.2487	0.95	-1.2222	0.4624
	Tv-Homo	0.8882	0.0349	0.8779	3.5329	0.95	-1.3042	0.5016
	Tv-Hetero	0.8885	0.0347	0.8782	3.5424	0.95	-1.3055	0.502
	Pooled	1.0947	0.6514	0.8741	4.9284	0.98	-1.5306	0.6253
	Flat	0.8395	-0.0657	0.8329	3.2674	0.95	-1.2472	0.4747
	Param	1.1030	0.6533	0.8829	7.8565	1	-1.7339	0.6855
$K^0 = 8$	Ti-Homo	0.8271	0.0064	0.8227	3.2357	0.95	-1.2315	0.4676
	Ti-Hetero	0.8314	0.0110	0.8271	3.2960	0.95	-1.2386	0.4703
	Tv-Homo	0.8983	0.0460	0.8874	3.5889	0.95	-1.3155	0.5076
	Tv-Hetero	0.8990	0.0459	0.8881	3.6023	0.95	-1.3171	0.5080
	Pooled	1.2428	0.8744	0.8783	5.8013	0.99	-1.6675	0.7153
	Flat	0.8392	-0.0567	0.8332	3.2702	0.95	-1.2468	0.4745
	Param	1.2505	0.8765	0.8872	8.4358	1	-1.8330	0.7674

Table A-8: Monte Carlo Experiment: Point Estimates, Different K^0, DGP2 (Grp Ti He.)

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
$K^0 = 2$	Ti-Homo	0.0313	0.0021	0.0209	0.0814	0.88	-0.0132	1.78
	Ti-Hetero	0.0235	0.0014	0.0171	0.0667	0.92	-0.0080	2.01
	Tv-Homo	0.0501	0.0177	0.0300	0.1160	0.84	-0.0776	8.44
	Tv-Hetero	0.0351	0.0052	0.0249	0.0961	0.88	-0.0242	2.34
	Pooled	0.1214	0.1152	0.0128	0.0498	0.21	-0.4354	1
	Flat	0.1150	-0.1127	0.0221	0.0858	0	0.4181	100
	Param	0.2265	0.1143	0.1716	0.6747	0.94	-0.4306	1
$K^0 = 4$	Ti-Homo	0.0221	0.0091	0.0151	0.0586	0.87	-0.0639	3.81
	Ti-Hetero	0.0111	0.0034	0.0083	0.0322	0.93	-0.0246	4.01
	Tv-Homo	0.1927	0.1896	0.0215	0.0822	0.11	-1.2371	11.51
	Tv-Hetero	0.0950	0.0878	0.0244	0.0926	0.33	-0.5819	3.42
	Pooled	0.2319	0.2317	0.0079	0.0308	0	-1.4910	1
	Flat	0.0495	-0.0471	0.0145	0.0563	0.07	0.2993	100
	Param	0.2566	0.2310	0.1073	0.4215	0.3	-1.4863	1
$K^0 = 6$	Ti-Homo	0.0159	0.0098	0.0102	0.0398	0.85	-0.0984	5.57
	Ti-Hetero	0.0076	0.0041	0.0049	0.0191	0.89	-0.0423	5.20
	Tv-Homo	0.2855	0.2849	0.0154	0.0587	0	-2.6521	14.68
	Tv-Hetero	0.2374	0.2365	0.0178	0.0660	0	-2.2017	3.51
	Pooled	0.2644	0.2643	0.0060	0.0233	0	-2.4582	1
	Flat	0.0246	-0.0222	0.0100	0.0389	0.38	0.2034	100
	Param	0.2749	0.2641	0.0747	0.2937	0.03	-2.4535	1
$K^0 = 8$	Ti-Homo	0.0155	0.0122	0.0078	0.0302	0.68	-0.1545	6.77
	Ti-Hetero	0.0080	0.006	0.004	0.0153	0.78	-0.0774	6.24
	Tv-Homo	0.3215	0.3213	0.0119	0.0452	0	-3.8562	15.33
	Tv-Hetero	0.2990	0.2986	0.0128	0.0471	0	-3.5879	3.60
	Pooled	0.2772	0.2771	0.0052	0.0202	0	-3.3266	1
	Flat	0.0153	-0.0129	0.0077	0.0298	0.63	0.1525	100
	Param	0.2830	0.2769	0.0574	0.2254	0.02	-3.3220	1

Table A-9: Monte Carlo Experiment: Forecast, Different K^0, DGP2 (Grp Ti He.)

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
$K^0 = 2$	Ti-Homo	1.2014	0.0055	1.1945	4.6638	0.94	-1.6015	0.665
	Ti-Hetero	1.1853	0.0042	1.1784	4.4307	0.95	-1.4876	0.6381
	Tv-Homo	1.3162	0.0050	1.3040	4.3151	0.91	-1.6805	0.7173
	Tv-Hetero	1.2017	0.0050	1.1883	4.4625	0.95	-1.4999	0.6464
	Pooled	1.2402	0.1183	1.2255	4.8993	0.94	-1.6374	0.6879
	Flat	1.2446	-0.1098	1.2331	4.7811	0.93	-1.6413	0.6909
	Param	1.2481	0.1223	1.2325	7.8794	0.99	-1.7799	0.7384
$K^0 = 4$	Ti-Homo	1.0599	0.0110	1.0537	4.0879	0.93	-1.4743	0.5798
	Ti-Hetero	1.0427	0.0020	1.0364	3.7691	0.95	-1.2646	0.5416
	Tv-Homo	1.2697	0.0075	1.2586	3.7160	0.88	-1.6528	0.6720
	Tv-Hetero	1.0809	0.0022	1.0691	3.8909	0.94	-1.3088	0.5624
	Pooled	1.1839	0.3976	1.1072	4.8767	0.94	-1.5952	0.6518
	Flat	1.0771	-0.0816	1.068	4.1807	0.93	-1.4965	0.5906
	Param	1.1914	0.3991	1.1148	7.8182	0.99	-1.7574	0.7111
$K^0 = 6$	Ti-Homo	0.9914	0.0155	0.9860	3.8175	0.93	-1.4066	0.5383
	Ti-Hetero	0.9813	0.0038	0.9759	3.4921	0.96	-1.1432	0.4979
	Tv-Homo	1.2778	0.0185	1.2680	3.3939	0.86	-1.7016	0.6664
	Tv-Hetero	1.0441	0.0185	1.0324	3.7406	0.95	-1.2183	0.5310
	Pooled	1.2388	0.6548	1.0457	5.3733	0.95	-1.6475	0.6875
	Flat	1.0035	-0.0574	0.9966	3.9066	0.93	-1.4264	0.5462
	Param	1.2472	0.6565	1.0545	8.1351	0.99	-1.8042	0.7455
$K^0 = 8$	Ti-Homo	0.9721	0.0279	0.9659	3.7423	0.93	-1.3865	0.5286
	Ti-Hetero	0.9608	0.0115	0.9552	3.4452	0.96	-1.1238	0.4876
	Tv-Homo	1.2746	0.0282	1.2644	3.3205	0.85	-1.7257	0.6624
	Tv-Hetero	1.0348	0.0308	1.0223	3.7104	0.95	-1.1883	0.5218
	Pooled	1.3526	0.8862	1.0174	6.1307	0.96	-1.7452	0.7616
	Flat	0.9700	-0.0444	0.9636	3.7966	0.93	-1.3927	0.5271
	Param	1.3601	0.8870	1.0264	8.6643	0.99	-1.8832	0.8133

Table A-10: Monte Carlo Experiment: Point Estimates, Different K^0, DGP3 (Grp Tv Ho.)

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
	Ti-Homo	0.2173	0.2169	0.0131	0.0509	0	-2.0461	1
	Ti-Hetero	0.2112	0.2108	0.0137	0.0529	0	-2.0054	1.69
	Tv-Homo	0.0321	0.0082	0.0234	0.091	0.95	-0.0548	2.01
$K^0 = 2$	Tv-Hetero	0.0325	0.0079	0.0233	0.0906	0.93	-0.0532	2.07
	Pooled	0.2171	0.2168	0.0097	0.0377	0	-2.0455	1
	Flat	0.2337	0.2333	0.0142	0.0552	0	-2.1542	100
	Param	0.2472	0.2163	0.1176	0.4618	0.48	-2.0408	1
	Ti-Homo	0.2741	0.2739	0.0119	0.0461	0	-2.2118	2.03
	Ti-Hetero	0.2750	0.2747	0.0120	0.0467	0	-2.2170	2.48
	Tv-Homo	0.0587	0.0536	0.0217	0.0837	0.33	-0.3749	3.94
$K^0 = 4$	Tv-Hetero	0.0605	0.0555	0.0219	0.0850	0.29	-0.3882	3.88
	Pooled	0.1929	0.1927	0.0081	0.0313	0	-1.6492	1
	Flat	0.3242	0.3239	0.0127	0.0494	0	-2.5618	100
	Param	0.2167	0.1922	0.0983	0.3865	0.39	-1.6445	1
	Ti-Homo	0.2897	0.2892	0.0158	0.0604	0	-2.1634	1.89
	Ti-Hetero	0.2984	0.2981	0.0137	0.0531	0	-2.2277	2.08
	Tv-Homo	0.0995	0.0971	0.0212	0.0819	0.01	-0.7081	4.19
$K^0 = 6$	Tv-Hetero	0.1003	0.0979	0.0211	0.0817	0	-0.7140	4.19
	Pooled	0.2012	0.2010	0.0078	0.0302	0	-1.5241	1
	Flat	0.3534	0.3531	0.0152	0.0592	0	-2.6306	100
	Param	0.2208	0.2006	0.0907	0.3564	0.2	-1.5195	1
	Ti-Homo	0.2666	0.2660	0.0171	0.0664	0	-2.0004	1.93
	Ti-Hetero	0.2834	0.2829	0.0157	0.0609	0	-2.1269	2.02
	Tv-Homo	0.1049	0.1023	0.0224	0.0865	0	-0.7698	5.96
$K^0 = 8$	Tv-Hetero	0.1091	0.1066	0.0225	0.0858	0.01	-0.8016	5.78
	Pooled	0.2021	0.2019	0.0088	0.0342	0	-1.5209	1
	Flat	0.3279	0.3276	0.0159	0.0619	0	-2.4628	100
	Param	0.2228	0.2015	0.0935	0.3673	0.28	-1.5162	1

Table A-11: Monte Carlo Experiment: Forecast, Different K^0, DGP3 (Grp Tv Ho.)

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
$K^0 = 2$	Ti-Homo	1.0475	0.1611	1.0286	4.8735	0.98	-1.4966	0.5988
	Ti-Hetero	1.0294	0.1392	1.0138	4.8565	0.98	-1.4885	0.5900
	Tv-Homo	0.9507	0.0186	0.9414	3.7414	0.95	-1.3715	0.5381
	Tv-Hetero	0.9515	0.0187	0.9422	3.7428	0.95	-1.3733	0.5387
	Pooled	1.0476	0.1609	1.0287	4.8783	0.98	-1.4967	0.5991
	Flat	1.0696	0.2195	1.0402	5.1782	0.98	-1.5300	0.614
	Param	1.0531	0.1601	1.031	7.6736	1	-1.7061	0.6607
$K^0 = 4$	Ti-Homo	1.3729	0.3097	1.3328	5.1851	0.94	-1.7468	0.7851
	Ti-Hetero	1.3553	0.3124	1.3137	5.1330	0.94	-1.7239	0.7724
	Tv-Homo	1.0879	0.0254	1.0792	3.9212	0.93	-1.5055	0.6150
	Tv-Hetero	1.0936	0.0257	1.0850	3.9377	0.93	-1.5109	0.6182
	Pooled	1.8946	0.1243	1.8874	5.5632	0.84	-2.1561	1.1083
	Flat	1.2398	0.4209	1.1606	5.2996	0.97	-1.6448	0.7048
	Param	1.8974	0.1246	1.8902	8.2086	0.98	-2.0727	1.0854
$K^0 = 6$	Ti-Homo	1.9244	0.4059	1.8756	5.6825	0.85	-2.1845	1.1369
	Ti-Hetero	1.8699	0.4200	1.8176	5.6099	0.87	-2.1166	1.0933
	Tv-Homo	1.2273	0.0505	1.2193	4.1107	0.90	-1.6523	0.7007
	Tv-Hetero	1.2294	0.0514	1.2213	4.1175	0.90	-1.6512	0.7014
	Pooled	2.5171	0.2716	2.5003	5.9269	0.71	-2.7056	1.5662
	Flat	1.6580	0.4999	1.5765	5.8450	0.92	-1.9382	0.9455
	Param	2.5192	0.2731	2.502	8.4685	0.93	-2.3940	1.4845
$K^0 = 8$	Ti-Homo	1.7915	0.1073	1.7850	6.1019	0.90	-2.0260	1.0236
	Ti-Hetero	1.7466	0.1157	1.7395	6.0285	0.91	-2.0106	0.9988
	Tv-Homo	1.1879	0.0152	1.1804	4.0868	0.91	-1.6074	0.6746
	Tv-Hetero	1.2079	0.0166	1.2005	4.1259	0.91	-1.6221	0.6861
	Pooled	2.2360	0.0755	2.2323	6.3016	0.83	-2.3547	1.3270
	Flat	1.5653	0.1355	1.5556	6.2445	0.95	-1.8694	0.8838
	Param	2.2393	0.0772	2.2350	8.7506	0.96	-2.2403	1.2913

E.4 Two-Step GRE estimator

Table A-12: Monte Carlo Experiment: Two-Step GRE with Kmean, Point Estimates

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.0627	0.0599	0.0114	0.0444	0.25	-0.3836	2.2
	Ti-Hetero	0.0611	0.0583	0.0116	0.0452	0.27	-0.3734	2.2
	Tv-Homo	0.1567	0.1544	0.0186	0.0721	0.11	-0.9927	2.2
	Tv-Hetero	0.1560	0.1537	0.0189	0.0738	0.10	-0.9887	2.2
DGP 2 (Grp Ti He.)	Ti-Homo	0.0550	0.0513	0.0133	0.0517	0.27	-0.3388	2.2
	Ti-Hetero	0.0456	0.0429	0.0099	0.0386	0.27	-0.2822	2.2
	Tv-Homo	0.1196	0.1143	0.0203	0.0789	0.21	-0.7572	2.2
	Tv-Hetero	0.1458	0.1427	0.0196	0.0764	0.15	-0.9400	2.2
DGP 3 (Grp Tv Ho.)	Ti-Homo	0.2863	0.2861	0.0114	0.0446	0.00	-2.2885	2.26
	Ti-Hetero	0.2807	0.2804	0.0115	0.0448	0.00	-2.2489	2.26
	Tv-Homo	0.1196	0.1171	0.0166	0.0648	0.08	-0.8168	2.26
	Tv-Hetero	0.1172	0.1144	0.017	0.0661	0.08	-0.7982	2.26
DGP 4 (Std Ti Ho.)	Ti-Homo	0.0832	0.0796	0.0210	0.0819	0.10	0.0014	2.03
	Ti-Hetero	0.0833	0.0797	0.0210	0.0819	0.08	0.0015	2.03
	Tv-Homo	0.0942	0.0908	0.0218	0.0851	0.05	0.0018	2.03
	Tv-Hetero	0.0943	0.0908	0.0218	0.0851	0.06	0.0019	2.03

Table A-13: Monte Carlo Experiment: Two-Step GRE with Kmean, Forecast

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.8607	0.0777	0.8502	3.3899	0.95	-1.2715	0.4866
	Ti-Hetero	0.8610	0.0751	0.8507	3.3949	0.95	-1.2714	0.4867
	Tv-Homo	0.8457	0.0084	0.8369	3.3324	0.95	-1.2545	0.4775
	Tv-Hetero	0.8457	0.0086	0.8369	3.3393	0.95	-1.2551	0.4774
DGP 2 (Grp Ti He.)	Ti-Homo	1.0969	0.0856	1.0860	4.2250	0.93	-1.5155	0.6031
	Ti-Hetero	1.0993	0.0723	1.0896	4.0301	0.94	-1.4115	0.5857
	Tv-Homo	1.0886	0.0054	1.0764	4.2115	0.93	-1.5073	0.5971
	Tv-Hetero	1.0900	0.0079	1.0775	3.9970	0.94	-1.3806	0.5758
DGP 3 (Grp Tv Ho.)	Ti-Homo	1.3326	0.3341	1.2852	5.0617	0.95	-1.7082	0.7589
	Ti-Hetero	1.3393	0.3219	1.2952	5.0363	0.95	-1.7071	0.7616
	Tv-Homo	1.3031	0.0273	1.2961	4.2823	0.90	-1.7175	0.7477
	Tv-Hetero	1.3052	0.0263	1.2982	4.2703	0.90	-1.7150	0.7483
DGP 4 (Std Ti Ho.)	Ti-Homo	0.8532	-0.0232	0.8488	3.2353	0.94	-1.2638	0.4822
	Ti-Hetero	0.8532	-0.0231	0.8488	3.2420	0.94	-1.2641	0.4823
	Tv-Homo	0.8537	-0.0018	0.8458	3.2581	0.94	-1.2640	0.4826
	Tv-Hetero	0.8537	-0.0014	0.8458	3.2654	0.94	-1.2646	0.4826

E.5 Subjective Priors With Knowledge on Groups

Table A-14: Monte Carlo Experiment: Estimates, SGM prior

		$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
		RMSE	Bias	Std	AvgL	Cov	PBias	Avg K
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.0186	0.0074	0.0120	0.0468	0.87	-0.0488	3.64
	Ti-Hetero	0.0216	0.0109	0.0121	0.0471	0.84	-0.0704	3.55
	Tv-Homo	0.1904	0.1871	0.0203	0.0781	0.17	-1.2004	2.08
	Tv-Hetero	0.1933	0.1901	0.0216	0.0809	0.14	-1.2186	2.15
DGP 2 (Grp Ti He.)	Ti-Homo	0.0208	0.0039	0.0149	0.0582	0.88	-0.0292	3.80
	Ti-Hetero	0.0120	0.0031	0.0083	0.0325	0.98	-0.0216	3.96
	Tv-Homo	0.1683	0.1625	0.0240	0.0914	0.19	-1.0689	11.82
	Tv-Hetero	0.0863	0.0779	0.0269	0.1029	0.41	-0.5174	3.48
DGP 3 (Grp Tv Ho.)	Ti-Homo	0.2408	0.2405	0.0114	0.0443	0	-1.9814	1.58
	Ti-Hetero	0.2345	0.2342	0.0112	0.0437	0	-1.9381	1.63
	Tv-Homo	0.0845	0.0805	0.0191	0.0742	0.19	-0.5619	2.91
	Tv-Hetero	0.0735	0.0682	0.0212	0.0815	0.26	-0.4763	3.12

Table A-15: Monte Carlo Experiment: Forecast, SGM prior

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
DGP 1 (Grp Ti Ho.)	Ti-Homo	0.8117	-0.0119	0.8073	3.2020	0.95	-1.2135	0.4587
	Ti-Hetero	0.8139	-0.0062	0.8089	3.2355	0.95	-1.2177	0.4602
	Tv-Homo	0.8518	0.0121	0.8429	3.3874	0.95	-1.2619	0.4807
	Tv-Hetero	0.8538	0.0128	0.8448	3.3970	0.95	-1.2647	0.4819
DGP 2 (Grp Ti He.)	Ti-Homo	1.0590	0.0042	1.0529	4.0942	0.93	-1.4735	0.5793
	Ti-Hetero	1.0429	0.0025	1.0365	3.7887	0.95	-1.2685	0.5422
	Tv-Homo	1.2587	0.0069	1.2478	3.6982	0.88	-1.6399	0.6662
	Tv-Hetero	1.0777	0.002	1.0662	3.8942	0.95	-1.3066	0.5609
DGP 3 (Grp Tv Ho.)	Ti-Homo	1.5915	0.2322	1.5649	5.3633	0.90	-1.9187	0.9205
	Ti-Hetero	1.6250	0.2189	1.5995	5.3665	0.89	-1.9410	0.9397
	Tv-Homo	1.2366	0.0325	1.2284	4.2045	0.91	-1.6532	0.7098
	Tv-Hetero	1.1791	0.0294	1.1707	4.1121	0.92	-1.5976	0.6724

Table A-16: Monte Carlo Experiment: Estimates, SGP prior, DGP3

	$\hat{\rho}$					$\hat{\alpha}_i$	Cluster
	RMSE	Bias	Std	AvgL	Cov	Bias	Avg K
SGP-RE1	0.0396	0.0294	0.0225	0.0871	0.75	-0.2072	4.00
SGP-RE2	0.0463	0.0378	0.0228	0.0892	0.64	-0.2650	3.98
SGP-RE3	0.0760	0.0716	0.0214	0.0834	0.26	-0.4993	3.58
SGP-RE4	0.0821	0.0793	0.0210	0.0818	0.02	-0.5533	4.00
SGP-RE5	0.0691	0.0654	0.0214	0.0834	0.09	-0.4583	6.00
TvHetero	0.0599	0.0549	0.0220	0.0849	0.31	-0.3842	3.85
Flat	0.3243	0.3240	0.0126	0.0493	0	-2.5626	100

Table A-17: Monte Carlo Experiment: Forecast, SGP prior, DGP3

	Point Forecast			Set Forecast		Density Forecast	
	RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
SGP-RE1	1.0266	0.0198	1.0178	3.7766	0.93	-1.4553	0.5823
SGP-RE2	1.0465	0.0214	1.0377	3.8382	0.93	-1.4716	0.5926
SGP-RE3	1.1337	0.0287	1.1250	4.1172	0.93	-1.5542	0.6455
SGP-RE4	1.0682	0.0310	1.0590	4.0305	0.94	-1.4934	0.6063
SGP-RE5	1.0784	0.0293	1.0696	3.9689	0.93	-1.5070	0.6114
Tv-Hetero	1.0952	0.0255	1.0866	3.9531	0.93	-1.6450	0.6192
Flat	1.2400	0.4211	1.1608	5.3189	0.97	-1.6450	0.7048

F Additional Empirical Results

In this section, we present the full result of empirical analysis in which detailed yearly estimate results are listed here.

Table A-18: Empirical Application: Predict Investment Rate, RMSFE

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	0.0917	0.1395	0.2625	0.1166	0.1108
	Ti-Hetero	0.0750	0.1159	0.3550	0.0674	0.0822
	Tv-Homo	0.0927	0.1382	0.2590	0.1165	0.1177
	Tv-Hetero	0.0783	0.1156	0.3686	0.0692	0.0812
	Pooled	0.0926	0.1386	0.2625	0.1160	0.1150
	Flat	0.1034	0.1491	0.2703	0.1328	0.1100
	Param	0.1958	0.2295	0.2466	0.2492	0.2043
Heterogenous Coef.	Ti-Homo	0.1103	0.1006	1.8575	0.1041	0.1144
	Ti-Hetero	0.1104	0.0999	1.8802	0.1028	0.1152
	Tv-Homo	0.1582	0.1729	0.2863	0.1782	0.1070
	Tv-Hetero	0.1097	0.1009	1.8644	0.1062	0.1101
	Flat	0.1649	0.1906	1.7937	0.1833	0.1164

Table A-19: Empirical Application: Predict Investment Rate, Average Number of Groups

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	2	2	1	2	2
	Ti-Hetero	6.62	7.8	6.66	7.79	7.86
	Tv-Homo	1	1	1	1	1
	Tv-Hetero	6.75	6.79	6.8	7.64	6.75
	Pooled	1	1	1	1	1
	Flat	337	337	337	337	337
	Param	1	1	1	1	1
Heterogenous Coef.	Ti-Homo	7.75	6.66	6.78	8.05	7.48
	Ti-Hetero	7.09	6.65	7.23	6.67	6.68
	Tv-Homo	1	1	1	1	1
	Tv-Hetero	6.07	6.11	6.79	6.57	7.46
	Flat	337	337	337	337	337

Table A-20: Empirical Application: Predict Investment Rate, Frequentist Coverage Rate

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	0.9822	0.9703	0.9733	0.9733	0.9614
	Ti-Hetero	0.9525	0.9466	0.9674	0.9585	0.9021
	Tv-Homo	0.9822	0.9644	0.9792	0.9733	0.9525
	Tv-Hetero	0.9466	0.9347	0.9496	0.9466	0.8813
	Pooled	0.9852	0.9644	0.9792	0.9763	0.9555
	Flat	0.9792	0.9792	0.9703	0.9733	0.9703
	Param	1	1	1	1	1
Heterogenous Coef.	Ti-Homo	0.9407	0.9585	0.9555	0.9525	0.8724
	Ti-Hetero	0.9436	0.9466	0.9674	0.9496	0.8724
	Tv-Homo	0.9852	0.9792	0.9792	0.9792	0.9703
	Tv-Hetero	0.9466	0.9466	0.9407	0.9258	0.8338
	Flat	0.9733	0.9822	0.9733	0.9763	0.9733

Table A-21: Empirical Application: Predict Investment Rate, Length of 95% Credible Set

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	0.5737	0.5692	0.5710	0.5912	0.5908
	Ti-Hetero	0.4149	0.3129	0.3223	0.3030	0.4012
	Tv-Homo	0.5665	0.5620	0.5628	0.5851	0.5875
	Tv-Hetero	0.2886	0.2801	0.2809	0.3810	0.3867
	Pooled	0.5759	0.5716	0.5716	0.5946	0.5966
	Flat	0.5709	0.5664	0.5729	0.5976	0.6041
	Param	6.7334	6.7548	6.9387	6.8507	6.8211
Heterogenous Coef.	Ti-Homo	0.2881	0.3008	0.403	0.2925	0.2837
	Ti-Hetero	0.2889	0.3024	0.4033	0.2921	0.2841
	Tv-Homo	0.6368	0.6605	0.7106	0.6414	0.6344
	Tv-Hetero	0.2827	0.2961	0.3978	0.2840	0.2741
	Flat	0.6660	0.7119	0.7948	0.6826	0.6753

Table A-22: Empirical Application: Predict Investment Rate, LPS

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	0.8021	0.5577	-0.3237	0.6752	0.7039
	Ti-Hetero	1.5788	1.5833	1.6552	1.7086	1.3724
	Tv-Homo	0.8044	0.5601	-0.3055	0.6754	0.6671
	Tv-Hetero	1.5618	1.5680	1.6063	1.6896	1.2981
	Pooled	0.7952	0.5556	-0.3197	0.6716	0.6746
	Flat	0.7532	0.4900	-0.5030	0.5868	0.6935
	Param	-1.2611	-1.2671	-1.2779	-1.2760	-1.2722
Heterogenous Coef.	Ti-Homo	1.3901	1.5670	1.2802	1.6146	1.1904
	Ti-Hetero	1.3059	1.5676	1.5278	1.6140	1.1883
	Tv-Homo	0.6598	0.5030	0.4711	0.5313	0.6879
	Tv-Hetero	1.5067	1.5490	1.5286	1.5520	1.1013
	Flat	0.5675	0.4889	0.4966	0.4385	0.6205

Table A-23: Empirical Application: Predict Investment Rate, CRPS

		2015	2016	2017	2018	2019
Homogenous Coef.	Ti-Homo	0.0529	0.0633	0.0712	0.0599	0.0605
	Ti-Hetero	0.0398	0.0399	0.0511	0.0315	0.0464
	Tv-Homo	0.0532	0.0635	0.0721	0.0599	0.0634
	Tv-Hetero	0.0354	0.0394	0.0517	0.0332	0.0454
	Pooled	0.0535	0.0637	0.0712	0.0603	0.0627
	Flat	0.0537	0.0640	0.0723	0.0620	0.0604
	Param	0.3510	0.3550	0.3649	0.3601	0.3554
Heterogenous Coef.	Ti-Homo	0.0389	0.0382	0.1257	0.0343	0.0485
	Ti-Hetero	0.0391	0.0381	0.1271	0.0346	0.0485
	Tv-Homo	0.0639	0.072	0.0790	0.0696	0.0615
	Tv-Hetero	0.0384	0.0379	0.1248	0.0353	0.0486
	Flat	0.0702	0.0751	0.1587	0.0721	0.065

Figure A-5: Posterior Predictive Density for each industries, Year = 2019