Incorporating Prior Knowledge of Latent Group Structure in Panel Data Models[†]

Boyuan Zhang[‡]

Amazon.com

First Version: June 14, 2022
This Version: January 21, 2024

Abstract

The assumption of group heterogeneity has become popular in panel data models. We develop a constrained Bayesian grouped estimator that exploits researchers' prior beliefs on groups in a form of pairwise constraints, indicating whether a pair of units is likely to belong to a same group or different groups. We propose a prior to incorporate the pairwise constraints with varying degrees of confidence. The whole framework is built on the nonparametric Bayesian method, which implicitly specifies a distribution over the group partitions, and so the posterior analysis takes the uncertainty of the latent group structure into account. Monte Carlo experiments reveal that adding prior knowledge yields more accurate estimates of coefficients and scores predictive gains over alternative estimators. We apply our method to two empirical applications. In a first application to forecasting U.S. CPI inflation, we illustrate that prior knowledge of groups improves density forecasts when the data is not entirely informative. A second application revisits the relationship between a country's income and its democratic transition; we identify heterogeneous income effects on democracy with five distinct groups over ninety countries.

JEL CLASSIFICATION: C11, C14, C23, E31

KEY WORDS: Grouped Heterogeneity; Bayesian Nonparametrics; Dirichlet Process Prior; Density Forecast; Inflation Rate Forecasting; Democracy and Development.

---

1 Introduction

Numerous studies have examined and demonstrated the important role of panel data models in empirical research throughout the social and business sciences, as the availability of panel data has increased. Using fixed-effects, panel data permits researchers to model unobserved heterogeneity across individuals, firms, regions, and countries as well as possible structural changes over time. As individual heterogeneity is often empirically relevant, fixed-effects are an objective of interest in numerous empirical studies. For example, teachers' fixed-effects are viewed as a measure of teacher quality in the literature on teacher valued-added (Rockoff, 2004; Chetty et al., 2014); the heterogeneous coefficients are crucial for panel data forecasting (Liu et al., 2020; Pesaran et al., 2022). In practice, however, researchers may face a short panel where N is large and T is short and fixed. When applying the least squares estimator for the fixed-effects, a significant number of noisy estimates are produced. To alleviate this issue, a popular and parsimonious assumption that has recently been used is to introduce a group pattern into the individual coefficients, so that units within each group have identical coefficients (Bonhomme and Manresa (2015, BM hereafter), Su et al. (2016), Bonhomme et al. (2022)).

To recover the group pattern, we essentially face a clustering problem, e.g., dividing N units into several unknown groups. All existing methods for estimating group heterogeneity solve a clustering problem by assuming that units are exchangeable and treating all units equally *a priori*. In a cross-country application of evaluating the impact of climate change on economic growth (Hsiang, 2016; Henseler and Schumacher, 2019; Kahn et al., 2021), countries in different climatic zones are assumed to have equal probabilities of being grouped together. The assumption of exchangeability might not be reasonable since correlations are common between observations at proximal locations and researchers could have knowledge of the underlying group structure based on theories or empirical findings. For instance, Sweden and Finland, which share a border, an economic structure, and weather conditions, may have a higher chance of being in the same group than African countries. In such a scenario, it is preferable to use additional information to break the exchangeability between countries to facilitate grouping as opposed to clustering based solely on observations in the sample. The availability of this information drives us to formalize such prior knowledge, which we wish to leverage to improve model performance.

In this paper, we focus on the group heterogeneity in the linear panel data model and develop a nonparametric Bayesian framework to incorporate prior knowledge of groups, which is considered additional information that does not enter the likelihood function. The prior knowledge aids in clustering units into groups and sharpens the inference of group-specific parameters, particularly when units are not well-separated.

The whole framework is built on the nonparametric Bayesian method, where we do not impose a restriction on the number of groups, and model selection is not required. The baseline model is a linear panel data model with an unknown group pattern in fixed-effects,

slope coefficients, and cross-sectional error variances. We estimate the model using the stick-breaking representation (Sethuraman, 1994) of the Dirichlet process (DP) (Ferguson, 1973, 1974) prior, a standard prior in nonparametric Bayesian inference. In this framework, the number of groups is considered a random variable and is subject to posterior inference. The number of groups and group membership are estimated together with the heterogeneous coefficients. Moreover, since the DP prior implicitly defines a prior distribution on the group partitionings, the posterior analysis takes the uncertainty of the latent group structure into account.

The derivation of the proposed prior starts with summarizing prior knowledge in the form of pairwise constraints, which describe a bilateral relationship between any two units. Inspired by the work of Wagstaff and Cardie (2000), we consider two types of constraints: *positive-link* and *negative-link* constraints, representing the preference of assigning two units to the same group or distinct groups. Instead of imposing these constraints dogmatically, each constraint is given a level of accuracy that shows how confident the researchers are in their choice. There is a hyperparameter that controls the overall strength of the prior knowledge: a small value partially recovers the exchangeability assumption on units, whereas a large value confines the prior distribution of group partitioning around group structure based on prior knowledge. We choose the optimal value for the hyperparameter by maximizing the marginal data density. Summarizing prior knowledge in the form of pairwise constraints is practical and flexible since it eliminates the need to predetermine the number of groups and focuses on the bilateral relationships within any *subset* of units.

The aforementioned pairwise constraints are used to modify the standard DP prior. In particular, the pairwise constraints are combined with the prior distribution of the group partitioning, shrinking the distribution toward my prior knowledge. We refer to the estimator using the proposed prior as the Bayesian group fixed-effects (BGFE) estimator.

We derive a posterior sampling algorithm for the framework with the modified DP prior. Adopting conjugate priors on group-specific coefficients allows for drawing directly from posteriors using a computationally efficient Gibbs sampler. With the newly proposed prior, it can be shown that, compared to the framework that uses a standard DP prior, all that is needed to implement pairwise constraints is a simple modification to the posterior of the group indices.

The pairwise constraint-based framework is closely related and applicable to other models where group structure plays a role. Although we concentrate primarily on the panel data model, the DP prior with pairwise constraints applies to models without the time dimension, such as the standard clustering problem and the estimation of heterogeneous treatment effects. The framework is also applicable to estimating panel VARs (Holland et al., 1983), which involves multiple dependent variables. The group structure is used to overcome overparameterization and overfitting issues by clustering the VAR coefficients into groups, and pairwise constraints add additional information to the highly parameterized

model. Moreover, the proposed Gibbs sampler with pairwise constraints is connected to the *KMeans*-type algorithm, motivating a frequentist's counterpart of our estimator with a fixed K. Essentially, the assignment step in the *Pairwise Constrained-KMeans* algorithm (Basu et al., 2004a), a constrained version of the *KMeans* algorithm (MacQueen et al., 1967), is remarkably similar to the step of drawing a group membership indicator from its posterior. The same exact equivalence can be achieved by applying small-variance asymptotics to the posterior densities under certain conditions. To obtain the frequentist's analog of our pairwise constrained Bayesian estimators, one can utilize the same approach in BM with the *Pairwise Constrained-KMeans* algorithm.

We compare the performance of the BGFE estimator to alternative estimators using simulated data. The Monte Carlo simulation demonstrates that the BGFE estimator generates more accurate estimates of the group-specific parameters and the number of groups than the BGFE estimator without including any constraints. The improved performance is mostly attributable to the precise group structure estimation. The BGFE estimator clearly dominates the estimators that omit the group structure by assuming homogeneity or full heterogeneity. We also evaluate the performance of one-step ahead point, set, and density forecasts. Unsurprisingly, the accurate estimates translate into the predictive power of the underlying model; the BGFE estimator outperforms the rest of the estimators.

We apply the proposed method to two empirical applications. An application to forecasting the inflation of the U.S. CPI sub-indices demonstrates that the suggested predictor yields more accurate density predictions. The better forecasting performance is mostly attributable to three key characteristics: the nonparametric Bayesian prior, prior belief on group structure, and grouped cross-sectional heteroskedasticity. In a second application, we revisit the relationship between a country's income and its democratic transition. This question was originally studied by Acemoglu et al. (2008), who demonstrate that the positive income effect on democracy disappears if country fixed effects are introduced into the model. The proposed framework recovers a group structure with a moderate number of groups. Each group has a clear and distinct path to democracy. In addition, we identify heterogeneous income effects on democracy and, contrary to the initial findings, show that a positive income effect persists in some groups of countries, though quantitatively small.

LITERATURE. This paper relates to the econometric literature on clustering in panel data models. Early contributions include Sun (2005) and Buchinsky et al. (2005). Hahn and Moon (2010) provide economic and theoretical foundations for fixed effects with a finite support. Most recent works focus on linear[1] panel data models with discrete unobserved group heterogeneity. Lin and Ng (2012) and Sarafidis and Weber (2015) apply the *KMeans* algorithm to identify the unobserved group structure of slope coefficients. Bonhomme and Manresa (2015) also use the *KMeans* algorithm to recover the group pattern, but they assume group structure in the additive fixed effects. Bonhomme et al. (2022) modify this method and split

---

[1]See Wang and Su (2021); Bonhomme et al. (2022), among others, for procedures to identify latent group structures in nonlinear panel data models.

3

the procedure into two steps. They first classify individuals into groups using *KMeans* algorithm and then estimate the coefficients. Ando and Bai (2016) improved on BM's approach by allowing for group structure among the interactive fixed effects. The underlying factor structure in the interactive fixed effects is the key to forming groups. Su et al. (2016) develop a new variant of Lasso to shrink individual slope coefficients to unknown group-specific coefficients. This method is then extended by Su and Ju (2018) and Su et al. (2019). Freeman and Weidner (2022) consider two-way grouped fixed effects that allow for different group patterns in time and cross-sectional dimensions. Okui and Wang (2021) and Lumsdaine et al. (2022) identify structure breaks in parameters along with grouped patterns. From the Bayesian perspective, Kim and Wang (2019), Zhang (2020), and Liu (2022) adopt the Dirichlet process prior to estimate grouped heterogeneous intercepts in linear panel data models in the semiparametric Bayesian framework. Moon et al. (2023) incorporate a version of a spike and slab prior to recover one core group of units. Alternative methods, such as binary segmentation (Wang et al., 2018) and assumptions, such as multiple latent groups structure (Cheng et al., 2019; Cytrynbaum, 2021) have also been explored to flourish group heterogeneity literature.

Our work concerns prior knowledge. Bonhomme and Manresa (2015)'s grouped fixed-effects (GFE) estimator is able to include prior knowledge, but it is plagued by practical issues to some extent. They add a collection of individual group probabilities as a penalty term in the objective function, which is a N by K matrix describing the probability of assigning each unit to all potential groups. This additional penalty term balances the respective weights attached to prior and data information in estimation. The main challenge is providing the set of individual group probabilities for each potential value of K as the underlying *KMeans* algorithm requires model selection. It is rather cumbersome to assess these probabilities for each possible K and to adjust for changes in reallocating probabilities across K.

None but Aguilar and Boot (2022) explore heterogeneous error variance, and they extend BM's GFE estimator to allow for group-specific error variances. They modify the objective function to avoid the singularity issue in pseudo-likelihood. Despite the fact that their work paves the way for identifying groups in the error variance, their framework is not yet ready to satisfactorily incorporate prior knowledge because they face the same issue as BM. Building on these works, we investigate the value of prior knowledge of group structure.

This paper also relates to the literature of constraint-based semi-supervised clustering in statistics and computer science. Pairwise constraints have been widely implemented in numerous models and have been shown to improve clustering performance. In the past two decades, various pairwise constrained *KMeans* algorithms using prior information have been suggested (Wagstaff et al., 2001; Basu et al., 2002, 2004a; Bilenko et al., 2004; Davidson and Ravi, 2005; Pelleg and Baras, 2007; Yoder and Priebe, 2017). Prior information is also introduced in the model-based method. Shental et al. (2003) develop a framework to incorporate prior information for the density estimation with Gaussian mixture models. The Dirichlet process mixture model with pairwise constraints has been discussed in Vlachos

et al. (2008), Vlachos et al. (2009), Orbanz and Buhmann (2008), Vlachos et al. (2010), Ross and Dy (2013). Lu and Leen (2004), Lu (2007) and Lu and Leen (2007) assume the knowledge on constraints is incomplete and penalize the constraints in accordance with their weights. Law et al. (2004) extents Shental et al. (2003) to allow for soft constraints in the mixture model by adding another layer of latent variables for the group label. Nelson and Cohen (2007) propose a new framework that samples pairwise constraints given a set of probabilities related to the weights of constraints.

Our paper is closely related to Paganin et al. (2021), who address a similar problem using a novel Bayesian framework. Their proposed method shrinks the prior distribution of group partitioning toward a *full* target group structure, which is an initial clustering of *all* units provided by experts. This is demanding since not every application can have a full target group structure, as their birth defect epidemiology study did. Our framework circumvents this problem by using pairwise constraints, which are flexibly assigned to any two units. In addition, the induced shrinkage of their framework is produced by the distance function defined by Variation of Information (Meilă, 2007). It can be demonstrated that a partition can readily become caught in local modes, preventing it from ever shrinking toward the prior partition. The use of pairwise relationships in this paper circumvents this issue as well. By fixing the group indices of other pairs, our framework makes sure that the partition with a specific pair that fits our prior belief has a higher prior probability than the partition with a pair that goes against our prior belief.

OUTLINE. In section 2, we present the specification of the dynamic panel data model with group pattern in slope coefficients and error variances and provide details on nonparametric Bayesian priors without prior knowledge, which are then extended to accommodate soft pairwise constraints. Section 3 focuses on the posterior analysis, where the posterior sampling algorithm is provided. We also highlight the posterior estimate of group structure and discuss the connection to constrained *KMeans* models. We briefly discuss the extensions of the baseline model in section 5. In section 4, we present empirical analysis in which we forecast the inflation rate of the U.S. CPI sub-indices and estimate the country's income effect on its democracy. Finally, we conclude in section 6. Monte Carlo simulations, additional empirical results, and proofs are relegated to the appendix.

2   Model and Prior Specification

We begin our analysis by setting up a linear panel data model with group heterogeneity in intercepts, slope coefficients, and cross-sectional innovation variance. We then elaborate a nonparametric Bayesian prior for the unknown parameters that takes prior beliefs in the group pattern into account. We briefly highlight several key concepts of a standard nonparametric Bayesian prior as our proposed prior inherits some of its properties.

2.1 A Basic Linear Panel Data Model

We consider a panel with observations for cross-sectional units $i = 1, \ldots, N$ in periods $t = 1, \ldots, T$. Given the panel data set (y_{it}, x'_{it}), a basic linear panel data model with grouped heterogeneous slope coefficients and grouped heteroskedasticity takes the following form:

$$y_{it} = \alpha'_{g_i} x_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N\left(0, \sigma^2_{g_i}\right), \tag{2.1}$$

where x_{it} are a $p \times 1$ vector of covariates, which may contain intercept, lagged y_{it}, other informative covariates. α_{g_i} denote the group-specific slope coefficients (including intercepts). $\sigma^2_{g_i}$ are the group-specific variance. $g_i \in \{1, ..., K\}$ is the latent group index with an unknown number of groups K. ε_{it} are the idiosyncratic errors that are independent across i and t conditional on g_i. They feature zero mean and grouped heteroskedasticity $\sigma^2_{g_i}$, with cross-sectional homoskedasticity being a special case where $\sigma^2_{g_i} = \sigma^2$. This setting leads to a heterogeneous panel with group pattern modeled through both α_{g_i} and $\sigma^2_{g_i}$.

It is convenient to reformulate the model in (2.1) in matrix form by stacking all observations for unit i:

$$y_i = x_i \alpha_{g_i} + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2_{g_i} I_T\right), \tag{2.2}$$

where $y_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]'$, $x_i = [x_{i1}, x_{i2}, \ldots, x_{iT}]'$, $\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}]'$, and $G = [g_1, \ldots, g_N]$ is a vector of group indices.

Group structure is the key element in our approach. It can be either represented as a vector of group indices G describing to which group each unit belongs or as a collection of disjoint blocks $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ induced by G, where B_k contains all the units in the k-th group and K is the number of groups in the sample of size N. $|B_k|$ denotes the cardinality of the set B_k with $\sum_{k=1}^{K} |B_k| = N$.

Remark 2.1. *Identification issues may arise with certain specifications. If the grouped fixed-effects in α_{g_i} are allowed to vary over time, for example, $\sigma^2_{g_i}$ cannot be identified when the group $g = g_i$ contains only one unit. Aguilar and Boot (2022) propose a solution, but this problem is beyond the scope of this work. They suggest using the square-root objective function rather than the pseudo-log-likelihood function as the objective function, which replaces the logarithm of σ^2_i with the square root of σ^2_i, to avoid the singularity problem.*

Following Sun (2005), Lin and Ng (2012) and BM, we assume that the composition of groups does not change over time. In addition, for any group $k \neq k'$, we assume that they have different slope coefficients, e.g., $\alpha_k \neq \alpha_{k'}$, and no single unit can simultaneously belong to these two groups: $B_k \bigcap B_{k'} = \varnothing$. Note that these assumptions are used to simplified the prior construction and are not necessary to incorporate prior knowledge. As we show in Section 5, both assumptions can be relaxed by using slightly different priors.

The primary objective of this paper is to estimate the group-specific slope coefficients α_{g_i}, group-specific variance $\sigma^2_{g_i}$, group membership G as well as the unknown number of

groups K using full sample and prior knowledge of the group structure. Given estimates of group-specific coefficients, we are able to offer the point, set, and density forecasts of y_{it+h} for each unit i. Throughout this paper, we will concentrate on the one-step ahead forecast where $h = 1$. For multiple-step forecasting, the procedure can be extended by iterating y_{iT+h} in accordance with (2.1) given the estimates of parameters or estimating the model in the style of direct forecasting. The method proposed in this paper is applicable beyond forecasting. In certain applications, the heterogeneous parameters themselves are the objects of interest. For example, the technique developed here can be adapted to infer group-specific heterogeneous treatment effects.

2.2 Nonparametric Bayesian Prior with Knowledge on G

We propose a nonparametric Bayesian prior for the unknown parameters with prior beliefs on the group pattern. Figure 1 provides a preview of the procedure for introducing prior knowledge into the model. We propose to use pairwise constraints to summarize researchers' prior knowledge, with each constraint accompanied by a hyperparameter W indicating the researchers' levels of confidence in their choice. The W is then incorporated directly in the prior distribution of the group partition G, which is induced from a standard nonparametric Bayesian prior, yielding a new prior. We will elaborate the details throughout this subsection and highlight the clustering properties of the underlying nonparametric Bayesian priors in Section 2.3.

Figure 1: Graphical Representation of the Proposed Prior
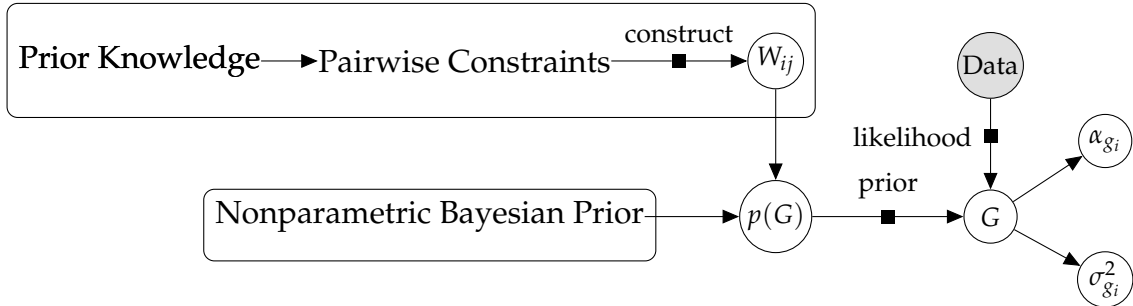
2.2.1   A New Prior with Soft Pairwise Constraints

The derivation of the proposed prior starts from summarizing prior knowledge in the form of pairwise constraints, which describe a bilateral relationship between any two units. Inspired by the literature on semi-supervised learning (Wagstaff and Cardie, 2000),[2] we consider two types of pairwise constraints: (1) positive-link (PL) constraints, $\mathcal{P}$, and (2) negative-

---

[2]We essentially follow the same idea of the pairwise constraints in Wagstaff and Cardie (2000). To better demonstrate the beliefs on constraints, we use different names: positive-link and negative-link, rather than must-link and cannot-link.

link (NL) constraints, $\mathcal{N}$. A positive-link constraint specifies that two units are more likely to be assigned to the same group, whereas a negative-link constraint indicates that the units are prone to be assigned to different groups.

Instead of imposing these constraints dogmatically, the constraint between units i and j is given a hyperparameter W_{ij} which describes how confident the researchers are in their choice for different types of constraints. W_{ij} is continuously valued on the real line, as depicted in Figure 2. On the one hand, the sign of W_{ij} specifies the constraint type, with a positive (negative) value indicating a PL (NL) constraint between i and j. On the other hand, the absolute value of W_{ij} reflects the strength of the prior belief. We become increasingly confident in our prior belief on units i and j as $|W_{ij}| \to \infty$. If $|W_{ij}| = \infty$, we essentially impose the constraint, which is known as a *hard* PL/NL constraint. Otherwise, it's a *soft* PL/NL constraint with a nonzero and finite W_{ij}. $W_{ij} = 0$ if there is no prior belief in units i and j.

Figure 2: Relationship Between W_{ij} and Pairwise Constraints
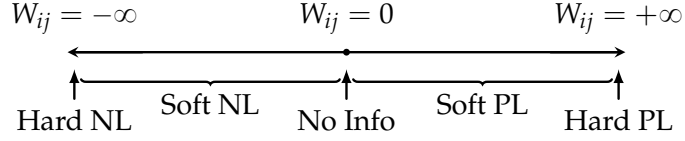
We assume the weight W_{ij} is a logit function of two user-defined hyperparameters[3], accuracy ψ_{ij} and type T_{ij}:

$$W_{ij} = T_{ij} \ln \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right). \tag{2.3}$$

Accuracy, $\psi_{ij} \in [0.5, 1)$, describes the user-specified probability of assigning a constraint for unit i and j being correct given our prior preference. Specifically, $\psi_{ij} = 1$ implies the constraint between i and j must be imposed since we confident that it is accurate, while specifying $\psi_{ij} = 0.5$ is equivalent to a random guess or no information is provided. ψ_{ij} is bounded below by 0.5, following the assumption that leaving the pair unrestricted is more rational than setting a less likely constraint. The type of constraints is denoted by T_{ij}. $T_{ij} = 1$ if unit i and j are specified to be positive-linked, and $T_{ij} = -1$ for a NL constraints. If the pair (i,j) doesn't involve any constraint, we assume $T_{ij} = 0$.

To incorporate these constraints into the prior, we propose modifying the exchangeable partition probability function (EPPF) or the prior distribution of group indices, $p(G)$, of the baseline Dirichlet process, which we will highlight in the Section 2.3. The resulting group partition will receive a strictly higher (lower) probability if it is (in)consistent with pairwise constraints. As a result, the induced prior on the group indices G directly depend on the

---

[3]This parametric form is related to the penalized probabilistic clustering proposed by Lu and Leen (2004, 2007). See detailed discussion in Appendix D.2.

characteristics of user-specific pairwise constraints and is able to increase or decrease the likelihood of a certain G.

In the presence of soft constraints, we modify the EPPF by multiplying a function of characteristics of constraints,

$$p(G|\psi,T) \propto p(G)\pi(\psi,T|G) = p(G)\prod_{i,j}\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right)^{cT_{ij}\delta_{ij}(G)}, \tag{2.4}$$

where $\psi_{ij}/(1-\psi_{ij})$ is the prior odds for the constraint between unit i and j, $\delta_{ij}(G)$ is a transformed Kronecker delta function such that

$$\delta_{ij}(G) = \begin{cases} 1 & \text{if } g_i = g_j \\ -1 & \text{if } g_i \neq g_j \end{cases}, \tag{2.5}$$

and c is a positive number that controls the overall strength of prior belief. For $c \to 0$, $p(G|\psi,T)$ corresponds to the baseline EPPF $p(G)$, while for $c \to \infty$, $p(G = G^*|\psi,T) \to 1$, where G^* satisfies all pairwise constraints.

Remark 2.2. *Due to the presence of pairwise constraints, the partition probability function presented in (2.4) no longer satisfies the exchangeable assumption as we now distinguish units within each group.*

Remark 2.3. *The current framework enables us to impose some constraints. It is an extreme case of soft constraint and thus handy to implement, requiring only setting $\psi_{ij} \to 1$ for the pair (i,j). Intuitively, any group partition violating the pairwise constraint between i and j (i.e., $T_{ij}\delta_{ij}(G) = -1$) will have zero probability, since for such a partition,*

$$\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right)^{cT_{ij}\delta_{ij}(G)} \to \left(\frac{1}{\infty}\right)^c = 0, \text{ this imples } p(G|\psi,T) = 0,$$

and hence the constraint on (i,j) is imposed and referred to as a hard constraint as opposed to soft constraint. By assigning proper ψ_{ij} for the pairs (i,j), we can flexibly combine soft and hard constraints inside a single specification.

Remark 2.4. *Soft pairwise constraints solve the transitivity issue that might be a problem for hard pairwise constraints. For instance, if we have $(1,2) \in \mathcal{P}$ and $(2,3) \in \mathcal{P}$, we can still have $(1,3) \in \mathcal{N}$ in the framework of soft pairwise constraints since it preserves the possibility of violating any of these constraints. This is not the case in hard pairwise constraints, as $(1,2) \in \mathcal{P}$ and $(2,3) \in \mathcal{P}$ implies $(1,3) \in \mathcal{P}$ by transitivity.*

With the definition of W_{ij}, we rewrite the partition probability function defined in (2.4) in terms of W_{ij} to ease notation,

$$p(G|\psi,T) = p(G|W) = \mathcal{C}(W,G,c)^{-1}p(G)\exp\left[c\sum_{i,j}W_{ij}\delta_{ij}(G)\right], \tag{2.6}$$

where

$$C(W, G, c) = \sum_{G'} p(G') \exp \left[c \sum_{i,j} W_{ij} \delta_{ij}(G') \right],$$ (2.7)

is a normalization constant and we will use the prior $p(G|W)$ hereinafter. In practice, we will first specify $(T_{ij}, \psi_{ij}) = $ (type, accuracy) for the constraint between unit i and j and then construct the corresponding weight W_{ij} via the equation (2.3).

Remark 2.5. *In the particular case where we don't have any constraint information,* $\exp \left(c \sum_{i,j} W_{ij} \delta_{ij} \right)$ *reduces to 1 as $W_{ij} = 0$ for all i and j, and recovers the original DP prior. Hence, our method can cater to all levels of supervision, ranging from hard constraints to a complete lack of constraints.*

2.2.2 The Effect of Constraints and Scaling Constant on Group Partitioning

The function $\pi(\psi, T|G)$ in Equation (2.4) is crucial in shifting the prior probability of G. By design, $T_{ij}\delta_{ij} = 1$ when the constraint between i and j is met in a group partitioning defined by G. The prior probability for G is therefore increased since $[\psi_{ij}/(1 - \psi_{ij})]^c > 1$. Similarly, if a group partitioning G violates the constraint between i and j, then $T_{ij}\delta_{ij} = -1$ and the prior probability for G drops due to $[\psi_{ij}/(1 - \psi_{ij})]^{-c} < 1$. Therefore, with $\pi(\psi, T|G)$, the resulting group partition is shrunk toward our prior knowledge without imposing any constraint.

To fix ideas, consider a simplified scenario with $N = 2$ units where there are at most two groups. For illustrative purposes, we set the concentration parameter to $a = 1$ so that $\Pr(g_1 = g_2) = \Pr(g_1 \neq g_2) = 0.5$[4] if no constraint exists. When $N = 2$, listing all partitions G is possible, $G \in \{(1,1), (1,2), (2,1), (2,2)\}$, and we can calculate the probabilities for each G using (2.6). As a result, we are able to derive the probability of units 1 and 2 belonging to the same or different groups, i.e., analytical formulae for $\Pr(g_1 = g_2)$ and $\Pr(g_1 \neq g_2)$, which neatly demonstrate the effect of ψ and c on group partitioning.

It is straightforward to show $\Pr(g_1 = g_2)$ as a function of ψ_{12}, T_{12}, and c:

$$\Pr(g_1 = g_2) = \frac{1}{1 + \exp(-4cW_{12})} = \frac{1}{1 + \left(\frac{\psi_{12}}{1 - \psi_{12}} \right)^{-4cT_{12}}}.$$ (2.8)

Figure 3 traces out the equation (2.8) for a range of c values. The left panel (a) displays the curve for a PL constraint. Firstly, observe that when $c = 0$, $\Pr(g_1 = g_2)$ remains unchanged at 0.5 regardless of the value of ψ. This is the situation in which c eliminates the constraint's effect on the prior. Next, given a particular c, $\Pr(g_1 = g_2)$ increases in ψ, which means that a stronger soft PL constraint between units 1 and 2 leads to higher chance of assigning both units to the same group. When ψ is fixed, increasing c easily results in a higher $\Pr(g_1 = g_2)$, indicating that a larger c value magnifies the effect of the PL constraint. In contrast, panel

---

[4]Antoniak (1974) provides analytical formulas for probabilities of more general events with larger N. In this example, $\Pr(g_1 = g_2) = \frac{1}{a+1}$.

(b) depicts the curve with a NL constraint. ψ_{12} and c clearly have the opposite effect on $Pr(g_1 = g_2)$: $Pr(g_1 = g_2)$ drops significantly as ψ_{12} or c increases. Notably, even with a large ψ_{12}, the soft constraint framework maintains the possibility of breaching the constraint, which is another important feature that preserves the chance of correctly assigning group indices even if the constraint is erroneous.

Figure 3: $Pr(g_1 = g_2)$ as a Function of ψ_{12} and c

(a) Positive-Link ($T_{12} = 1$)   (b) Negative-Link ($T_{12} = -1$)
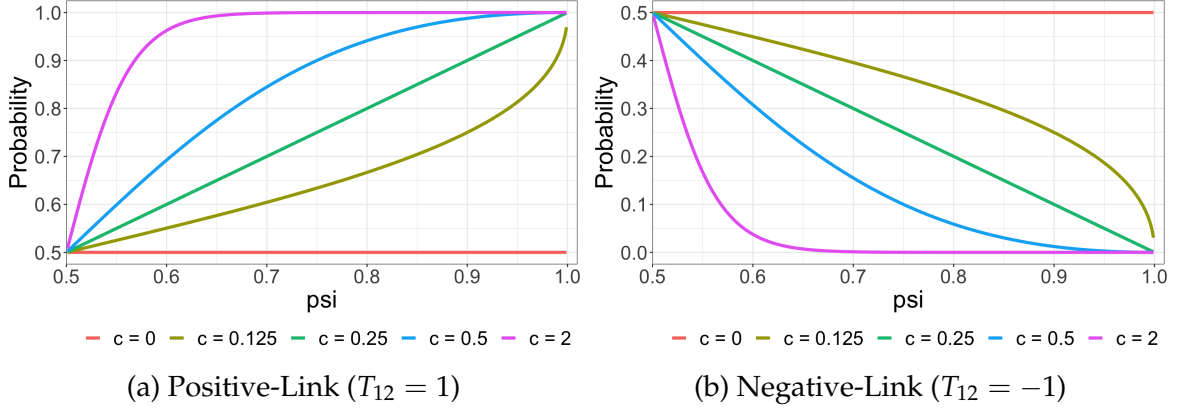
In the general case where numerous PL and NL constraints are enforced, c concurrently affects all constraints. In other words, the value of c determines the overall "strength" of the prior belief of G. If the prior belief is coherent with the real group partition, it would be preferable to have a large c to intensify the effect on constraints, allowing prior information to take precedence over data information, and vice versa.

Remark 2.6. *We propose to find the optimal c that maximizes marginal data density using grid search, see details in Appendix 4.1.2. Alternatively, the scaling constant c can be pair-specific and data-driven. Basu et al. (2004b), for instance, assume that c_{ij} is a function of observables, i.e., $c_{ij} = \psi(x_i, x_j; T_{ij})$. $\psi(x_i, x_j; T_{ij})$ is monotonically increasing (decreasing) in the distance between units i and j if they are involved in positive-link (negative-link) constraints. This reflects the belief that if two more distant units are assumed to be positive-linked (negative-linked), their constraint should be given more (less) weight.*

2.2.3 Specification of Soft Pairwise Constraints

In reality, it is practical to establish soft pairwise constraints based on existing information on group, even if it is not the genuine group partitioning. In the empirical analysis, for instance, we use the official expenditure categories of CPI sub-indices to construct soft pairwise constraints. When information on group partitioning is insufficient, especially when the number of units is large, these official expenditure categories may serve as a trustworthy starting point. Before formalizing the idea, we first introduce the prior similarity matrix

11

$\underline{\pi}^S$ which is a $N \times N$ symmetric matrix describing the prior probability of any two units belonging to the same group, i.e., $\underline{\pi}_{ij}^S = \Pr(g_i = g_j)$ conditional on all hyperparameters in the prior.

The general idea is to derive soft pairwise constraints using the existing information on a preliminary group partitioning $\underline{G}$, which is allowed to involve only a subset of units. We start with the type of constraints T_{ij} between any two units. Given the preliminary group structure, such as expenditure categories, we specify PL constraints for all pairs of units within the same group and NL constraints for all pairs of units from different groups. This means that we believe the preliminary group structure is correct *a priori*. Despite the fact that more elaborate and subtle constraints might be implemented, this rough specification is usually a great starting point.

The accuracy ψ_{ij} for constraints is then specified. When our prior knowledge is limited or the number of units is large, we cannot specify ψ_{ij} for all pairs with solid knowledge of them. Instead, one desirable yet simple choice is to assume ψ_{ij} again based on preliminary group partitioning $\underline{G}$. More specifically, all units in the same group are positive-linked with identical ψ_{ij}^{PL}, i.e., for units i and j from the group $\underline{g}_i = \underline{g}_j = \underline{g}$, we have $\psi_{ij}^{PL} = c_{\underline{g}}$. Units from different groups are assumed to be negative-linked with identical ψ_{ij}^{NL}, i.e., for units i and j from distinct groups, we assume $\psi_{ij}^{NL} = c_{\underline{g}_i \underline{g}_j}$ and $c_{\underline{g}_i \underline{g}_j} = c_{\underline{g}_j \underline{g}_i}$. Following this strategy, ψ_{ij} depends solely on $\underline{G}$ and hence two units from the same group would have identical soft pairwise constraints with other units. Notice that the number of possible distinct ψ_{ij} reduces from $N(N-1)/2$ to $\bar{K}(\bar{K}+1)/2$, where $\bar{K}$ is the number of groups in $\underline{G}$ and $\bar{K} \ll N$.

This framework permits no prior belief in certain units. If at least one unit in a pair (i, j) is not included in $\underline{G}$, we assume that this pair of units is free of constraints and we set T_{ij} to 0 or ψ_{ij} to 0.5 in the prior. Note that the absence of a constraint does not ensure that the units i and j are completely unrelated. Instead, if both i and j are involved in constraints with a third unit k, or are connected through a series of l constraints, $i \leftrightarrow k_1 \leftrightarrow \cdots \leftrightarrow k_l \leftrightarrow j$, then the prior probability of i and j belonging to the same group differs from the prior probability without any constraints. If we wish to prevent two units from linking *a priori*, they must not be subject to any constraints with the remaining units.

The aforementioned specification strategy induces a block prior similarity matrix, i.e., for an unit i, $\underline{\pi}_{ij}^S = \underline{\pi}_{ik}^S$ if $\underline{g}_j = \underline{g}_k$. Intuitively, if two units have identical soft pairwise constraints and hence posit an identical relationship with all other units, they are equivalent and exchangeable. As a result, these units should have an equal prior probability of sharing the same group index with any other units. More formally,

Theorem 1 (Stochastic Equivalence). *Given two units j, k from the same prior group, if $\psi_{jm} = \psi_{km}$ for all $m = 1, 2, .., N$, then $\Pr(g_i = g_j) = \Pr(g_i = g_k)$ for all unit i in the prior, given weights W.*

Theorem 1 echos the concept of *stochastic equivalence* (Nowicki and Snijders, 2001) in

stochastic block model[5] (SBM) (Holland et al., 1983). In less technical terms, for nodes p and q in the same group, p has the same (and independent) probability of connecting with node r, as q does. Interestingly, this relationship is not coincidental. The prior draw of group membership with the aforementioned specification of T_{ij} and ψ_{ij} can be viewed as a simulation of a simple SBM. In a simple SBM, there are two essential components: a vector of group memberships and a block matrix, each element of which represents the edge probability of two nodes, given their group memberships. In our case, the preliminary group structure serves as the group membership in SBM. The DP prior and the weight (or T_{ij} and ψ_{ij}) of each constraint induce a prior similarity probability comparable to the block matrix.

Let's consider an example of 90 units. The preliminary group structure divides units into 3 groups, with groups 1, 2 and 3 containing 25, 30 and 35 units, respectively. Figure 4 shows the prior similarity matrix, which is based on the aforementioned specification strategy, so it becomes a block matrix with equal entries in each of nine blocks. Units within the same group are stochastically equivalent, as their prior probabilities of being grouped not only with each other but also with units from other groups are the same. As a result, the similarity probability of each pair depends solely on their preliminary membership (and ψ).

Figure 4: Prior Similarity Matrix under Stochastic Equivalence
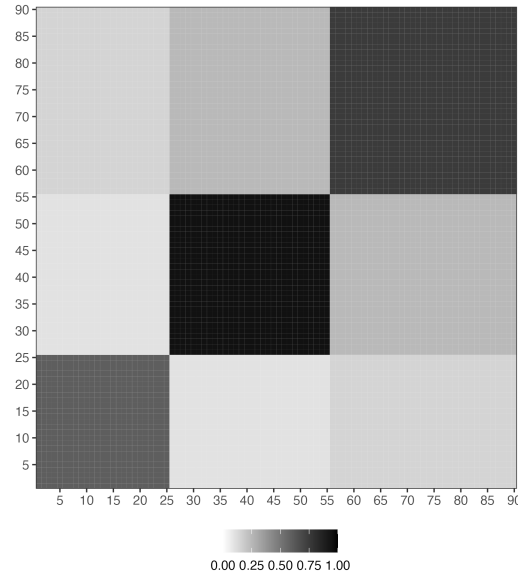
2.2.4 Comparison to Existing Methods

Bonhomme and Manresa (2015) incorporates prior knowledge of group membership by adding a penalty term to the objective function. They assume that prior information is in the form of probabilities which describe the prior probability of unit i belonging to group k

---

[5]For a more comprehensive review of the stochastic block model, see Lee and Wilkinson (2019).

13

with at most K groups as $\omega_{ik}^{(K)}$. Consequently, the estimated group index is given by:

$$\hat{g}_i(\beta, \alpha) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_{it} - \beta^{(K)\prime} x_{it} - \alpha_{kt}^{(K)} \right)^2 - C \ln \omega_{ik}^{(K)}, \qquad (2.9)$$

where $C > 0$ is a hyperparameter need to be tuned further and K is the predetermined number of groups.

The penalty determines the weights assigned to prior and data information in estimation. Due to the fact that K is frequently unknown in advance, this method requires model selection to determine the ideal number of groups. Assume we have n_K alternative options for K. We have a $N \times K$ matrix $\omega^{(K)} = \left\{ \omega_{ik}^{(K)} \right\}$ for prior information for a given K. As a result, in order to pick a model, we must therefore provide n_K sets of prior probability matrix $\omega^{(K)}$, which is cumbersome and inconvenient. For instance, if K has values ranging from $3, 5, 10$ and $N = 200$, there are 3,600 entries for ω, none of which can be missing or undefined. In addition, the information criteria may be unreliable in the finite-sample results, necessitating further care when selecting an appropriate variant for empirical application.

Summarizing prior knowledge through pairwise constraints is often more practical than the penalty function approach in BM and solves the aforementioned practical issues. Pairwise relationships can be derived intuitively from researchers' input without requiring in-depth knowledge of the underlying groups; researchers do not need to fix the number of groups K or group membership *a priori*. One only needs to focus on a pair of units each time and specify the preference of assigning them to the same or different groups. Moreover, since the pairwise constraints are incorporated into the DP prior, which implicitly defines a prior distribution on the group partitions, model selection is not required and the posterior analysis also takes the uncertainty of the latent group structure into account.

Paganin et al. (2021) offer a statistical framework for including concrete prior knowledge on the partition. Their proposed method aims to shrink the prior distribution towards a *complete* prior clustering structure, which is an initial clustering involving *all* units provided by experts. Specifically, they suggest a prior on group partition that is proportional to a baseline EPPF of the DP prior multiplied by a penalization term,

$$p\left(G | G_0, m\right) \propto p(G) e^{-m d(G, G_0)} \qquad (2.10)$$

with $m > 0$ is a penalization parameter, $d\left(G, G_0\right)$ is a suitable distance measuring how far G is from G_0, and $p(G)$ indicates the same EPPF as in (2.4). Because of the penalization term, the resulting group indices G shrink toward the initial target G_0.

This framework is parsimonious and easy to implement, but it comes with a cost. The method is incapable of coping with an initial clustering in a subset of the units under study or multiple plausible prior partitions; otherwise, the distance measure is not well-defined. In addition, the authors suggest utilizing Variation of Information (Meilă, 2007) as the distance measure. It can be shown that the resulting partition can easily become trapped in local

modes, leading the partition to never shrink toward G_0. They also argue that other available distance methods have flaws. As a result, the penalty term does not function as anticipated.

Our proposed framework with pairwise constraints is more flexible than adopting a target partition. Actually, the target partition in Paganin et al. (2021) can be viewed as a special case of the pairwise constants, in which every unit must be involved in at least one PL constraint. Our framework could manage partitions involving arbitrary subsets of the units by tactically specifying the bilateral relationships. Most importantly, when the group indices of other pairs are fixed, our framework ensures that the partition containing a specific pair that is consistent with our prior belief receives a strictly greater prior probability than the partition that is inconsistent with our prior belief. This guarantees that the generated G shrinks in the direction of our prior belief.

2.3 Nonparametric Bayesian Prior

The baseline model contains the parameters listed below: $(\alpha, \sigma^2, \pi, \xi, a, \phi)$. We rely mostly on nonparametric Bayesian models.[6] Bayesian nonparametric models have emerged as rigorous and principled paradigms to bypass the model selection problem in parametric models by introducing a nonparametric prior distribution on the unknown parameters. The prior assumes that a collection of α and σ^2 is drawn from the Dirichlet process prior.[7] π is a vector of mixture probabilities in Dirichlet process that is produced by the stick-breaking approach with stick length ξ. a is the concentration parameter in the Dirichlet process, whereas ϕ is a collection of hyperparameters in the base measure B_0. We consider prior distributions in the partially separable form,[8]

$$p(\alpha, \sigma^2 | a, \phi, \xi) p(\xi | a) p(a).$$

We tentatively focus on the random coefficients model where, conditional on G, α_{g_i} and $\sigma^2_{g_i}$ are independent to the conditional set that includes initial value of each unit y_{i0}, the initial values of predetermined variables, and the whole history of exogenous variables. The assumption guarantees that α_{g_i} and σ_{g_i} can be sampled separately and simplifies the inference of the underlying distribution of α_{g_i} and σ_{g_i} to an unconditional density estimation problem, therefore lowering computational complexity. The joint distribution of heterogeneous parameters as a function of the conditioning variables can then be modeled to extend the model to the correlated random coefficient model, which is briefed in Section 2.3.3. A

---

[6]For a more comprehensive review of the nonparametric Bayesian literature, see Ghosal and Van der Vaart (2017) and Müeller et al. (2018).

[7]There have been some empirical works that use Dirichlet process model with panel data. Dirichlet process mixture prior is specified for either the distribution of the innovations (Hirano, 2002) or intercepts (Fisher and Jensen, 2022).

[8]The joint prior includes ξ but not π. Because the stick-breaking formulation of ξ is a deterministic transformation of ξ, knowing ξ is identical to knowing π.

full explanation and derivation for the correlated random coefficient model are provided in the online appendix.

2.3.1 Prior on Group-Specific Parameters

In the nonparametric Bayesian literature, the Dirichlet Process (DP) prior (Ferguson, 1973, 1974; Sethuraman, 1994) is a canonical choice, notable for its capacity to construct group structure and accommodate an infinite number of possible group components.[9] The DP mixture is also known as a "infinite" mixture model due to the fact that the data indicate a finite number of components, but fresh data can uncover previously undetected components (Neal, 2000). When the model is estimated, it chooses automatically an appropriate subset of groups to characterize any finite data set. Therefore, there is no need to determine the "proper" number of groups.

The DP prior can be written as an infinite mixture of point mass with the probability mass function:

$$\left(\alpha_i, \sigma_i^2\right) \sim \sum_{k=1}^{\infty} \pi_k \delta_{\left(\alpha_k, \sigma_k^2\right)} \text{ with } \left(\alpha_k, \sigma_k^2\right) \sim B_0(\phi), \tag{2.11}$$

where δ_x denotes the Dirac-delta function concentrated at x and B_0 is the base distribution. We adopt an Independent Normal Inverse-Gamma (INIG) distribution for the base distribution B_0:

$$B_0(\phi) := INIG\left(\mu_\alpha, \Sigma_\alpha, \frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right), \tag{2.12}$$

with a set of hyperparameters $\phi = \left(\mu_\alpha, \Sigma_\alpha, \frac{v_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$.

The group probabilities π_k are constructed by an infinite-dimensional stick-breaking process (Sethuraman, 1994) governed by the concentration parameter a,

$$\pi_k \equiv \xi_k \prod_{j<k}(1 - \xi_j) \text{ for } k > 1, \text{ and } \pi_1 = \xi_1, \tag{2.13}$$

where stick lengths ξ_k are independent random variables drawn from the beta distribution,[10] $Beta(1, a)$. The group probability will be random but still satisfy $\sum_{k=1}^{\infty} \pi_k = 1$ almost surely.

Equation (2.13) is essential to understanding how the DP prior controls the number of groups. The building of group probabilities is compared to the breaking of a stick of unit length sequentially, in which the length of each break is assigned to the current value of π_k. As the number of groups increases, the probability created by the stochastic process decreases because the remaining stick becomes shorter with each break. In practice, the number of groups does not increase as fast as N due to the characteristic of the stick-breaking process that leads the group probability to soon approach zero.

---

[9]See Appendix A.1 for a brief overview of the DP and Appendix A.1.3 for its clustering properties.
[10]Recall that a $Bata(m, n)$ distribution is supported on the unit interval and has mean $m/(m + n)$.

Although in principle we do not restrict the maximum number of groups and allow the number to rise as N increases, a finite number of instances will only occupy a finite number of K components. The concentration parameter a in the prior of ξ_k determines the degree of discretization – the complexity of the mixture and, consequently, K, as also revealed in (A.3). As $a \to 0$, the realizations are all concentrated at a single value, however as $a \to \infty$, the realizations become continuous-valued as its based distribution. Specifically, Antoniak (1974) derives the relationship between a and the number of unique groups,

$$E\left(K|a\right) \approx a \log\left(\frac{a+N}{a}\right) \quad \text{and} \quad \text{Var}\left(K|a\right) \approx \alpha \left[\log\left(\frac{\alpha+N}{\alpha}\right) - 1\right],$$

that is, the expected number of unique groups is increasing in both a and the number of units N.

Escobar and West (1995) highlights the importance of specifying a when imposing prior smoothness on an unknown density and demonstrates that the number of estimated groups under a DP prior is sensitive to a. This suggests that a data-driven estimate of a is more reasonable. Moreover, Ascolani et al. (2022) emphasizes the importance of introducing a prior for a as it is crucial for learning the true number of groups as N increases and hence establishing the posterior consistency. We define a gamma hyperprior for a and update it based on the observed data in order to alter the smoothness level. This step generates a posterior estimate of a, which indirectly determines the number of groups K without reestimating the models with different group sizes. Essentially, this represents "automated" model selection.

Collectively, we specify a DP prior for (α_i, σ_i^2). The DP prior is a mixture of an infinite number of possible point masses, which can be constructed through the stick-breaking process. The discretization of the underlying distribution is governed by the concentration parameter a. With a hyerprior on a, we permit the data to determine the number of groups K present in the data, which can expand unboundedly along with the data.

2.3.2   Prior on Group Partitions

In a formal Bayesian formulation, a prior distribution is specified to partition $\mathcal{B}$ with associated indices G. Despite the fact that DP prior does not specify this prior distribution explicitly, we can characterize it using the exchangeable partition probability function (EPPF) (Pitman, 1995). As we briefly mentioned in the last subsection, the EPPF plays a significant role in connecting the prior belief on group structure to the DP prior, which is included as part of our proposed prior distribution in Equation (2.4).

The EPPF characterizes the distribution of a partition $\mathcal{B} = \{B_1, B_2, \dots, B_K\}$ induced by G. As the generic Dirichlet process assumes units are exchangeable, any permutation has no effect on the joint probability distribution of G; hence, the EPPF is determined entirely by the number of groups and the size of each group. Pitman (1995) demonstrates that the EPPF

of the Dirichlet process has the closed form,

$$p(G) = \frac{\Gamma(a)}{\Gamma(a+N)} a^K \prod_{k=1}^{K} \Gamma\left(|B_k|\right), \tag{2.14}$$

where a is the concentration parameter and $\Gamma(x) = (x-1)!$ denotes the Gamma function. Note that the partition $\mathcal{B}$ is conceived as a random object and hence the group number K is not predetermined, but rather is a function of G, $K = K(G)$.

Sethuraman (1994) and Pitman (1996) constructively show that group indices/partitions can be drawn from the EPPF for DP using the stick-breaking process defined in (2.13). As a result, the EPPF does not explicitly appear in the posterior analysis in the current setting so long as the priors for the stick lengths are included.

2.3.3 Correlated Random Coefficients Model

As suggested by Chamberlain (1980), allowing the individual effects to be correlated with the initial condition can eliminate the omitted variable bias. This subsection presents the first attempt to introduce dependence between grouped effects and covariates, under the presence of group structure in both heterogeneous slope coefficients and cross-sectional variance. The underlying theorems, such as posterior consistency, and the performance of the framework are left for future study.

We primarily follow the proposed framework in Liu (2022) and utilize Mixtures of Gaussian Linear Regressions (MGLRx) for the group-specific parameters. MGLRx prior is discussed in Pati et al. (2013) and can be viewed as a Dirichlet Process Mixture (DPM) prior that takes the dependence of covariates into account. Notice that the correlated random coefficients model requires a DPM-type prior for α_i and σ_i^2. This is because α_i and σ_i^2 are assumed to be correlated with covariates of each individual, and as such, they are not identical within a group.

Following Liu (2022), we first transform σ_i^2 and define $l_i = \log \frac{\bar{\sigma}^2(\sigma_i^2 - \underline{\sigma}^2)}{\bar{\sigma}^2 - \sigma_i^2}$, where $\underline{\sigma}^2$ $(\bar{\sigma}^2)$ is some small (large) positive number. This transformation simplifies the prior for σ_i^2, which is now dependent on covariates, and ensures that a similar prior structures can be applied to both α_i and l_i.

In the correlated random coefficients model, the DPM prior for α_i or σ_i^2 is an infinite mixture of Normal densities with the probability density function:

$$\alpha_i \sim \sum_{k=1}^{\infty} \pi_k(w_{i0}) N\left(\mu_k^\alpha [1\ w_{i0}']', \Omega_k^\alpha\right), \tag{2.15}$$

$$\sigma_i^2 \sim \sum_{k=1}^{\infty} \pi_k(w_{i0}) N\left(\mu_k^\sigma [1\ w_{i0}']', \Omega_k^\sigma\right), \tag{2.16}$$

where $w_{i0} = [1, y_{i0}, x_{i,0:T}]'$ is the conditioning set at period 0, which includes initial value of each unit y_{i0}, the initial values of predetermined variables, and the whole history of exogenous variables. Notice that α_i and σ_i^2 share the same set of group probabilities $\pi_k(w_{i0})$.

Similar but not identical to the DP prior, it is the component parameters $(\mu_k^\alpha, \Omega_k^\alpha)$ or $(\mu_k^\sigma, \Omega_k^\sigma)$ that are directly drawn from the base distribution G_0. G_0 is assumed to be a conjugate Matricvariate-Normal-Inverse-Wishart distribution.

The group probabilities are now characterized by a *probit* stick-breaking process (Rodriguez and Dunson, 2011),

$$\pi_k(w_{i0}) = \Phi\left(\zeta_k\left(w_{i0}\right)\right) \prod_{j<k} \left(1 - \Phi\left(\zeta_j\left(w_{i0}\right)\right)\right), \tag{2.17}$$

where the stochastic function ζ_k is drawn from a Gaussian process, $\zeta_k \sim GP\left(0, V_k\right)$ for $k = 1, 2, \cdots$. The Gaussian process is assumed to have zero mean and the covariance function V_k. defined as follows,

$$V_k(x, x') = \tau_v \exp\left(-A_k \|x - x'\|_2^2\right), \tag{2.18}$$

where $\tau_v \sim IG(\frac{v_v}{2}, \frac{v_v}{2})$ and A_k has its own hyperprior, see details in Pati et al. (2013).

3 Posterior Analysis

This section describes the procedure for analyzing posterior distributions for the baseline model described in (2.1) with the priors specified in Section 2.2.1. The joint posterior distribution of model parameters is

$$p(\alpha, \sigma^2, \Xi, a, G | Y, X, W, \phi)$$
$$\propto p(Y|X, \alpha, \sigma^2, G) p(\alpha, \sigma^2 | \phi) p(G|\Xi) p(W|G) p(\Xi|a) p(a), \tag{3.1}$$

where $p(Y|X, \alpha, \sigma^2, G)$ is the likelihood function given by equation (2.1) for an i.i.d. model conditional on group indices G, and $p(W|G)$ is the additional term of pairwise constraints with the form $p(W|G) = \prod_{i=1}^{N} \prod_{j=1}^{N} \exp\left(cW_{ij}\delta_{ij}\right)$.

3.1 Posterior Sampling

Draws from the joint posterior distribution can be obtained by using blocked Gibbs sampling. The algorithm is derived from Ishwaran and James (2001) and Walker (2007). Due to the use of a finite-dimensional prior and truncation, the method described in Ishwaran and James (2001) cannot truly address our demand for estimating the number of groups without a predetermined value or upper bound. We employ the slice sampler (Walker, 2007), which is the exact block Gibbs sampler for the posterior computation in infinite-dimensional Dirichlet process models, modifying the block Gibbs sampler of Ishwaran and

James (2001) to avoid truncation approximations. Walker (2007) augments the posterior distribution with a set of auxiliary variables consisting of i.i.d. standard uniform random variables, i.e., $u_i \overset{iid}{\sim} U(0,1)$ for $i = 1, 2, .., N$. The augmented posterior is then represented as

$$p(\alpha, \sigma^2, \Xi, a, G, u | Y, X, W, \phi)$$
$$\propto p(Y|X, \alpha, \sigma^2, G) p(\alpha, \sigma^2 | \phi) p(W|G) p(\Xi|a) p(a) \prod_i \mathbf{1}(u_i \leqslant \pi_{g_i}). \tag{3.2}$$

where $\prod_i \mathbf{1}(u_i \leqslant \pi_{g_i})$ is substituted for $p(G|\Xi)$ in the equation (3.1).

To roughly see how slice sampling works, recall that the group probabilities are constructed in a sequential manner, following a stick-breaking procedure. The leftover of the stick after each break gets smaller and smaller. Given the finite number of units, we can always find the smallest K^* such that for all groups $k \geqslant K^*$, the minimum of u_i among all units is larger than π_k, which is bounded above by the length of the leftover after k breaks, $1 - \sum_{j=1}^{k} \pi_j$. More formally,

$$K^* = \min_k \left\{ u^* > 1 - \sum_{j=1}^{k} \pi_j \right\}, \text{ with } u^* = \min_{1 \leqslant i \leqslant N} u_i. \tag{3.3}$$

As a result, all units receive strictly zero probability of being assigned to any group $k = K^* + 1, K^* + 2, ..., N$ since the indicator function $\mathbf{1}(u_i \leqslant \pi_k)$ is zero.

There are two advantages to incorporating the auxiliary variable u into the model. First and foremost, u directly determines the largest possible number of groups in each sampling iteration. This reduces the support of G and Ξ to a finite space, enabling us to solve a problem of finite dimensions without truncation. Furthermore, u have no effect on the joint posterior of other parameters because the original posterior can be restored by integrating out u_i for $i = 1, 2, ..., N$.

The Gibbs sampler are used to simulate the joint posterior distribution of $(\alpha, \sigma^2, \Xi, u, a, G)$. We break this vector into blocks and sequentially sampling for each block conditional on the current draws for the other parameters and the data. The full conditional distributions for each block are easily derived using the conjugate priors specified in Section 2.

For the group-specific parameters, we directly draw samples from their posterior densities as we adapt conjugate priors. The posterior inference with respect to (α, σ^2) becomes standard once we condition on the latent group indices G. It is essentially a Bayesian panel data regression for each group. The conditional posterior for the stick length Ξ is a beta distribution given G, and hence direct sampling is possible.

We follow Walker (2007) to derive the posterior of auxiliary variable u. As u are standard uniformly distributed, the posterior is a uniform distribution defined on $(0, \pi_{g_i})$, conditional on the group probabilities and group indices. In terms of the concentration parameter a, we use a 2-step procedure proposed by Escobar and West (1995). Following their approach, we

first draw a latent variable η from $Beta(a+1, N)$. Then, given η and number of groups K^a in the current iteration, we directly draw a from a mixture of two Gamma distribution.

It is worth noting that the steps for implementing the DP prior with or without soft pairwise constraints are the same for all parameter besides the group indices G. This is due to the fact that soft pairwise constraints only affect other parameters through the group indices. It is handy to sample group indices with soft pairwise constraints conditional on other parameters. The posterior probability of assigning unit i to group k includes additional term $p(W_i|G) = \prod_{j \neq i, g_j = k} \exp\left(2cW_{ij}\delta_{ij}\right)$ to rewards (penalizes) the abidance (violation) of constraints,

$$\bar{\pi}_k = p\left(g_i = k | \alpha, \sigma^2, G^{(i)}, u, Y, X, W\right) = \frac{p\left(y_i | \alpha_k, \sigma_k^2, Y, X\right) \mathbf{1}\left(u_i \leqslant \pi_k\right) p(W_i|G)}{\sum_{k'=1}^{K^*} p\left(y_i | \alpha_{k'}, \sigma_{k'}^2, Y, X\right) \mathbf{1}\left(u_i \leqslant \pi_{k'}\right) p(W_i|G)}, \tag{3.4}$$

where K^* is the maximal number of groups after generating potential new group-specific slope coefficients and variance. We then draw the group index for unit i from a multinomial distribution:

$$g_i = k, \text{ with probability } \bar{\pi}_k. \tag{3.5}$$

Algorithm 1 below summarizes the algorithm for the proposed Gibbs sampling. For illustrative purposes, we focus primarily on the posterior densities of major parameters and omit details on step (vii). In short, step (vii) creates potential groups by sampling new (α_k, σ_k^2) from the prior if the latest K^* based on newly drawn u_i and π_k is larger than previous K^*, which indicate the current iteration permits more groups. This Detailed derivations and explanation of each step are provided in Appendix C.2.

Algorithm 1. *(Gibbs Sampler for Random Coefficients Model with Soft Pairwise Constraints)*
For each iteration $s = 1, 2, .., N_{sim}$,

(i) *Calculate number of active groups:* $K^a = \max_{1 \leqslant i \leqslant N} g_i^{(s-1)}$.

(ii) *Group-specific slope coefficients: draw* $\alpha_k^{(s)}$ *from* $p\left(\alpha_k | \sigma^{2(s-1)}, G^{(s-1)}, Y, X\right)$ *for* $k = 1, 2, ..., K^a$.

(iii) *Group-specific variance: draw* $\sigma_k^{2(s)}$ *from* $p\left(\sigma_k^2 | \alpha^{(s)}, G^{(s-1)}, Y, X\right)$ *for* $k = 1, 2, ..., K^a$.

(iv) *Group "stick length": draw* $\xi_k^{(s)}$ *from* $p\left(\xi_k | a^{(s-1)}, G^{(s-1)}\right)$ *for* $k = 1, 2, ..., K^a$ *and update group probability in accordance to the stick-breaking procedure.*

(v) *Auxiliary variable: draw* $u_i^{(s)}$ *from* $p\left(u_i | \Xi^{(s)}, G^{(s-1)}\right)$ *for* $i = 1, 2, ..., N$ *and calculate* $u^* = \min_{1 \leqslant i \leqslant N} u_i$.

(vi) *DP concentration parameter: draw a latent variable* η *from* $Beta(a+1, N)$ *and draw* $a^{(s)}$ *from* $p\left(a | \eta, K^a\right)$.

(vii) *Generate potential groups based on* u^* *and find the maximal number of groups* K^*.

(xi) Group indices: draw g_i from $p\left(g_i = k | \alpha^{(s)}, \sigma^{2(s)}, G^{(i)}, u, Y, X, W\right)$ for $i = 1, 2, ..., N$ and $k = 1, 2, ..., K^*$.

3.2 Determining Partition

In contrast to popular algorithms such as agglomerative hierarchical clustering or the *KMeans* algorithm, which return a single clustering solution, Bayesian nonparametric models provide a posterior over the entire space of partitions, enabling the assessment of statistical properties, such as the uncertainty on the number of groups.

However, when the group structure is part of the major conclusion of an empirical analysis, the point estimate of group structure becomes crucial. Wade and Ghahramani (2018) discuss in detail an appropriate point estimate of the group partitioning based on the posterior draws. From the decision theory, the point estimate G^* minimizes the posterior expected loss,

$$G^* = \underset{\widehat{G}}{\operatorname{argmin}} \, \mathbb{E}\left[L(G, \widehat{G}) | Y\right] = \underset{\widehat{G}}{\operatorname{argmin}} \sum_G L(G, \widehat{G}) p(G|Y),$$

where the loss function $L(G, \widehat{G})$ is the variation of information by Meilă (2007), which measures the amount of information lost and gained in changing from partition G to $\widehat{G}$.[11] The Variation of Information is based on the Shannon entropy $H(\cdot)$, and can be computed as

$$\text{VI}\left(G, \widehat{G}\right) = -H(G) - H\left(\widehat{G}\right) + 2H\left(G \wedge \widehat{G}\right)$$

$$= \sum_{j=1}^{K} \frac{\lambda_j}{N} \log\left(\frac{\lambda_j}{N}\right) + \sum_{l=1}^{K'} \frac{\lambda_l'}{N} \log\left(\frac{\lambda_l'}{N}\right) - 2 \sum_{j=1}^{K} \sum_{l=1}^{K'} \frac{\lambda_{jl}^{\wedge}}{N} \log\left(\frac{\lambda_{jl}^{\wedge}}{N}\right),$$

where log denotes log base 2, $\lambda_j = |B_j|$ is the cardinality of the group j, and $\lambda_{jl}^{\wedge}$ the size of blocks of the intersection $G \wedge \widehat{G}$ and hence the number of indices in block j under partition G and block l under $\widehat{G}$.

Wade and Ghahramani (2018) show that the optimal group partitioning can be identified based on the posterior similarity matrix,

$$g^* = \underset{\widehat{g}}{\operatorname{argmin}} \sum_{i=1}^{N} \log\left(\sum_{j=1}^{N} \mathbf{1}\left(\widehat{g}_j = \widehat{g}_i\right)\right) - 2 \sum_{i=1}^{N} \log\left(\sum_{j=1}^{N} P\left(g_j = g_i | Y, X, W\right) \mathbf{1}\left(\widehat{g}_j = \widehat{g}_i\right)\right)$$

$$(3.6)$$

where $P\left(g_j = g_i | Y, X, W\right)$ is the (i, j) entry of the posterior similarity matrix. We refer to Wade and Ghahramani (2018) for additional properties and empirical evaluations.

---

[11] Another possible loss function is the 0-1 loss function $L(G, \widehat{G}) = \mathbf{1}(G = \widehat{G})$, which leads to G^* being the posterior mode. This loss function is undesirable since it ignores similarity between two partitions. For instance, a partition that deviates from the truth in the allocation of only one unit is penalized the same as a partition that deviates from the truth in the allocation of numerous units. Furthermore, it is generally recognized that the mode may not accurately reflect the distribution's center.

3.3 Connection to Constrained KMeans Algorithm

The procedure of Gibbs sampling with soft constraints in Algorithm 1 is closely related to constrained clustering in the computer science literature. In this parallel literature, constrained clustering refers to the process of introducing prior knowledge to guide a clustering algorithm. For a subset of the data, the prior knowledge takes the form of constraints that supplement the information derived from the data via a distance metric. As we shall see below, under several simplifying assumptions, our framework could be reduced to a deterministic method for estimating group heterogeneity using a constrained *KMeans* algorithm. Though this deterministic method may address the practical issues in BM, it only works for certain restricted models and hence is not as general as our proposed framework.

We start with a brief review of the Pairwise Constrained KMeans (*PC-KMeans*) clustering algorithm by Basu et al. (2004a), which is a well-known clustering algorithm in the field of semi-supervised machine learning. It's a pairwise constrained variant of the standard *KMeans* algorithm in which an augmented objective function is used in the assignment step. Given a collection of observations $(y_1, y_2, \ldots, y_N)$, a set of positive-link constraints $\mathcal{P}$, a set of negative-link constraints $\mathcal{N}$, the cost of violating constraints $w = \{w_{ij}^p, w_{ij}^n\}$ and the number of groups K, the *PC-KMeans* algorithm divides N observations into K groups (the *assignment* step) so as to minimize the following objective function,

$$\underbrace{\frac{1}{2} \sum_{k=1}^{K} \sum_{i \in B_k} \|y_i - \mu_k\|^2}_{\text{within-cluster sum of squares}} + \underbrace{\sum_{(i,j) \in \mathcal{P}} w_{ij}^p \mathbf{1} \left(g_i \neq g_j\right) + \sum_{(i,j) \in \mathcal{N}} w_{ij}^n \mathbf{1} \left(g_i = g_j\right)}_{\text{cost of violation}}, \qquad (3.7)$$

where μ_k is the centroid of group k, i.e., $\mu_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i$, B_k is the set of units assigned to group k, and $|B_k|$ is the size of group k. The first part is the objective function for the conventional *KMeans* algorithm, while the second part accounts for the incurred cost of violating either PL constraints (w_{ij}^p) or NL constraints (w_{ij}^n).

Similar to *KMeans*, *PC-KMeans* alternates between reassigning units to groups and recomputing the means. In the assignment step, it determines a disjoint K partitioning that minimizes (3.7). Then the update step of the algorithm recalculates centroids of observations assigned to each cluster and updates μ_k for all k.

By applying asymptotics to the variance of distributions within the model, we demonstrate linkages between the posterior sampler of our constrained BGFE estimator and *KMeans*-type algorithms in Theorem 2. We investigate small-variance asymptotics for posterior densities, motivated by the asymptotic connection between the Gibbs sampling algorithm for the Dirichlet process mixture model and *KMeans* (Kulis and Jordan, 2011), and demonstrate that the Gibbs sampling algorithm for the CBG estimator with soft constraints encompasses the constrained clustering algorithm *PC-KMean* in the limit.

Theorem 2. (Equivalency between BGFE with Soft Constraints and *PC-KMeans*)
If the following conditions hold,

(i) *group pattern is in fixed-effects but not in slope coefficients, i.e., $x_{it} = 1$. Other covariates might be introduced, but they cannot have grouped effects on y_{it};*

(ii) *The number of group is fixed at K;*

(iii) *Homoscedasticity: $\sigma_k^2 = \sigma^2$ for all $k = 1, 2, ..., K$;*

(iv) *Constraint weights is scaled by the variance of errors: $W_{ij} \rightarrow W_{ij}/\sigma^2$;*

then the proposed Gibbs sampling algorithm for the BGFE estimator with soft constraint embodies the PC-KMeans clustering algorithm in the limit as $\sigma^2 \rightarrow 0$. In particular, the posterior draw of group indices G is the solution to the PC-KMeans algorithm.

We return to the world of grouped fixed-effects models. In fact, the clustering algorithm is essential for BM and Bonhomme et al. (2022), who use the *KMeans* algorithm to reveal the group pattern in the fixed-effects. With the theorem described above, it motivates a *constrained* version of BM's GFE estimator. We show that it is straightforward to incorporate prior knowledge in the form of soft paired restrictions into the GFE estimator. The *soft pairwise constrained* grouped fixed-effects (SPC-GFE) estimator is defined as the solution to the following minimization problem given the number of groups K:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{G}\right) = \underset{\theta, \alpha, G}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2 + c \left[\sum_{(i,j) \in \mathcal{P}} w_{ij}^p \mathbf{1}\left(g_i \neq g_j\right) + \sum_{(i,j) \in \mathcal{N}} w_{ij}^n \mathbf{1}\left(g_i = g_j\right) \right], \quad (3.8)$$

where the minimum is taken over all possible partitions G of the N units into K groups, common parameters θ, and group-specific time effects α. w_{ij}^p and w_{ij}^n are the user-specified costs on PL and NL constraints.

For given values of θ and α, the optimal group assignment for each individual unit is

$$\widehat{g}_i(\theta, \alpha) = \underset{g \in \{1, ..., K\}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2 + c \left[\sum_{(i,j) \in \mathcal{P}} w_{ij}^p \mathbf{1}\left(g_i \neq g_j\right) + \sum_{(i,j) \in \mathcal{N}} w_{ij}^n \mathbf{1}\left(g_i = g_j\right) \right], \quad (3.9)$$

where we essentially apply the *PC-KMeans* algorithm to get the group partition. The SPC-GFE estimator of (θ, α) in (3.8) can then be written as

$$(\widehat{\theta}, \widehat{\alpha}) = \underset{\theta, \alpha}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{\widehat{g}_i t}\right)^2, \quad (3.10)$$

where $\widehat{g}_i = \widehat{g}_i(\theta, \alpha)$ is given by (3.9). θ and α are computed using an OLS regression that controls for interactions of group indices and time dummies. The SPC-GFE estimate of g_i is then simply $\widehat{g}_i(\widehat{\theta}, \widehat{\alpha})$.

Remark 3.1. *While the SPC-GFE estimator implements soft constraints, it still requires a predetermined number of group K and model selection.*

4 Empirical Applications

We apply our panel forecasting methods to the following two empirical applications: inflation of the U.S. CPI sub-indices and the income effect on democracy. The first application focuses mostly on predictive performance, whereas the second application focuses primarily on parameter estimation and group structure.

4.1 Model Estimation and Measures of Forecasting Performance

To accommodate richer assumptions on model, we use variants of the baseline model in Equation (2.1) in this section, either by adding common regressors or allowing for time-variation in the intercept. We use the conjugate prior for all parameters, see details in Appendix B.

4.1.1 Specification of Constraints

In both applications, the prior knowledge of the latent group structure or the pre-grouping structure covers all units. In the first application, CPI sub-indices can be clustered by expenditure category, whereas countries in the second application may be grouped according to their geographic regions. We build positive-link and negative-link constraints given the prior knowledge: all units within the same group are presumed to be positive-linked, while units from different categories are believed to be negative-linked. In terms of the accuracy of constraints, ψ_{ij}^{PL} and ψ_{ij}^{NL} are equal for all constraints with the same type, following the strategy described in Section 2.2.3. In the applications below, we fix $\psi_{ij}^{PL} = 0.65$ and $\psi_{ij}^{NL} = 0.55$ to reflect the belief that PL constraints (attracting forces) play slightly more important role than the NL constraints (repelling forces) and NL constraints cannot be ignored. Finally, we construct weights W_{ij} using (2.3). Notice that these assumptions on prior and hyperparameters are an example to showcase how the proposed framework works with real data. In practice, we may specify constraints for a subset of units with different levels of weights, either in a data-driven manner (for instance, highly correlated units may fall into the same group with a high level of confidence) or in a model-based manner (i.e., W is a function of covariates).

4.1.2 Determining the Scaling Constant c

Given that the dimension of the space of group partitions increases exponentially with the number of units N, attention must be given while selecting c across analyses with different N. As suggested by Paganin et al. (2021), calibrating the modified prior is computationally intensive. We are facing a trade-off between investing time to get the prior "exactly right"

and letting the constant c be an estimated model parameter. As such, we propose to find the optimal c that maximizes marginal data density using grid search.

In the Monte Carlo simulation, the value of c is fixed for simplicity, but in the empirical applications, c is determined by marginal data density (MDD). We calculate MDD using the harmonic mean estimator (Newton and Raftery, 1994), which defined as

$$\hat{m}^{HM}(y) = \left[\frac{1}{S} \sum_{j=1}^{S} \frac{1}{p\left(y|\theta^{(j)}\right)} \right]^{-1}, \tag{4.1}$$

given a sample $\theta^{(j)}$ from the posterior $p(\theta|y)$. The simplicity of the harmonic mean estimator is its main advantage over other more specialized techniques. It uses only within-model posterior samples and likelihood evaluations, which are often available anyway as part of posterior sampling. We finally choose the optimal value for c that maximizes MDD.

4.1.3 Estimators

We consider six estimators in the section. The first three estimators are our proposed Bayesian grouped fixed-effects (BGFE) estimator with different assumptions on cross-sectional variance and pairwise constraints. The last three estimator ignore the group structure.

(i) *BGFE-he-cstr*: group-specific slope coefficients and heteroskedasticity *with* constraints.

(ii) *BGFE-he*: group-specific slope coefficients and heteroskedasticity *without* constraints.

(iii) *BGFE-ho*: homoskedastic version of *BGFE-he*.

(iv) *Pooled OLS*: fully homogeneous estimator

(v) *AR-he*: flat-prior estimator that assumes $p(\alpha_i) \propto 1$ corresponds to standard AR model with additional regressor u_t in this environment.

(vi) *AR-he-PC*: *AR-he* with the lagged value of the first principal component as additional regressor.

In the first application, we focus on inflation forecasting. For the most recent advances in this topic, Faust and Wright (2013) provide a comprehensive overview of a large set of traditional and recently developed forecasting methods. Among many candidate methods, we choose the AR model as the benchmark and exclusively include it as an alternative estimator in this exercise. This is because the AR is relatively hard to beat and, notably, other popular methods, such as the Atkeson–Ohanian version random walk model (Atkeson et al., 2001), UCSV (Stock and Watson, 2007), and TVP-VAR (Primiceri, 2005), generally do as reasonably well as the AR model, according to Faust and Wright (2013).

4.1.4 Posterior Predictive Densities

We 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 variables x_{iT+1} at $T + 1$.

The posterior predictive distribution for unit i is given by

$$p(y_{iT+1}|Y, X) = \int p(y_{iT+1}|Y, X, \Theta) p(\Theta|Y, X) d\Theta, \tag{4.2}$$

where Θ is a vector of parameters $\Theta = \left(\alpha_{g_i}, \sigma_{g_i}^2, g_i \right)$. This density is the posterior expectation of the following function:

$$p(y_{iT+1}|Y, X, \Theta) = \sum_{k=1}^{K(G)} \mathbf{1}(g_i = k) p(y_{iT+1}|Y, X, \Theta), \tag{4.3}$$

which is invariant to relabeling the components of the mixture and $K(G)$ is the number of groups in G. Given S posterior draws, the posterior predictive distribution estimated from the MCMC draws is

$$\hat{p}(y_{iT+1}|Y, X) = \frac{1}{S} \sum_{j=1}^{S} \left[\sum_{k=1}^{K^{(j)}(G)} \mathbf{1}(g_i = k) p \left(y_{iT+1}|Y, X, \Theta^{(j)} \right) \right]. \tag{4.4}$$

where

$$p \left(y_{iT+1}|Y, X, \Theta^{(j)} \right) = \phi \left(y_{iT+1}; \alpha_{g_i}^{(j)'} x_{it+1} + \gamma^{(j)} z_{it+1}, \sigma_{g_i}^{(j)2} \right). \tag{4.5}$$

We can therefore draw samples from $\hat{p}(y_{iT+1}|Y, X)$ by simulating (2.1) forward conditional on the posterior draws of Θ and observations. Note that MCMC exhibits the true Bayesian predictive distribution, implicitly integrating over the entire underlying parameter space.

4.1.5 Point Forecasts

We evaluate the point forecasts via the real-time recursive out-of-sample Root Mean Squared Forecast Error (RMSFE) under the quadratic compound loss function averaged across units. Let $\hat{y}_{iT+1|T}$ represent the predicted value conditional on the observed data up to period T, the loss function is written as

$$L \left(\hat{y}_{1:N,T+1|T}, y_{1:N,T+1} \right) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_{iT+1|T} - y_{iT+1} \right)^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_{iT+1|T}^2, \tag{4.6}$$

where $y_{i,T+1}$ is the realization at $T + 1$ and $\hat{\varepsilon}_{iT+1|T}$ denote the forecast error.

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

$$\hat{y}_{1:N,T+1|T} = \underset{\hat{y} \in \mathbb{R}^N}{\operatorname{argmin}} \int_{-\infty}^{\infty} L\left(\hat{y}, y_{1:N,T+1}\right) p(y_{1:N,T+1}|Y,X)dy_{1:N,T+1}$$

$$= \underset{\hat{y} \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} E\left[\left(\hat{y} - y_{iT+1}\right)^2 | Y, X\right]. \tag{4.7}$$

This implies optimal posterior forecast is the posterior mean,

$$\hat{y}_{i,T+1|T} = E\left(y_{iT+1}|Y,X\right), \text{ for } i = 1, \ldots, N. \tag{4.8}$$

Conditional on posterior draws of parameters, the mean forecast can be approximated by the Monte Carlo averaging,

$$\hat{y}_{i,T+1|T} \approx \frac{1}{S} \sum_{j=1}^{S} \hat{y}_{iT+1|T}^{(j)} = \frac{1}{S} \sum_{j=1}^{S} \hat{\alpha}_{g_i}^{(j)'} x_{iT+1}. \tag{4.9}$$

Finally, the RMSFE across units is given by

$$RMSFE_{T+1} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_{i,T+1} - \hat{y}_{i,T+1|T}\right)^2}. \tag{4.10}$$

4.1.6 Density Forecasts

To compare the performance of density forecasts for various estimators, we report the average log predictive scores (LPS) to assess the performance of the density forecast from the view of the probability distribution function. As suggested in Geweke and Amisano (2010), the LPS for a panel reads as,

$$LPS_{T+1} = -\frac{1}{N} \sum_{i=1}^{N} \ln \int p\left(y_{iT+1}|Y,X,\Theta\right) p(\Theta|Y,X)d\Theta, \tag{4.11}$$

where the expectation can be approximated using posterior draws:

$$\int p\left(y_{iT+1}|Y,X,\Theta\right) p(\Theta|Y,X)d\Theta \approx \frac{1}{S} \sum_{j=1}^{S} p\left(y_{iT+1}|Y,X,\Theta^{(j)}\right). \tag{4.12}$$

The following results are also robust to other metrics such as the continuous ranked probability score (Matheson and Winkler, 1976; Hersbach, 2000).

4.2 Inflation of the U.S. CPI Sub-Indices

Policymakers and market participants are very interested in the abilities to reliably predict the future disaggregated inflation rate. Central banks predict future inflation trends to jus-

tify interest rate decisions, control and maintain inflation around their targets. The Federal Reserve Board forecasts disaggregated price categories for short-term inflation forecasting (Bernanke, 2007). They rely primarily on the bottom-up approach that focuses on estimating and forecasting price behavior for the various categories of goods and services that make up the aggregate price index. Moreover, investors in fixed-income markets in the private sector wish to forecast future sectoral inflation in order to anticipate future trends in discounted real returns. Some private firms also need to predict specific inflation components in order to forecast price dynamics and reduce risks accordingly.

In this section, we demonstrate the use of constrained BGFE estimators with prior knowledge on the group pattern to forecast inflation rates for the sub-indices of U.S. Consumer Expenditure Index (CPI). We focus primarily on the one-step ahead point and density forecast. Due to space constraints, we only report the group partitioning for the most recent month in the main text.

4.2.1 Model Specification and Data

Model: We start by exploring the out-of-sample forecast performance of a simple, generic Phillips curve model. It is a panel autoregressive distributed lag (ADL) model with a group pattern in the intercept, coefficients, as well as error variance. The model is given by

$$y_{it+1} = \alpha_{g_i} + \sum_{j=0}^{p-1} \rho_{g_i,j} y_{it-j} + \beta_{g_i} u_t + \varepsilon_{it+1}, \quad \varepsilon_{it+1} \sim N(0, \sigma_{g_i}^2). \tag{4.13}$$

where y_{it} is year-over-year inflation rate, i.e., $y_{it} = \log(\text{price}_{it}/\text{price}_{it-12})$, and u_t is the slack measure for the labor market, the unemployment gap. We fix p at 3 because the benchmark AR model would have the best predictive performance.

Data: We collect the sub-indices of CPI for all urban consumer (CPI-U) that include food and energy. The raw data is obtained from the U.S. Bureau of Labor Statistics (BLS), which is recorded on a monthly basis from January 1947 to August 2022. The CPI-U is a hierarchical composite index system that partitions all consumer goods and services into a hierarchy of increasingly detailed categories. It consists of eight major expenditure categories (1) Apparel; (2) Education and Communication; (3) Food and Beverages; (4) Housing; (5) Medical Care; (6) Recreation; (7) Transportation; (8) Other Goods and Services. Each category is composed of finer and finer sub-indexes until the most detailed levels or "leaves" are reached. This hierarchical structure can be represented as a tree structure, as shown in Figure 5. It is important to note that the parent series and its child series may be highly correlated and readily form a group due to the fact that parent series are generated from child series. For instance, the *Energy Services* is expected to correlated with its child series *Utility gas service* and *Electricity*. Due to our focus on group structure, it is vital to eliminate all parent series in

order to prevent not just double-counting but also dubious grouping results. More details regarding the data are provided in Appendix F.1.

Figure 5: Hierarchical Structure of CPI
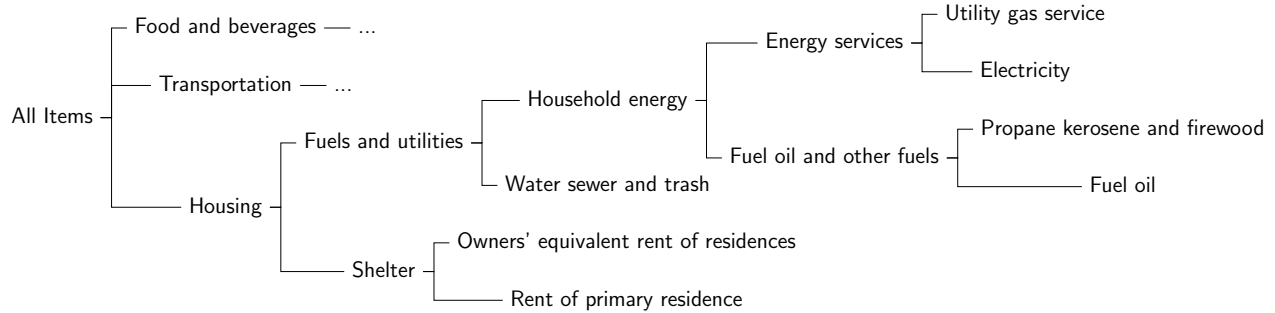
Pre-grouping: The official expenditure categories are used to build PL and NL constraints: all units within the same categories are presumed to be positive-linked, while units from different categories are believed to be negative-linked.

We focus on the CPI sub-indices after January 1990 for two reasons: (1) the number of sub-indices before 1990 was relatively small, diminishing the importance of the group structure; and (2) the consumption has been changed and more expenditure series were introduced in the 1990s as a result of the popularity of electronic products, food care, etc. After the elimination of all parent nodes, the unbalanced panel consists of 156 sub-indices in eight major expenditure categories. We employ rolling estimation windows of 48 months[12] and require each estimation sample to be balanced, removing individual series lacking a complete set of observations in a given window. Finally, we generate 329 samples with the first forecast produced for April 1995.

4.2.2   Results

We begin the empirical analysis by comparing the performance of point and density forecasts across 329 samples. Throughout the analysis, the *AR-he* estimator serves as the benchmark as it essentially assumes individual effects.

In Figure 6, we present the frequency of each estimator with the lowest RMSFE in the panel (a) and the boxplot[13] of the ratio of RMSFE relative to the *AR-he* estimator in the panel (b). First, the AR-he and AR-he-PC estimators, which rely only on an individual's own past data, are not competitive in point forecasts and perform considerably worse than the others. This implies that it is highly advantageous to explore cross-sectional information to improve

---

[12]The benchmark *AR-he* model scores the best overall performance with a window size of 48.

[13]The boundaries of the whiskers is based on the 1.5 IQR value. All other points outside the boundary of the whiskers are plotted as outliers in red crosses.

point forecasts. Moreover, the BGRE-he-cstr estimator scores the highest frequency of being the best estimator despite the fact that BGFE-he-cstr, BGFE-he, BGFE-ho, and pooled OLS estimators all utilize cross-sectional information. Examining the box plot, we find that the BGFE-ho and pooled OLS estimators, which overlook heteroskedasticity, can achieve greater performance in some samples, but make poorer forecasts more often than the other estimators. BGFE-he-cstr and BGFE-he estimator, on the other hand, typically outperform the others and the benchmark in terms of median RMSFE and the ability to produce forecasts with the lowest RMSFE without also increasing the risk of generating the least accurate forecasts.

Figure 6: RMSFE - All Samples

(a) Freq. lowest RMSFE

(b) Boxplot: $RMSE_m$ / $RMSE_{AR-he}$
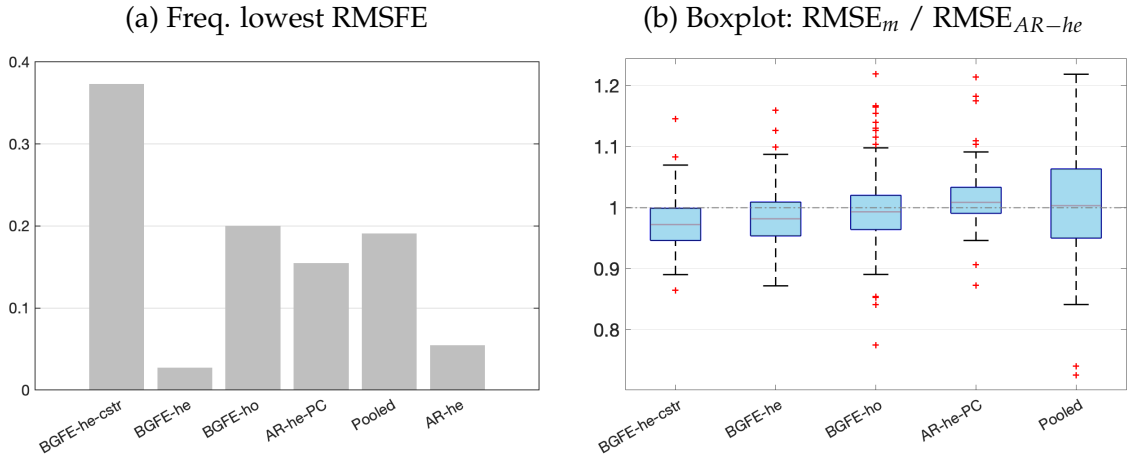
The revealing patterns of the density forecast are significantly distinct from those of the point forecast. Figure 7 depicts the log predictive score (LPS) for density forecast. The most notable pattern from the panel (a) is that the BGFE-he-cstr estimator, which incorporates prior knowledge, is dominating and outperforms the rest in over 80% of the samples. It emerges as the apparent winner in this case. Furthermore, when generating density forecast, the BGFE-ho and pooled OLS are not as accurate as they are in point forecast: they never have the lowest LPS across samples. This also confirms that the heteroscedasticity[14] is a well-known feature of the inflation time series (Clark and Ravazzolo, 2015). In the boxplot, we ignore BGFE-ho and pooled OLS and show the differences in LPS between the respective estimators and the AR-he estimator. As LPS differences represent percentage point differences, BGFE-he-cstr can provide density forecasts that are up to 22% more accurate compared to the benchmark model. Finally, despite the fact that the BGFE-he-cstr and BGFE-he estimators are mainly based on the same algorithm, the use of prior knowledge on group pattern further enhances the performance, resulting in the BGFE-he-cstr estimator having a lower LPS and scoring the best model with the highest frequency.

Next, we assess the value of adding prior information about groups by comparing the performance of the BGFE-he-cstr and BGFE-he estimators exclusively. The solid black line

---

[14]We provide more results in Section G.1.2 to explore the importance of heteroskedasticity in density forecast for the inflation.

Figure 7: Log Predictive Scores - All Samples

(a) Freq. lowest LPS

(b) Boxplot: LPS$_m$ - LPS$_{AR}$
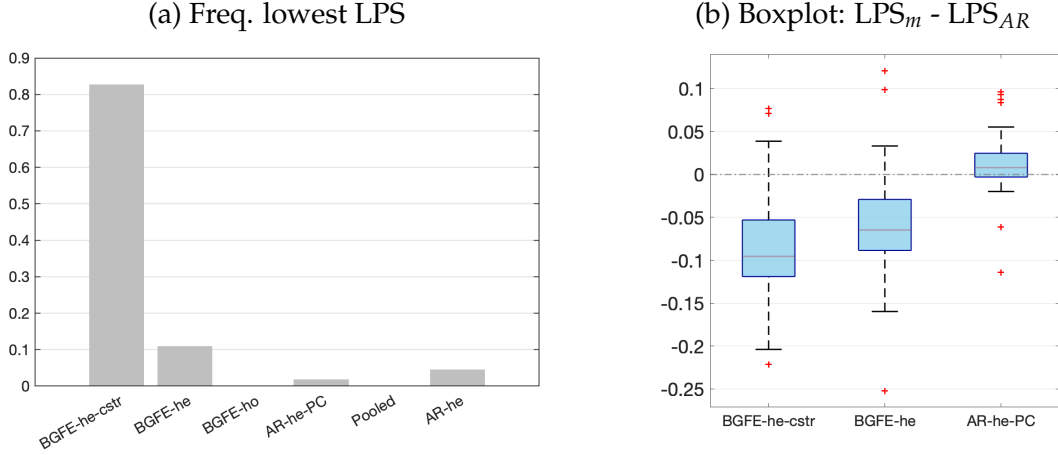
in Figure 8 represents the ratio of RMSE between BGFE-he-cstr and BGFE-he. The periods during which BGFE-he-cstr, BGFE-he, and all other estimators achieve the lowest RMSE are indicated by pink, blue, and green shaded areas, respectively. Though the BGFE-he-cstr estimator is not always the best across samples, the prior information improves the performance of the Bayesian grouped estimator. The BGFE-he-cstr estimator performs better than the BGFE-he estimator in most samples, with an average improvement of 2%.

Figure 8: Relative RMSE, BGFE-he-cstr

Adding prior information on groups substantially improves the accuracy of density forecasts. Figure 9 shows the comparison between BGFE-he-cstr and BGFE-he in terms of the difference in LPS. We find the prior information valuable as BGFE-he-cstr outperforms BGFE-he in more than 98% of the samples. Clearly, the majority of the figure is covered by a pink background, showing that BGFE-he-cstr is typically the best choice. All of these facts demonstrate that adding prior informatio is favorable and essential, especially for density forecasting.

Having specified pairwise constraints across sub-indices, we provide a prior on G that shrinks the group structure toward the eight expenditure categories with equal accuracy for all pairs inside each category. As Theorem 1 suggested, our prior specification essentially

Figure 9: Relative LPS, BGFE-he-cstr
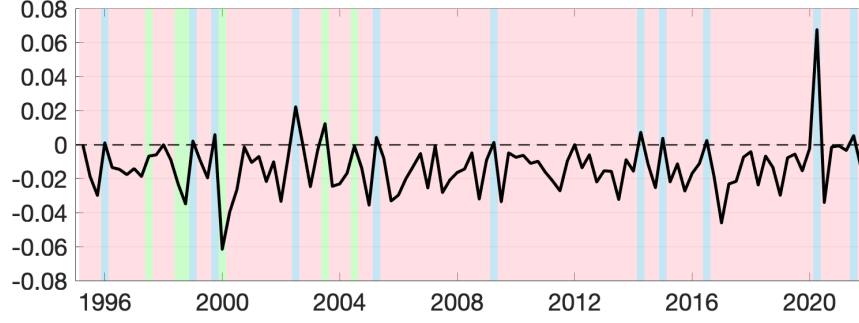
assumes that the prior probability of any two units that come from the same expenditure category being in the same group is equal, and the prior group pattern is actually the expenditure category. We now examine the posterior of group structure to demonstrate how the distribution of G gets updated by data. In order to accomplish this, we construct a posterior similarity matrix (PSM), whose (i, j)-element records the posterior p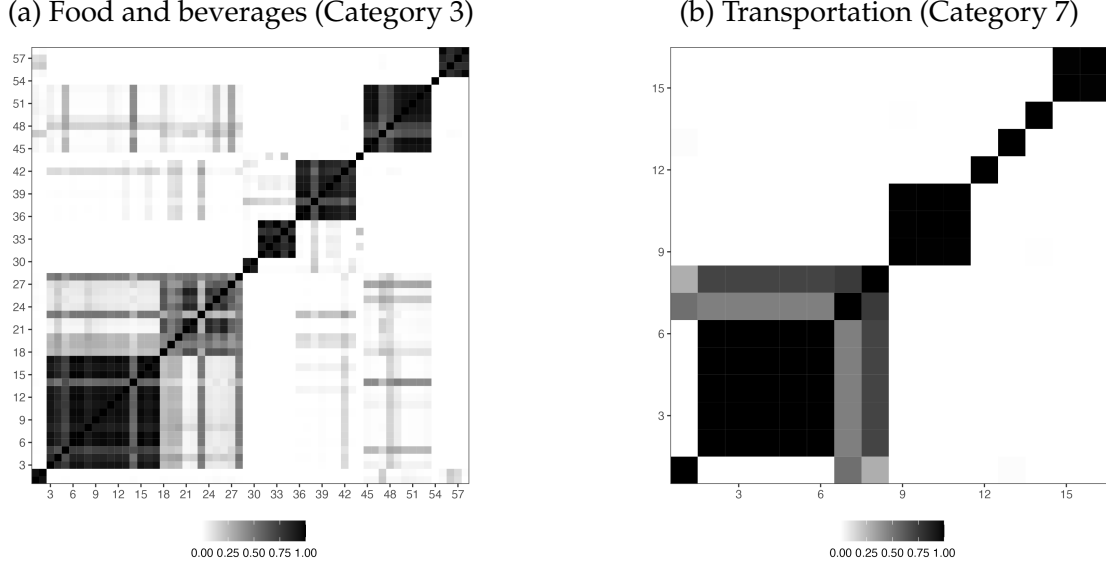robabilities of units i and j being in the same group. For illustrative purposes, we present the results for the last sample, in which we forecast CPI in August 2022. Figure 10 depicts the PSM generated by BGFE-he-crst for the series in the categories of *Food and Beverages* and *Transportation*. A darker block indicates a higher posterior probability of being in one group. A common pattern emerges: even though some sub-indices are joined together frequently, as shown in the dark diagonal blocks, it is extremely unlikely that all series within the same category belong to the same group. Some series have relatively low or zero probabilities of being grouped together, as suggested by the white and gray off-diagonal blocks. This indicates that the group structure based on official expenditure categories is not optimal, which may result in inaccurate forecasting. Instead, our suggested framework uses information from both prior beliefs and data to reinvent the group pattern, leading to improved forecasting performance.

Finally, we restrict our analysis to the point estimate of group partition, i.e., the single grouping solution, rather than the posterior over the whole universe of partitions. Figure 11 depicts the posterior point estimate of G for the last sample ended in August 2022, derived using the approach described in Section 3.2. Eight expenditure categories are divided into twelve groups of varied sizes. Two different forms of groups are generated based on the arrangement of their components. Groups 2, 3, 4, and 5 contain sub-indices from a variety of categories, with no clear dominance. In contrast, the majority of the series in groups 1 and 8, for example, belong to a certain category. Group 1 may refer to a *Food* group, whereas group 8 is a *Transportation* group. The detailed group 8 components are depicted in Figure 12. There are seven sub-indices from *Transportation*, including car and truck rentals, gasoline (regular, midgrade, and premium), other motor fuels, airline fares, and ship fares, and one series from *Housing* - fuel oil (for residential heating). Clearly, all sub-indices share a common trend and have a close relationship with energy and oil prices, which have increased since

Figure 10: Posterior Similarity Matrices for Selected Categories

(a) Food and beverages (Category 3)

(b) Transportation (Category 7)

Notes: This is not the direct output of the algorithm, the ordering of series is changed so that cluster are lying on the diagonals.

the Pandemic. This is an example demonstrating that our proposed algorithm exploits cross-sectional information, not limited to our prior knowledge, and forms meaningful groups for forecasting.

Figure 11: Posterior Point Estimate of the Group Partition, August 2022
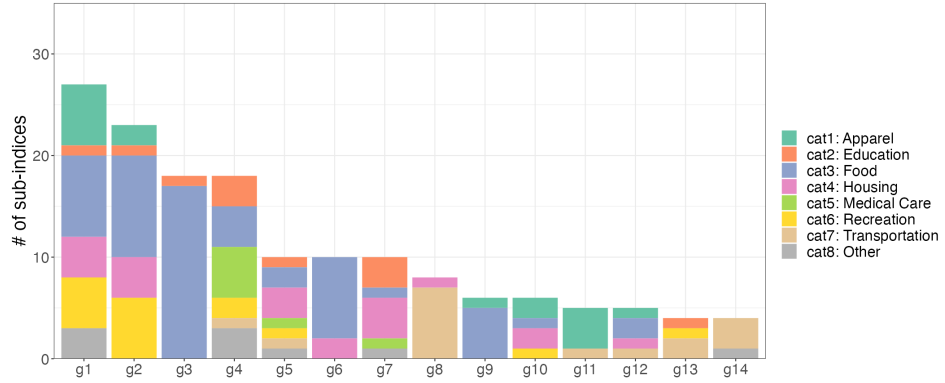
4.2.3 Impact of the Accuracy of Constraints

We examine how the accuracy of pairwise constraints influences the point estimate of group partitioning. For demonstration purposes, we restrict our analysis to PL constraints solely by setting $\psi_{ij}^{NL} = 0.5$ and changing ψ_{ij}^{PL}. We do not select the constant c in the setup since it would balance the impact of the PL restrictions with a different level of accuracy. We set c to 0.5. Again, PL constraints are derived from the official expenditure categories, with the

34

Figure 12: Components in the Group #9, August 2022
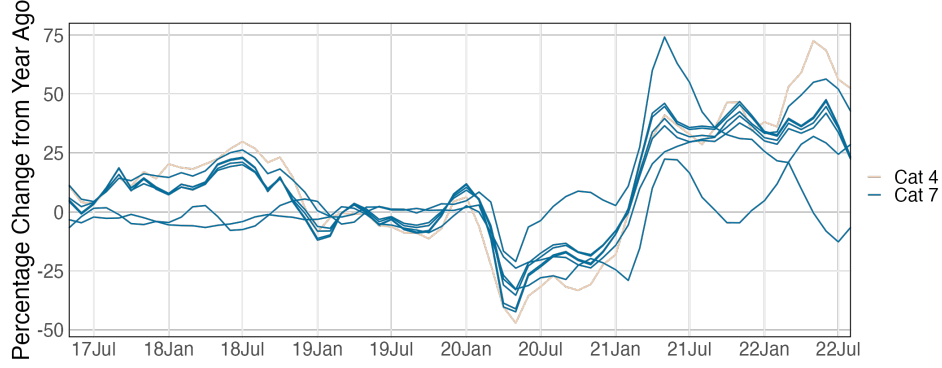
assumption that all units within the same category are positive-linked with equal probability of being in the same group.

Figure 13: Impact of the Strength of Constraints

(a) Weaker PL $\left(\psi_{ij}^{PL} = 0.55 \right)$ (b) Stronger PL $\left(\psi_{ij}^{PL} = 0.95 \right)$
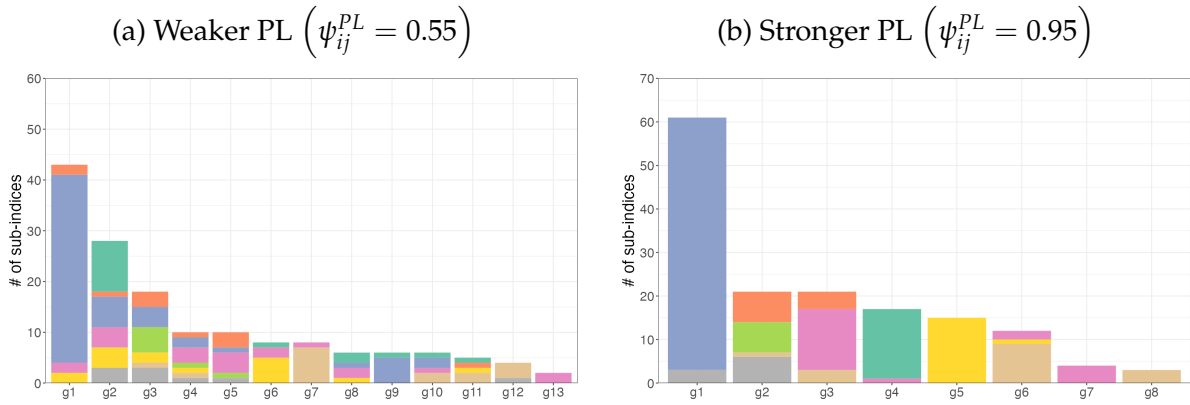
Figure 13 presents the point estimates of the group structure with two different levels of accuracy. The "weaker" PL constraints with $\psi_{ij}^{PL} = 0.55$, as shown in panel (a), demonstrate a limited influence of prior knowledge on the group structure. The eleven groups are composites of CPI sub-indices from the various categories, which is diverse from the official spending categories. Panel (b), on the other hand, illustrates the group structure with "stronger" PL constraints. By setting a high level of accuracy for PL constraints, such as $\psi_{ij}^{PL} = 0.95$, the prior knowledge dominates and pushes the group structure towards the official expenditure categories. As anticipated, panel (b) shows fewer groups, and the majority of CPI sub-indices within each group belong to the same category, bringing the group structure closer to that of the prior.

4.3 Income and Democracy

It is well-known in the literature that income per capita is strongly correlated with the level of democracy across countries. This strong empirical regularity is often known as "modernization theory" or Lipset hypothesis (Lipset, 1959). The theory claims a causal relation: democratic regimes are created and consolidated in affluent societies (Lipset, 1959; Przeworski et al., 1995; Barro, 1999; Epstein et al., 2006).

In an influential paper, Acemoglu et al. (2008) challenge the casual effect of countries' income on the level of democracy. They argue that it is essential to take into account other factors that affect both economic and political development simultaneously. Their analysis, based on panel data, indicates that the positive relationship between income and democracy disappears when fixed effects are included in the regression. They suggest that this finding is due to historical events, such as the end of feudalism, industrialization, or colonization, which have led countries to follow distinct paths of development. The fixed effects are meant to capture these persistent events. The finding is robust, as it holds for different measures of democracy, various econometric specifications, and additional covariates. Another study by Bonhomme and Manresa (2015) uses a different econometric model but arrives at the same conclusion. Their analysis highlights the presence of diverse group-specific paths of democratization in the data, consistent with the observation that regime types and transitions tend to cluster in time and space (Gleditsch and Ward, 2006; Ahlquist and Wibbels, 2012).

The seminal work of Acemoglu et al. (2008) has been subject to critical scrutiny by several recent works, including Moral-Benito and Bartolucci (2012), Benhabib et al. (2013), and Cervellati et al. (2014). Moral-Benito and Bartolucci (2012) contend that a nonlinear relationship between income and democracy exists, even after accounting for country-specific effects. They show that a positive income-democracy relationship holds only in countries with low levels of income. Benhabib et al. (2013) use panel estimation methods that adjust for the censoring of democracy measures at their upper and lower bounds and find that the positive relationship between income and democracy withstands the inclusion of country fixed effects. Cervellati et al. (2014) extend the linear estimation framework of Acemoglu et al. (2008) and unveil the presence of significant heterogeneity in the income effect on democracy across different subsamples. Specifically, they demonstrate that this effect exhibits an opposite sign for colonies and non-colonies, is substantially different from zero, and is of considerable magnitude. They also argue that the existence of a heterogeneous effect of income suggests that results from a linear framework, such as the finding of a zero effect, may lack robustness since they depend on the composition of the sample.

In this section, we contribute to the literature on the relationship between income and democracy by employing a novel grouped fixed-effects approach. Specifically, we expand on the econometric model proposed by Bonhomme and Manresa (2015) to incorporate group structure not only in time fixed-effects, but also in slope coefficients and the variance of errors. This more complex model allows for a more detailed analysis of the heterogeneous ef-

fects of income on democracy across different groups of countries. To identify these groups, we incorporate prior knowledge about the latent group structure by clustering countries based on either geographic location or initial levels of democracy score. By leveraging this information, we are able to identify a moderate number of groups, each of which exhibits a distinct path to democracy.

Our results indicate that the effect of income on democracy is highly varied across countries, a finding which is consistent with previous research by Cervellati et al. (2014). Furthermore, we find that the positive cumulative effect of income on democracy exists in groups of countries with a medium or relative low level of income, in line with the findings of Moral-Benito and Bartolucci (2012). However, the effect could be relatively small for some groups, suggesting that other factors beyond income may also play important roles in democratic development.

4.3.1 Model Specification and Data

Model: To accommodate richer assumptions on models for the real-world applications, we extend the baseline model in Chapter 1 either by adding common regressors and allowing for time-variation in the fixed-effects. Time-variation are essential to this analysis as they capture highly persistent historical shocks. Following Bonhomme and Manresa (2015), we introduce group-specific time patterns of heterogeneity $\alpha_{g_i t}$ and consider the following two specifications:

SP1: *Time-varying GFE + grouped slope coefficients*

$$y_{it} = \alpha_{g_i t} + \rho_{g_i} y_{it-1} + \beta_{g_i} x_{it-1} + \varepsilon_{it}, \ \varepsilon_{it} \sim N(0, \sigma_{g_i}^2), \tag{4.14}$$

SP2: *Time-varying GFE*

$$y_{it} = \alpha_{g_i t} + \rho y_{it-1} + \beta x_{it-1} + \varepsilon_{it}, \ \varepsilon_{it} \sim N(0, \sigma_{g_i}^2), \tag{4.15}$$

where y_{it} is the democracy score of country i at time t. The lagged value of democracy score y_{it-1} is included to capture persistence in democracy and also potentially mean-reverting dynamics (i.e., the tendency of the democracy score to return to some equilibrium value for the country). The coefficient of main interest is β_{g_i} and reflects the effect of the lagged value of log income per capita x_{it-1} on democracy. In addition, $\alpha_{g_i t}$ denote a set of group-specific time fixed-effects; ε_{it} is an error term with grouped variance $\sigma_{g_i}^2$, capturing additional transitory shocks to democracy and other omitted factors. We use the conjugate prior for all parameters, see details in Appendix B.

Specification 2 in (4.15) nests the linear dynamic panel data model in BM as a special case. If we assume homoskedasticity, it is the Equation (22) in BM. This specification enables us to reproduce BM's results and provide fresh insight into their framework. Specification

1 in (4.14), on the other hand, generalizes specification 2 by introducing group-dependent slope coefficients. As we shall demonstrate in the following section, specification 1 yields a more refined group structure and provides a clearer view of the income effects.

Data: We use the Freedom House (FH) Political Rights Index as a benchmark for measuring democracy. To standardize the index, we normalize it between 0 and 1, with higher scores indicating higher levels of democracy. FH assesses a country's political rights based on a checklist of questions, such as the presence of free and fair elections, the role of elected officials, the existence of competitive parties or other political groupings, the power of the opposition, and the extent of self-government or participation by minority groups. We measure countries' income using the logarithm of GDP per capita, which is adjusted for purchasing power parity (PPP) in 1996 prices, using data from the Penn World Tables 6.1 (Heston et al., 2002). Details on the data can be found in Section 1 of Acemoglu et al. (2008), and all data in this section are from the replication files of Bonhomme and Manresa (2015).[15] Our analysis is based on a five-year panel dataset that includes all independent countries since the postwar period, with observations taken every fifth year from 1970 to 2000. We chose this period for comparability with previous studies. The final dataset consists of a balanced panel of 89 countries.

Prior group structure: We propose two prior grouping strategies as specified below and assume all units within the same prior group are presumed to be positive-linked, while units from different prior groups are believed to be negative-linked.

(i) Given the countries available in the dataset, we form six groups according to their geographic locations:[16] (1) North America; (2) Europe; (3) Latin America and the Caribbean; (4) Asia and Australasia; (5) Sub-Saharan Africa; (6) Middle East and North Africa. We refer this prior to *geo-prior*.

(ii) Alternately, countries could be categorized according to their initial level of democracy in year 1970. As the Freedom House Index has six possible values, we cluster countries into three primary groups with a reasonable number of countries in each: (1) low democracy, $y_{i,1970} = 0$ or 0.166; (2) medium democracy, $y_{i,1970} = 0.333$, 0.5, or 0.667; (3) high democracy, $y_{i,1970} = 0.833$ or 1. We refer this prior to *dem-prior*.

Figure 14 presents the world maps with countries colored differently according to their respective groups. The panel on the left illustrates the geographic groups, while the panel on the right depicts the democratic groups. All gray nations/regions are excluded from the

---

[15]https://www.dropbox.com/s/ssjabvc2hxa5791/Bonhomme_Manresa_codes.zip?dl=0

[16]The regions are assigned by the Economist Intelligence Unit, and may slightly differ from conventional classifications.

dataset. We concentrate primarily on the first pre-grouping strategy, as it needs no country-specific knowledge beyond geographic information. We then compare the results using different pre-grouping strategies in Section 4.3.3.

Figure 14: Specifications of Prior Grouping

(a) Geographic Location (Geo-Prior)   (b) Initial Level of Democracy (Dem-Prior)
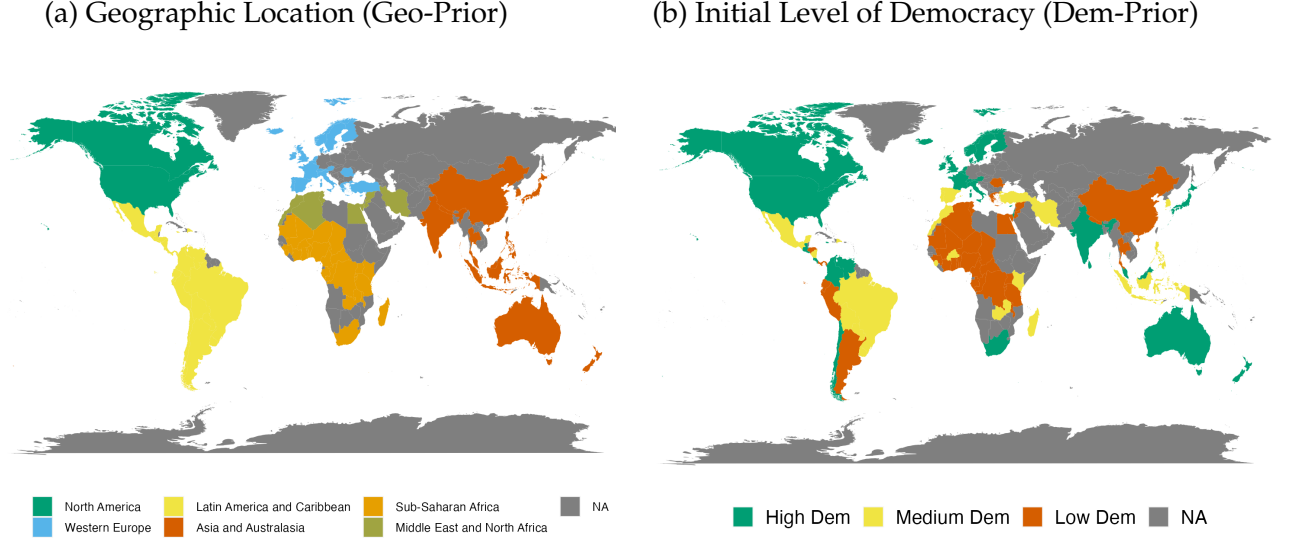
4.3.2 Results

Specification 1: We begin with specification 1 where group-specific slope coefficients are allowed and new findings emerge. Table 1 presents the posterior probability of the number of groups utilizing various estimators. BGFE-ho creates more than 5 groups in all posterior draws. Intriguingly, accounting for heteroskedasticity drastically reduces the number of groups, with BGFE-he identifying four groups. Adding pairwise constraints based on geographic information increases the number of groups to five, whereas six groups are expected in the prior.

Table 1: Probability for Number of Groups, Geo-Prior

	BGFE-he-cstr	BGFE-he
$Pr(K < 4)$	0.000	0.000
$Pr(K = 4)$	0.000	**1.000**
$Pr(K = 5)$	**1.000**	0.000
$Pr(K > 5)$	0.000	0.000

The marginal data density (MDD) of each estimator in Table 2 provides some insight on different models. Among all the estimators, the BGFE-ho estimator has the lowest MDD; it is even lower than that of specification 1. BGFE-he-cstr and BGFE-he, on the other hand,

39

benefit from the introduction of group-specific slope coefficients, since both achieve substantially greater MDD than in specification 1. BGFE-he-cstr has the highest MDD since the pairwise constraints give direction on grouping and identify the ideal group structure, which BGFE-he cannot uncover without our prior knowledge.

Table 2: Marginal Data Density, Geo-Prior

	BGFE-he-cstr	BGFE-he	BGFE-ho
SP1	544.324	501.904	327.077
SP2	413.476	381.218	368.918

We concentrate on the BGFE-he-cstr estimator and use the approach outlined in Section 3.2 to identify the unique group partitioning $\hat{G}$. The left panel of Figure 15 presents the world map colored by $\hat{G}$, while the right panel present the group-specific averages of democracy index over time. The estimated group structure $\hat{G}$ features five distinct groups which we refer to as the "high-democracy", "low-democracy", "flawed-democracy", "late-transition" and "progressive-transition" group, respectively. With the exception of the "flawed-democracy" and "progressive-transition" group, the group-specific averages of the democracy index are comparable to those in BM for all other groups. BGFE-he-cstr does not identify the "early transition" group in comparison to BM but instead produces two new groups. Group 3 ("flawed-democracy") comprises primarily of relatively democratic but not the most democratic nations, including India, Sri Lanka, and Venezuela, among others. Group 5 ("progressive transition") contains 30 countries that have had a steady expansion of democracy, including Argentina, Greece, and Panama. Consequently, by incorporating group-specific slope coefficients, we recover a more refined group structure than that of BM.

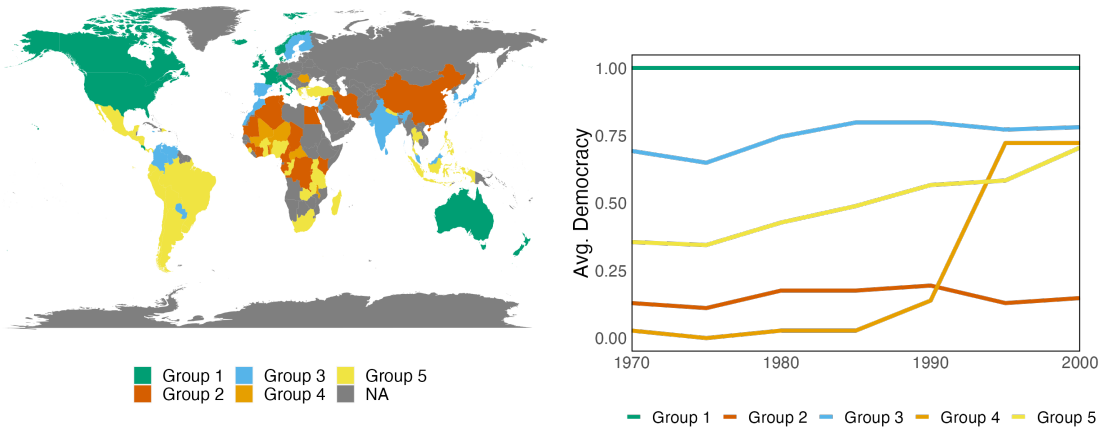Figure 15: Posterior Point Estimate of Group Partitioning and Average Democracy

Table 3 presents the posterior mean and 90% credible set for each coefficient across all groups, with G fixed at the point estimate $\hat{G}$. The key feature of using the specification 1 is

that we are able to see distinct (cumulative) income effects across groups as group-specific coefficients are allowed.

The effect of income on democracy is negligible for group 1 ("full-democracy") and group 4 ("late-transition") as the posterior means of β are close to 0 and the associated credible intervals for $\hat{\beta}$ contain 0. Group 1, which we refer to as the "full-democracy" group, mostly contains high-income, high-democracy countries. It includes the United States, Canada, UK, most of European countries, Australia, and New Zealand, but also Costa Rica and Uruguay. These country kept their democracy index at the highest level throughout the sample, demonstrating that income has no effect on democracy. Group 4 is referred to as the "late-transition" group, which consists of Benin, Central African Republic, Mali, Malawi, Niger, and Romania. The transition to democracy for countries in group 4 was primarily driven by historical events in the 90s: Romania began a transition towards democracy after the 1989 Revolution; all other countries involved in the third wave of democratization in sub-Saharan Africa beginning in 1989. The impact of historical events is primarily captured by time fixed-effects, as the credible intervals for $\hat{\beta}$ and $\hat{\rho}$ well cover zero.

Group 2 ("low-democracy") and group 3 ("flawed-democracy") are two groups that have stable Freedom House scores. Lagged democracy is highly significant and indicates that there is a considerable degree of persistence in democracy. Log income per capita is also significant and illustrates the well-documented positive relationship between income and democracy. Though statistically significant, the effect of income is quantitatively small. For example, the coefficient of 0.055 for the group 3 implies that a 10 percent increase in GDP per capita is associated with an increase in the Freedom House score of 0.0055, which is very small. Group 2 includes low-democratic countries, China, Singapore, Iran, and a fraction of African countries, whose Freedom House scores remain relatively low throughout the sample. Countries in group 3, however, have Freedom House score stay in the relatively high level. The group covers countries with almost democratic but minor flaws in certain aspects, including India, Japan, South Korea, Finland, Sweden, and Portugal, among others. The cumulative income effects for these two groups are, however, different - it is negligible for group 2 (0.079) and modest for group 3 (0.244). Group 5 ("progressive-transition"), on the other hand, experiences a continuous increase in Freedom House score from a 0.35 in 1970 to 0.7 in 2020. It has the largest positive income coefficient, although the cumulative income effect is modest (0.156).

Figure 16 depicts the historical average log income for each group of countries between 1970 and 2000. Except for the high income group (group 1) and low income group (group 4), all other three groups with medium or relatively low levels of income reveal positive cumulative income effect, as indicated in Table 3. This finding is generally in line with the results of Moral-Benito and Bartolucci (2012), who observe a positive effect in low-income nations. Notice that, their definition of low-income country is quite board, encompassing countries with a GDP per capita below the 80th percentile of the empirical cross-sectional

41

Table 3: Coefficient Estimates across Groups, Geo-Prior

	Lagged democracy (ρ)		Lagged Income (β)		Income Effect ($\beta/(1-\rho)$)		Error variance (σ^2)	
	Coef.	Cred. Set	Coef.	Cred. Set	Coef.	Cred. Set	Coef.	Cred. Set
Group 1 (16)	0.058	[-0.263, 0.360]	0.000	[-0.012, 0.012]	0.000	[-0.013, 0.013]	0.001	[0.001, 0.001]
Group 2 (18)	0.484	[0.354, 0.606]	0.041	[0.019, 0.062]	0.079	[0.044, 0.117]	0.010	[0.008, 0.011]
Group 3 (19)	0.775	[0.703, 0.850]	0.055	[0.031, 0.078]	0.249	[0.143, 0.352]	0.013	[0.011, 0.016]
Group 4 (6)	-0.178	[-0.468, 0.115]	-0.025	[-0.066, 0.017]	-0.020	[-0.054, 0.015]	0.008	[0.005, 0.011]
Group 5 (30)	0.206	[0.091, 0.310]	0.125	[0.090, 0.163]	0.157	[0.118, 0.199]	0.057	[0.048, 0.066]
Pooled OLS	0.667	[0.614, 0.717]	0.081	[0.064, 0.099]	0.244	[0.207, 0.281]	0.039	[0.035, 0.043]

density.. As a result, beside six counties in the group 4 that had the lowest income on average, other countries fit within this definition and confirm that the positive income effect is not prevalent in high-income countries.

Figure 16: Average Log Income by Groups
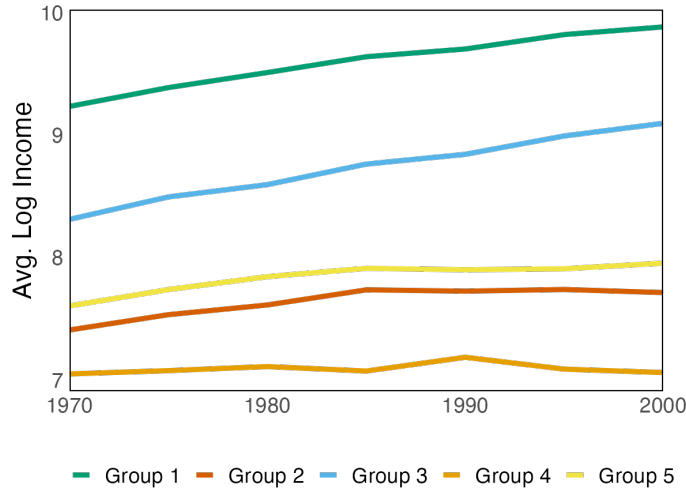
Specification 2: The results for specification 2 are reported in Appendix G.2. In short, the results are comparable to the key findings in BM. BGFE-ho in specification 1 is identical to the main model in BM; it produces eight groups, which is consistent with the upper bound on the number of groups in BM based on BIC. BGFE-he-cstr, on the other hand, is more preferable and has the highest marginal data density as shown in Table 2. The point estimate of group partitioning based on BGFE-he-cstr consists of four groups that all have the similar pattern as BM's group structure. This justifies BM's subjective choice of four groups. Regarding the estimated coefficients, there is moderate persistence and a positive effect of income on democracy, but the cumulative effect of income is quantitatively small: $\beta/(1-\rho) = 0.08$.

4.3.3 Impacts of Different Pre-Grouping Strategies

All results presented thus far are based on pairwise constraints derived from spatial information. We now implement the alternative pre-grouping strategy based on the initial level of democracy. We stick with the BGFE-he-cstr estimator under specification 1.

Different pre-grouping strategies yields different estimates of group patterns. We consider the BGFE-he estimator with three prior group structures: geo-prior, dem-prior, and no prior knowledge, with point estimates of the group partition presented in panel (a), (b), and (c) of Figure 17, respectively. Geo-prior and dem-prior produce comparable group structures, however certain nations are assigned to distinct groups.. They are encircled in the black dashed rectangle, including Portugal, Spain, Romania, Mali, Niger, Central African Republic, Benin, Malawi, and Jordan. Another country is South Korea. As depicted in panel (c), without any prior knowledge of groups, the group pattern is quite different, particularly for countries in Asia, Africa, and Latin America.

The difference in group patterns result in discrepancies in MDD, which are listed in Table 4. The BGFE-he estimator with geo-prior has the highest MDD in both specifications, whereas the dem-prior is only informative in specification 1 in comparison to the BGFE-he estimator without prior knowledge.

Table 4: Marginal Data Density, Different Priors

	geo-prior	dem-prior	no prior
SP1	544.324	527.517	501.904
SP2	413.476	309.369	381.218

Using the initial level of democracy as prior knowledge results in four groups, as indicated in the panel (b). The dem-prior has two major impacts on the group structure comparing with the geo-prior. It combines the "late-transition" group (group 4 in geo-prior) with the "progressive-transition" group (group 5 in geo-prior) to form a bigger and boarder "progressive-transition" group. Additionally, Portugal and Spain are no longer categorized as "flawed-democracy" countries, but rather as "progressive-transition" group in a boarder sense. In terms of the posterior estimates, as shown in Table 5, we observe similar value for the first three groups because they are merely subject to minor changes in the group structure. However, as "late-transition" and "progressive-transition" groups are agglomerated together under the dem-prior, the income effects of countries in the new group becomes larger, forcing some countries to exhibit strong and positive effects even though they are not under the geo-prior.

Figure 17: Posterior Point Estimate of Group Structure

(a) Geo-Prior

(b) Dem-Prior

(c) No Prior Knowledge
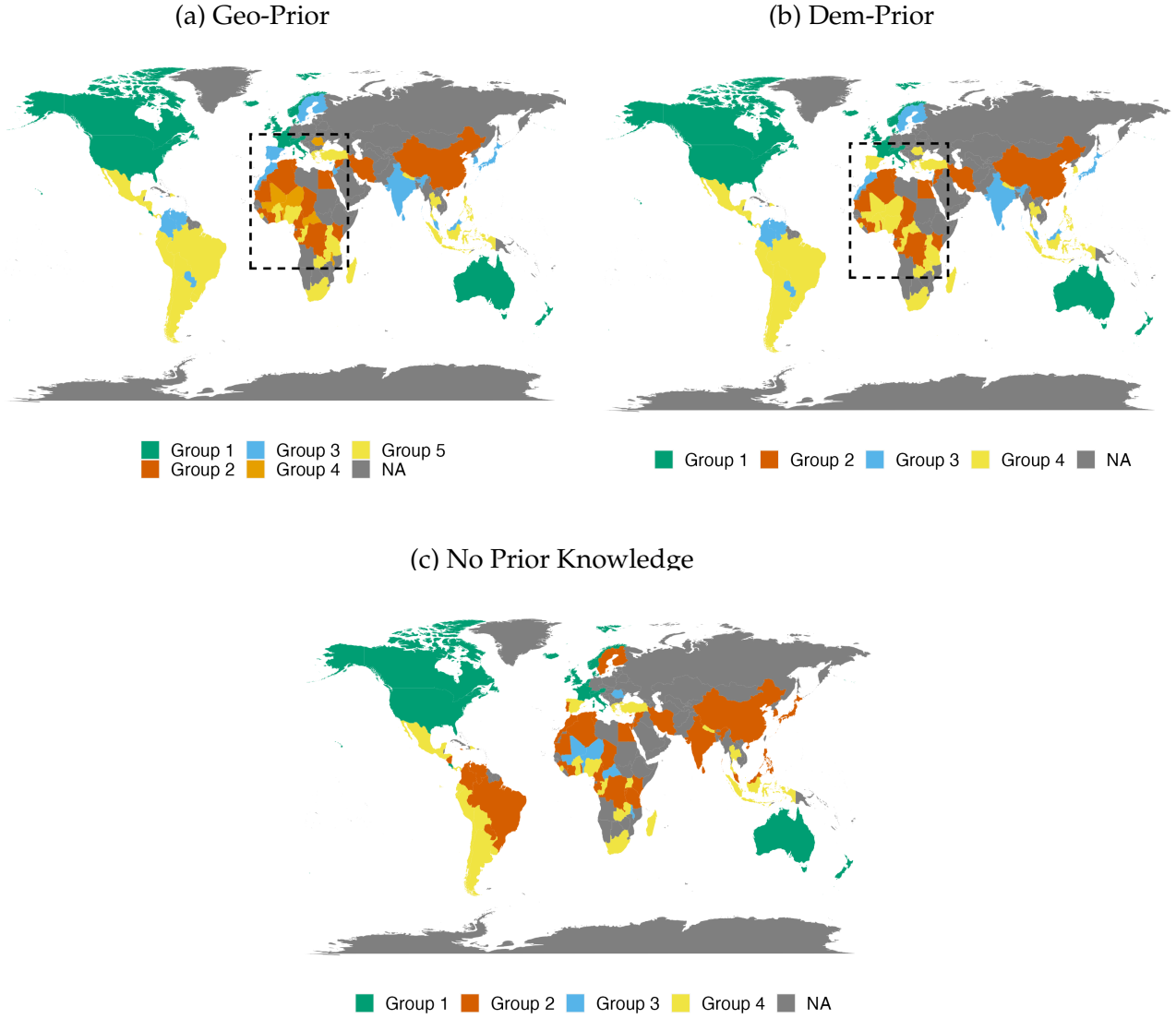
Table 5: Coefficient Estimates across Groups, Dem-Prior

	Lagged democracy (ρ)		Lagged Income (β)		Income Effect ($\beta/(1-\rho)$)		Error variance (σ^2)	
	Coef.	Cred. Set	Coef.	Cred. Set	Coef.	Cred. Set	Coef.	Cred. Set
Group 1 (16)	0.062	[-0.262, 0.369]	0.000	[-0.012, 0.012]	0.000	[-0.014, 0.013]	0.001	[0.001, 0.001]
Group 2 (19)	0.513	[0.395, 0.640]	0.043	[0.022, 0.064]	0.087	[0.047, 0.126]	0.010	[0.008, 0.012]
Group 3 (15)	0.802	[0.704, 0.904]	0.040	[0.016, 0.064]	0.223	[0.079, 0.337]	0.012	[0.009, 0.015]
Group 4 (39)	0.302	[0.205, 0.388]	0.120	[0.092, 0.151]	0.172	[0.136, 0.210]	0.054	[0.047, 0.062]
Pooled OLS	0.667	[0.614, 0.717]	0.081	[0.064, 0.099]	0.244	[0.207, 0.281]	0.039	[0.035, 0.043]

5 Extensions

Within the domain of panel data models, the proposed constrained-based BGFE framework can be extended in multiple directions to allow for more subtle group structures or more covariates. In addition, the DP prior with soft pairwise constraints also applies to other related topics and models, such as clustering problems, heterogeneous treatment effects, and panel VARs.

5.1 Subtle Group Structure

Through the Dirichlet process defines a prior that possesses the clustering property and is flexible enough to incorporate pairwise constraints, the group structure itself is elementary. Aside from our prior belief on the group, the group structure, which is introduced in all α_i and σ_i^2, is entirely governed by the stick-breaking process defined in Equation (2.13). The stick length ξ_k, on which we have a prior, is independent of any regressors or time. Consequently, each unit is associated with a single group, and the membership remains constant across time.

To create an even more flexible and richer group structure, we provide insight into three possible extensions, each of which requires a set of more distinctive nonparametric priors. (1) overlapping group and (2) time-varying group and (3) dependent group.

Overlapping group structures allow for multi-dimensional grouping. This is a natural extension without having to greatly modify the proposed DP prior. Following Cheng et al. (2019), each of α_i's and σ_i^2 may have its own group structure and a separate Dirichlet process is specified to each of them. As a result, units simultaneously belong to multiple groups based on the heterogeneous effects among regressors or cross-sectional heteroskedasticity.

Time-varying group structures allow the membership of the group to change over time. We could replace the DP by variants of the hierarchical Dirichlet process (Teh et al., 2006) to achieve this feature. In short, the hierarchical Dirichlet process (HDP), a nonparametric Bayesian approach to clustering grouped data, is now the foundation of the prior. The time dimension naturally divides the panel data into T groups, and a Dirichlet process is assumed for each group, with all Dirichlet processes having the same base distribution, which is distributed according to a global base distribution. The HDP allows each group to have its own cluster, but most importantly, these clusters are shared across groups. This lays the groundwork for time-varying group structures, as it assumes that the number of clusters remains constant over time, while cluster memberships are subject to change. Variants of the HPD are then proposed to capture the time-persistence in group structures, including dynamic HDP (Ren et al., 2008) and sticky HDP (Fox et al., 2008, 2011). A closely related area in the frequentists' methods is to identify structure breaks in parameters with grouped patterns, see Okui and Wang (2021); Lumsdaine et al. (2022).

Dependent group structures allow the prior group probability to rely directly on a collection of characteristics. The dependence is introduced through a modification of the stick-breaking representation for DPs, where the group probabilities vary with the characteristics. Rodriguez and Dunson (2011) introduced the probit-stick breaking (PSB) process where the Beta random variables are replaced by normally distributed random variables transformed using the standard normal CDF. The PSB is defined by,

$$\pi_k(w_i) = \Phi\left(\zeta_k(w_i)\right) \prod_{j<k} \left[1 - \Phi\left(\zeta_j(w_i)\right)\right], \tag{5.1}$$

where stochastic function ζ_k is drawn from Gaussian process $\zeta_k \sim GP(0, V_k)$ for $k = 1, 2, \cdots$ and w_i is the set of characteristics that are informative to the latent group. Other forms of dependence are also available, see Quintana et al. (2022) for a comprehensive review. A caveat of this approach is that analysis of group structure is confined to w_i observed by the researcher. The approach requires researchers to know possible key characteristics, be able to observe them and ensure they are informative. In many cases, however, these characteristics might be hard to justify by researchers.

5.2 Beyond Panel Data Models

Although we concentrate on panel data model, our framework of the DP prior with soft pairwise constraints applies to other models where the group structure are crucial.

Gaussian Mixture Model

If we ignore covariates and focus exclusively on group membership, we essentially face a classical clustering problem with an infinite-dimensional mixture model. A typical probabilistic model is the infinite Gaussian mixture model (Rasmussen, 1999), where the data itself is assumed to be drawn from a mixture of Gaussian components

$$y_i \sim \sum_{k=1}^{\infty} \pi_k N(\mu_k, \Sigma_k), \tag{5.2}$$

where π_k are the mixture weights. With soft pairwise constraints, observations are clustered in accordance with prior belief.

Heterogeneous Treatment Effects

Following the potential outcomes framework of Rubin (1974), we posit the existence of potential outcomes $y_i(1)$ and $y_i(0)$ corresponding respectively to the response the i th subject would have experienced with and without the treatment, and define the treatment effect at x as

$$\tau(x) = \mathbb{E}\left[y_i(1) - y_i(0)|x_i = x\right]. \tag{5.3}$$

Previous works on Bayesian analysis for the treatment effects include Chib and Hamilton (2000, 2002); Chib and Jacobi (2007); Chib (2007); Heckman et al. (2014), etc. Another strand of literature propose to estimate (5.3) by using several machine learning algorithms (Hill, 2011; Athey and Imbens, 2016; Wager and Athey, 2018). These methods are built on the idea that researchers find the subsamples across which the effect of a treatment differs out of all possible subsamples on the basis of the values of x_i. Instead of trying to discover valid subsets of the data, Shiraito (2016) directly models the outcome as a function of the treatment and pre-treatment covariates x_i and estimate of the distribution of conditional average treatment effects (CATE) across units by employing the Dirichlet process.

$$y_i = \tau'_{g_i} T_i + \gamma'_{g_i} x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_{g_i}\right), \tag{5.4}$$

where T is the binary treatment variable. Our approaches fit in this methods by adding side information on the treatment groups into the prior.

Panel VARs

Panel VARs (Holtz-Eakin et al., 1988; Canova and Ciccarelli, 2013, and references therein.) has been widely used in macroeconomic analysis and policy evaluations to capture the interdependency across sectors, markets, and countries. Nevertheless, the large dimension of panel VARs typically makes the curse of dimensionality a severe problem. Billio et al. (2019) propose nonparametric Bayesian priors that cluster the VAR coefficients and induce group-level shrinkage. Our paradigm with the DP prior with soft pairwise constraints is applicable to their method and injects prior information on groups into the underlying Granger causal networks.

Panel VARs have the same structure as VAR models, in the sense that all variables are assumed to be endogenous and interdependent, but a cross-sectional dimension is added to the representation. Thus, let Y_t be the stacked version of y_{it}, the vector of J variables for each unit $i = 1, \ldots, N$, i.e., $Y_t = (y'_{1t}, y'_{2t}, \ldots y'_{Nt})'$. Then a panel VAR is

$$Y_t = A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + u_t, \quad i = 1, \ldots, N, \tag{5.5}$$

where u_t is a $J \times 1$ vector of idiosyncratic errors and A_0 and A_j are $NJ \times NJ$ matrices of coefficients.

The main feature of Billio et al. (2019) is to specify a prior that blends the DP prior with Lasso prior for each of A_0 and A_j, such that the VAR coefficients are either shrunk toward 0 or clustered at multiple non-zero locations. Our proposed DP prior with soft pairwise constraints, in the meantime, fit into their framework by replacing the original DP prior and permitting richer structure within each coefficient matrix. As the nonzero coefficients form Granger causal networks, equipping with soft pairwise constraints may result in a more plausible network by taking researchers' expertise into account.

6 Concluding Remarks

This paper proposes a Bayesian framework for estimating and forecasting in panel data models when prior group knowledge is available and informative for the group pattern. We include prior knowledge in the form of soft pairwise constraints into the Dirichlet process prior. Then, an intuitive and coherent prior is presented. The constrained grouped estimator proposed examines both heteroskedasticity and heterogeneous slope coefficients to endogenously reveal group structure. Our framework immediately estimates the number of groups as opposed to relying on ex-post model selection, and the structure of pairwise restrictions circumvents the computational difficulties and limitations that afflict conventional approaches. In addition, when utilizing small-variance asymptotics, the suggested Gibbs sampler with pairwise constraint contains a clustering procedure comparable to that of the constrained *KMeans* algorithm. In Monte Carlo simulations, we demonstrate that constrained Bayesian grouped estimators outperform conventional estimators even in the presence of incorrect prior knowledge. Our empirical application to forecasting sub-indices of CPI inflation rates demonstrates that incorporating prior knowledge on the latent group structure yields more accurate density predictions. The better forecasting performance is mostly attributable to the key characteristics: nonparametric Bayesian prior and grouped cross-sectional variance. The method proposed in this paper is applicable beyond forecasting. In a second application, we revisit the relationship between a country's income and its democratic transition, where estimation of heterogeneous parameters is the object of interest. We recover a reasonable cluster pattern with a moderate number of groups and identify heterogeneous income effects on democracy.

The current work raises exciting questions for future research. It is desirable to investigate overlapping group structures, in which a unit might belong to many groups. This would allow us to increase the flexibility of a panel data model, potentially enhancing its predictive performance. Second, the assumption that an individual cannot change its group identity for the entire sample time can be amended, resulting in a specification that is even more flexible. Thirdly, our method is applicable to other econometric models, such as panel VARs with latent group structures in macro series.

References

ACEMOGLU, D., S. JOHNSON, J. A. ROBINSON, AND P. YARED (2008): "Income and Democracy," *American Economic Review*, 98, 808–42.

AGUILAR, J. AND T. BOOT (2022): "Grouped Heterogeneity in Linear Panel Data Models with Heterogeneous Error Variances," *Available at SSRN 4031841*.

AHLQUIST, J. S. AND E. WIBBELS (2012): "Riding the Wave: World Trade and Factor-Based Models of Democratization," *American Journal of Political Science*, 56, 447–464.

ALDOUS, D. J. (1985): "Exchangeability and Related Topics," in *École d'Été de Probabilités de Saint-Flour XIII—1983*, Springer, 1–198.

ANDERSON, T. W. AND C. HSIAO (1982): "Formulation and Estimation of Dynamic Models using Panel Data," *Journal of Econometrics*, 18, 47–82.

ANDO, T. AND J. BAI (2016): "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Journal of Applied Econometrics*, 31, 163–191.

ANTONIAK, C. E. (1974): "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

ASCOLANI, F., A. LIJOI, G. REBAUDO, AND G. ZANELLA (2022): "Clustering Consistency with Dirichlet Process Mixtures," *arXiv preprint arXiv:2205.12924*.

ATHEY, S. AND G. IMBENS (2016): "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

ATKESON, A., L. E. OHANIAN, ET AL. (2001): "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 25, 2–11.

BARRO, R. J. (1999): "Determinants of Democracy," *Journal of Political economy*, 107, S158–S183.

BASU, S., A. BANERJEE, AND R. MOONEY (2002): "Semi-Supervised Clustering by Seeding," in *Proceedings of 19th International Conference on Machine Learning*, Citeseer.

BASU, S., A. BANERJEE, AND R. J. MOONEY (2004a): "Active Semi-Supervision for Pairwise Constrained Clustering," in *Proceedings of the 2004 SIAM international conference on data mining*, SIAM, 333–344.

BASU, S., M. BILENKO, AND R. J. MOONEY (2004b): "A Probabilistic Framework for Semi-Supervised Clustering," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 59–68.

BENHABIB, J., A. CORVALAN, AND M. M. SPIEGEL (2013): "Income and Democracy: Evidence from Nonlinear Estimations," *Economics Letters*, 118, 489–492.

BERNANKE, B. S. (2007): "Inflation Expectations and Inflation Forecasting," in *Speech at the Monetary Economics Workshop of the National Bureau of Economic Research Summer Institute, Cambridge, Massachusetts*, vol. 10.

BILENKO, M., S. BASU, AND R. J. MOONEY (2004): "Integrating Constraints and Metric Learning in Semi-Supervised Clustering," in *Proceedings of the 21st International Conference on Machine Learning*, 81–88.

BILLIO, M., R. CASARIN, AND L. ROSSINI (2019): "Bayesian Nonparametric Sparse VAR Models," *Journal of Econometrics*, 212, 97–115.

BLACKWELL, D. AND J. B. MACQUEEN (1973): "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2022): "Discretizing Unobserved Heterogeneity," *Econometrica*, 90, 625–643.

BONHOMME, S. AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184.

BUCHINSKY, M., J. HAHN, AND V. HOTZ (2005): "Cluster Analysis: A Tool for Preliminary Structural Analysis," *Unpublished manuscript*.

CANOVA, F. AND M. CICCARELLI (2013): "Panel Vector Autoregressive Models: A Survey," in *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*, Emerald Group Publishing Limited.

CERVELLATI, M., F. JUNG, U. SUNDE, AND T. VISCHER (2014): "Income and Democracy: Comment," *American Economic Review*, 104, 707–719.

CHAMBERLAIN, G. (1980): "Analysis of Covariance with Qualitative Data," *The Review of Economic Studies*, 47, 225–238.

CHENG, X., F. SCHORFHEIDE, AND P. SHAO (2019): "Clustering for Multi-Dimensional Heterogeneity," *Manuscript*.

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

CHIB, S. (2007): "Analysis of Treatment Response Data without the Joint Distribution of Potential Outcomes," *Journal of Econometrics*, 140, 401–412.

CHIB, S. AND B. H. HAMILTON (2000): "Bayesian Analysis of Cross-Section and Clustered Data Treatment Models," *journal of Econometrics*, 97, 25–50.

——— (2002): "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models," *Journal of Econometrics*, 110, 67–89.

CHIB, S. AND L. JACOBI (2007): "Modeling and Calculating the Effect of Treatment at Baseline from Panel Outcomes," *Journal of Econometrics*, 140, 781–801.

CLARK, T. E. AND F. RAVAZZOLO (2015): "Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility," *Journal of Applied Econometrics*, 30, 551–575.

CYTRYNBAUM, M. (2021): "Blocked Clusterwise Regression," *arXiv preprint arXiv:2001.11130*.

DAVIDSON, I. AND S. RAVI (2005): "Clustering with Constraints: Feasibility Issues and the K-Means Algorithm," in *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, 138–149.

EPSTEIN, D. L., R. BATES, J. GOLDSTONE, I. KRISTENSEN, AND S. O'HALLORAN (2006): "Democratic Transitions," *American journal of political science*, 50, 551–569.

ESCOBAR, M. D. AND M. WEST (1995): "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

FAUST, J. AND J. H. WRIGHT (2013): "Forecasting Inflation," in *Handbook of Economic Forecasting*, Elsevier, vol. 2, 2–56.

FERGUSON, T. S. (1973): "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

——— (1974): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

FISHER, M. AND M. J. JENSEN (2022): "Bayesian Nonparametric Learning of How Skill is Distributed across the Mutual Fund Industry," *Journal of Econometrics*, 230, 131–153.

FOX, E. B., E. B. SUDDERTH, M. I. JORDAN, AND A. S. WILLSKY (2008): "An HDP-HMM for Systems with State Persistence," in *Proceedings of the 25th International Conference on Machine Learning*, 312–319.

——— (2011): "A Sticky HDP-HMM with Application to Speaker Diarization," *The Annals of Applied Statistics*, 5, 1020–1056.

FREEMAN, H. AND M. WEIDNER (2022): "Linear Panel Regressions with Two-Way Unobserved Heterogeneity," *arXiv preprint arXiv:2109.11911*.

FRUCHTERMAN, T. M. AND E. M. REINGOLD (1991): "Graph Drawing by Force-Directed Placement," *Software: Practice and Experience*, 21, 1129–1164.

GEWEKE, J. AND G. AMISANO (2010): "Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns," *International Journal of Forecasting*, 26, 216–230.

GHOSAL, S. AND A. VAN DER VAART (2017): *Fundamentals of Nonparametric Bayesian Inference*, vol. 44, Cambridge University Press.

GLEDITSCH, K. S. AND M. D. WARD (2006): "Diffusion and the International Context of Democratization," *International Organization*, 60, 911–933.

HAHN, J. AND H. R. MOON (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory*, 26, 863–881.

HAMILTON, J. D. (2018): "Why You Should Never Use the Hodrick-Prescott Filter," *Review of Economics and Statistics*, 100, 831–843.

HECKMAN, J. J., H. F. LOPES, AND R. PIATEK (2014): "Treatment Effects: A Bayesian Perspective," *Econometric Reviews*, 33, 36–67.

HENSELER, M. AND I. SCHUMACHER (2019): "The Impact of Weather on Economic Growth and its Production Factors," *Climatic Change*, 154, 417–433.

HERSBACH, H. (2000): "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, 15, 559–570.

HESTON, A., R. SUMMERS, B. ATEN, ET AL. (2002): "Penn world table version 6.1," *Center for International Comparisons at the University of Pennsylvania (CICUP)*, 18.

HILL, J. L. (2011): "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240.

HIRANO, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

HOLLAND, P. W., K. B. LASKEY, AND S. LEINHARDT (1983): "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137.

HOLTZ-EAKIN, D., W. NEWEY, AND H. S. ROSEN (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–1395.

HSIANG, S. (2016): "Climate Econometrics," *Annual Review of Resource Economics*, 8, 43–75.

ISHWARAN, H. AND L. F. JAMES (2001): "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

KAHN, M. E., K. MOHADDES, R. N. NG, M. H. PESARAN, M. RAISSI, AND J.-C. YANG (2021): "Long-Term Macroeconomic Effects of Climate Change: A Cross-Country Analysis," *Energy Economics*, 104, 105624.

KIM, J. AND L. WANG (2019): "Hidden Group Patterns in Democracy Developments: Bayesian Inference for Grouped Heterogeneity," *Journal of Applied Econometrics*, 34, 1016–1028.

KOOP, G. (2003): *Bayesian Econometrics*, John Wiley & Sons.

KULIS, B. AND M. I. JORDAN (2011): "Revisiting K-Means: New Algorithms via Bayesian Nonparametrics," *arXiv preprint arXiv:1111.0352*.

LAIO, F. AND S. TAMEA (2007): "Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables," *Hydrology and Earth System Sciences*, 11, 1267–1277.

LAW, M. H., A. TOPCHY, AND A. K. JAIN (2004): "Clustering with Soft and Group Constraints," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, 662–670.

LEE, C. AND D. J. WILKINSON (2019): "A Review of Stochastic Block Models and Extensions for Graph Clustering," *Applied Network Science*, 4, 1–50.

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, 42–55.

LIPSET, S. M. (1959): "Some Social Requisites of Democracy: Economic Development and Political Legitimacy," *American Political Science Review*, 53, 69–105.

LIU, L. (2022): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *Journal of Business & Economic Statistics*, 0, 1–15.

LIU, L., H. R. MOON, AND F. SCHORFHEIDE (2020): "Forecasting with Dynamic Panel Data Models," *Econometrica*, 88, 171–201.

LU, Z. (2007): "Semi-Supervised Clustering with Pairwise Constraints: A Discriminative Approach," in *Artificial Intelligence and Statistics*, PMLR, 299–306.

LU, Z. AND T. K. LEEN (2004): "Semi-Supervised Learning with Penalized Probabilistic Clustering," in *NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems*, 849–856.

——— (2007): "Penalized Probabilistic Clustering," *Neural Computation*, 19, 1528–1567.

LUMSDAINE, R. L., R. OKUI, AND W. WANG (2022): "Estimation of Panel Group Structure Models with Structural Breaks in Group Memberships and Coefficients," *Journal of Econometrics*, Forthcoming.

MACQUEEN, J. ET AL. (1967): "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA, vol. 1, 281–297.

MATHESON, J. E. AND R. L. WINKLER (1976): "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22, 1087–1096.

MEILĂ, M. (2007): "Comparing Clusterings—An Information Based Distance," *Journal of Multivariate Analysis*, 98, 873–895.

MILLER, J. W. (2019): "An Elementary Derivation of the Chinese Restaurant Process from Sethuraman's Stick-Breaking Process," *Statistics & Probability Letters*, 146, 112–117.

MOON, H. R., F. SCHORFHEIDE, AND B. ZHANG (2023): "Bayesian Estimation of Panel Models under Potentially Sparse Heterogeneity," *arXiv preprint arXiv:2310.13785*.

MORAL-BENITO, E. AND C. BARTOLUCCI (2012): "Income and Democracy: Revisiting the Evidence," *Economics Letters*, 117, 844–847.

MÜELLER, P., F. A. QUINTANA, AND G. PAGE (2018): "Nonparametric Bayesian Inference in Applications," *Statistical Methods & Applications*, 27, 175–206.

NEAL, R. M. (1992): "Bayesian Mixture Modeling," in *Maximum Entropy and Bayesian Methods*, Springer, 197–211.

——— (2000): "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

NELSON, B. AND I. COHEN (2007): "Revisiting Probabilistic Models for Clustering with Pairwise Constraints," in *Proceedings of the 24th International Conference on Machine Learning*, 673–680.

NEWTON, M. A. AND A. E. RAFTERY (1994): "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society: Series B (Methodological)*, 56, 3–26.

NOWICKI, K. AND T. A. B. SNIJDERS (2001): "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

OKUI, R. AND W. WANG (2021): "Heterogeneous Structural Breaks in Panel Data Models," *Journal of Econometrics*, 220, 447–473.

ORBANZ, P. AND J. M. BUHMANN (2008): "Nonparametric Bayesian Image Segmentation," *International Journal of Computer Vision*, 77, 25–45.

PAGANIN, S., A. H. HERRING, A. F. OLSHAN, AND D. B. DUNSON (2021): "Centered Partition Processes: Informative Priors for Clustering," *Bayesian Analysis*, 16, 301–370.

PATI, D., D. B. DUNSON, AND S. T. TOKDAR (2013): "Posterior Consistency in Conditional Distribution Estimation," *Journal of multivariate analysis*, 116, 456–472.

PELLEG, D. AND D. BARAS (2007): "K-Means with Large and Noisy Constraint Sets," in *European Conference on Machine Learning*, Springer, 674–682.

PESARAN, M. H., A. PICK, AND A. TIMMERMANN (2022): "Forecasting with Panel Data: Estimation Uncertainty versus Parameter Heterogeneity," *CEPR Discussion Paper No. DP17123*.

PITMAN, J. (1995): "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*, 102, 145–158.

——— (1996): "Some Developments of the Blackwell-MacQueen Urn Scheme," *Lecture Notes-Monograph Series*, 245–267.

PITMAN, J. AND M. YOR (1997): "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator," *The Annals of Probability*, 25, 855–900.

PRIMICERI, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852.

PRZEWORSKI, A., F. LIMONGI, AND S. GINER (1995): "Political Regimes and Economic Growth," in *Democracy and Development*, Springer, 3–27.

QUINTANA, F. A., P. MÜLLER, A. JARA, AND S. N. MACEACHERN (2022): "The Dependent Dirichlet Process and Related Models," *Statistical Science*, 37, 24–41.

RASMUSSEN, C. (1999): "The Infinite Gaussian Mixture Model," in *Advances in Neural Information Processing Systems*, ed. by S. Solla, T. Leen, and K. Müller, MIT Press, vol. 12, 554–560.

REN, L., D. B. DUNSON, AND L. CARIN (2008): "The Dynamic Hierarchical Dirichlet Process," in *Proceedings of the 25th International Conference on Machine Learning*, 824–831.

ROCKOFF, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94, 247–252.

RODRIGUEZ, A. AND D. B. DUNSON (2011): "Nonparametric Bayesian Models through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6.

ROSS, J. AND J. DY (2013): "Nonparametric Mixture of Gaussian Processes with Constraints," in *International Conference on Machine Learning*, PMLR, 1346–1354.

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

SARAFIDIS, V. AND N. WEBER (2015): "A Partially Heterogeneous Framework for Analyzing Panel Data," *Oxford Bulletin of Economics and Statistics*, 77, 274–296.

SETHURAMAN, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

SHENTAL, N., A. BAR-HILLEL, T. HERTZ, AND D. WEINSHALL (2003): "Computing Gaussian Mixture Models with EM using Equivalence Constraints," *Advances in Neural Information Processing Systems*, 16, 465–472.

SHIRAITO, Y. (2016): "Uncovering Heterogeneous Treatment Effects," *Visited on. https://shiraito. github. io/research/files/jmp. pdf*.

SMITH, A. F. (1973): "A General Bayesian Linear Model," *Journal of the Royal Statistical Society: Series B (Methodological)*, 35, 67–75.

STOCK, J. H. AND M. W. WATSON (2007): "Why has US Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking*, 39, 3–33.

SU, L. AND G. JU (2018): "Identifying Latent Grouped Patterns in Panel Data Models with Interactive Fixed Effects," *Journal of Econometrics*, 206, 554–573.

SU, L., Z. SHI, AND P. C. PHILLIPS (2016): "Identifying Latent Structures in Panel Data," *Econometrica*, 84, 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, 334–349.

SUN, Y. (2005): "Estimation and Inference in Panel Structure Models," *Available at SSRN 794884*.

TEH, Y. W., M. I. JORDAN, M. J. BEAL, AND D. M. BLEI (2006): "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

VLACHOS, A., Z. GHAHRAMANI, AND T. BRISCOE (2010): "Active Learning for Constrained Dirichlet Process Mixture Models," in *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, 57–61.

VLACHOS, A., Z. GHAHRAMANI, AND A. KORHONEN (2008): "Dirichlet Process Mixture Models for Verb Clustering," in *Proceedings of the ICML Workshop on Prior Knowledge for Text and Language*.

VLACHOS, A., A. KORHONEN, AND Z. GHAHRAMANI (2009): "Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering," in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 74–82.

WADE, S. AND Z. GHAHRAMANI (2018): "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)," *Bayesian Analysis*, 13, 559–626.

WAGER, S. AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

WAGSTAFF, K. AND C. CARDIE (2000): "Clustering with Instance-Level Constraints," *AAAI/IAAI Proceedings*, 1097, 577–584.

WAGSTAFF, K., C. CARDIE, S. ROGERS, S. SCHROEDL, ET AL. (2001): "Constrained K-Means Clustering with Background Knowledge," in *Proceedings of the 18th International Conference on Machine Learning*, vol. 1, 577–584.

WALKER, S. G. (2007): "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics—Simulation and Computation®*, 36, 45–54.

WANG, L. AND D. B. DUNSON (2011): "Fast Bayesian Inference in Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 20, 196–216.

WANG, W., P. C. PHILLIPS, AND L. SU (2018): "Homogeneity Pursuit in Panel Data Models: Theory and Application," *Journal of Applied Econometrics*, 33, 797–815.

WANG, W. AND L. SU (2021): "Identifying Latent Group Structures in Nonlinear Panels," *Journal of Econometrics*, 220, 272–295.

YODER, J. AND C. E. PRIEBE (2017): "Semi-Supervised K-Means++," *Journal of Statistical Computation and Simulation*, 87, 2597–2608.

ZHANG, B. (2020): "Forecasting with Bayesian Grouped Random Effects in Panel Data," *arXiv preprint arXiv:2007.02435*.

Supplemental Appendix to
"Unobserved Grouped Patterns in Panel Data and Prior Wisdom"

Boyuan Zhang

A   Definitions and Terminology

A.1   Dirichlet Process and Related Stochastic Processes

All unknown quantities in a model must be assigned prior distributions in Bayesian inference. A nonparametric prior can be used to reflect uncertainty about the parametric form of the prior distribution. Because of its richness, computational ease, and interpretability, the Dirichlet process (DP) is one of the most often used random probability measures. It can be used to model the uncertainty about the functional form of the prior distribution for parameters in a model.

The DP, which was first established using Kolmogorov consistency conditions (Ferguson, 1973), can be defined from a number of views. Ferguson (1973) shows that the DP can be obtained by normalizing a gamma process. By using exchangeability, the Pólya urn method leads to the GP (Blackwell and MacQueen, 1973). The Chinese restaurant process (CRP) (Aldous, 1985; Pitman, 1996), a distribution over partitions, is a similarly related sequential process that produces the DP when each partition is assigned an independent parameter with a common distribution. Sethuraman (1994) provided a constructive definition of the DP, which led to the characterization as a stick-breaking prior (Ishwaran and James, 2001).

Construction of the DP using a stick-breaking process or a gamma process represents the DP as a countably infinite sum of atomic measures. These approaches suggest that a DPM model can be seen as a mixture model with infinitely many components. The distribution of parameters imposed by a DP can also be obtained as a limiting case of a parametric mixture model (Neal, 1992; Rasmussen, 1999; Neal, 2000). This approach shows that a DPM can easily be defined as an extension of a parametric mixture model without the need to do model selection for determining the number of components to be used.

A.1.1   Dirichlet Process

Ferguson (1973) defines a DP with two parameters, a positive scalar a and a probability measure B_0, referred to as the concentration parameter and the base measure, respectively. The base distribution B_0 is the parameter on which the nonparametric distribution is centered, which can be thought of as the prior guess (Antoniak, 1974). The concentration parameter a expresses the strength of belief in B_0. For small values of a, samples from a DP is likely to be

composed of a small number of atomic measures with large weights. For large values, most samples are likely to be distinct, thus concentrated on B_0.

Technically, a nonparametric prior is a probability distribution on $\mathcal{P}$, the space of all probability measures (say on the real line). Measurable sets in $\mathcal{P}$ are of the form $\{A : P(A) < 1\}$. We could specify a prior distribution over $(P(A_1), P(A_2), \ldots, P(A_K))$ where A_1, A_2,..., A_K are measurable finite partition of the measurable set A. Denote

$$P \sim DP(a, B_0)$$

for all partition $(A_1, \cdots, A_K)$, then,

$$(P(A_1), \cdots, P(A_K)) \sim \text{Dir}(aB_0(A_1), \cdots, aB_0(A_K))$$

Dir $(\cdot)$ stands for the Dirichlet distribution with probability distribution function being

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

where $x_i \in (0, 1)$ and $\sum_{i=1}^{K} x_i = 1$. This is a multivariate generalization of the Beta distribution and the infinite-dimensional generalization of the Dirichlet distribution is the Dirichlet process.

We next define a typical DP prior for (α, σ^2):

$$(\alpha_i, \sigma_i^2) \sim B,$$
$$B \sim DP(a, B_0), \qquad (A.1)$$

where B is a random distribution. There are two parameters. The base distribution B_0 is a distribution over the same space as B. For example, if B_0 is a distribution on reals then B must be a distribution on reals too. The concentration parameter a is a positive scalar. One property of the DP is that random distributions B are discrete, and each places its mass on a countably infinite collection of atoms drawn from B_0.

We adopt an Independent Normal Inverse-Gamma (INIG) distribution for the base distribution B_0:

$$B_0(\phi) := INIG\left(\mu_\alpha, \Sigma_\alpha, \frac{\nu_\sigma}{2}, \frac{\delta_\sigma}{2}\right), \qquad (A.2)$$

with a set of hyperparameters $\phi = \left(\mu_\alpha, \Sigma_\alpha, \frac{\nu_\sigma}{2}, \frac{\delta_\sigma}{2}\right)$.

The form of the base distribution and the value of the concentration parameter are critical aspects of model selection that influence modeling performance. Given a murky prior distribution, the concentration parameter's value can be derived from the data. It is more

difficult to choose the base distribution because the model's performance is largely dependent on its parametric form, even if it is constructed hierarchically for robustness. The choice of the base distribution is determined largely by mathematical and practical convenience. For computational ease, conjugate distributions are recommended.

A draw from DP is, by definition, a discrete distribution. In this sense, given the baseline model, imposing a DP prior on the distribution of α_{g_i} entails limiting the intercepts to some discrete values and assuming that intercepts within a group are identical, which may not be appealing for some empirical applications. A natural extension is to suppose that α_i has a continuous parametric distribution $f(\alpha_i; \theta)$, with θ as parameters, and to use a DP prior for the distribution of θ. The parameters θ are then discrete and has group structure, whereas group heterogeneity α_{g_i} has a continuous distribution, i.e., α_i within a group can be different, but they are all derived from the same distribution. This additional layer of mixing is the general idea of the Dirichlet process mixture (DPM) model.

A.1.2 Stick Breaking Process

A nonparametric prior can also be defined as the distribution of a random variable P taking values in $\mathcal{P}$. A construction of DP follows the stick-breaking process (Sethuraman, 1994),

$$P(A) = \sum_{k=1}^{\infty} \pi_k \mathbf{1}_{\alpha_k}(A),$$

$$\alpha_k \sim B_0, \quad k = 1, 2, \cdots,$$

$$\pi_k = \begin{cases} \zeta_1, & k = 1 \\ \prod_{j<k} (1 - \zeta_j) \zeta_k, & k = 2, 3, \cdots \end{cases}$$

$$\text{where } \zeta_k \sim \text{Beta}(1, a), \quad k = 1, 2, \cdots$$

The stick breaking process distinguishes the roles of B_0 and a in that the former governs component value α_k while the latter guides the choice of component probability π_k. Roughly speaking, the DP concentration parameter a is linked to the number of unique components in the mixture density and thus determines and reflects the flexibility of the mixture density.

A.1.3 Chinese Restaurant / Pólya Urn Process

Another widely used representation of the DP prior is the Chinese restaurant process (CRP). To set the stage, imagine that we have a Chinese restaurant that has infinitely many tables that can each seat infinitely many customers. When a new customer, say the n-th, enters the restaurant, the probability of them sitting at the table k with n_k other customers proportional

to n_k, and the probability of this customer sitting alone at a new table is related to a (the concentration parameter in DP),

$$p\left((\alpha_i, \sigma_i^2) = (\alpha_k, \sigma_k^2)|\alpha_{1:i-1}, \sigma_{1:i-1}^2, a, B_0\right) \propto \begin{cases} n_k & \text{if } k \text{ is an existing table} \\ a & \text{if } k \text{ is a new table.} \end{cases}$$

Upon marginalizing out the random mixing measure B, one obtains the conditional distribution of (α_i, σ_i^2) given $(\alpha_{1:i-1}, \sigma_{1:i-1}^2)$, which follows a Polya urn distribution,

$$\alpha_i, \sigma_i^2|\alpha_{1:i-1}, \sigma_{1:i-1}^2, a, B_0 \sim \frac{1}{a+i-1}\sum_{j=1}^{i-1}\delta_{(\alpha_j, \sigma_j^2)} + \frac{a}{a+i-1}B_0(\cdot). \tag{A.3}$$

Equation (A.3) reveals the *clustering property* of the DP prior: there is a positive probability that each (α_i, σ_i^2) will take on the value of another (α_j, σ_j^2), leading some of the variables to share values. The probability of sharing values is proportional to n_{g_j}. This self-reinforcing property is sometimes expressed as *the rich get richer*. This equation also reveals the roles of scaling parameter a and base distribution B_0. The unique values contained in $(\alpha_{1:N}, \sigma_{1:N}^2)$ are drawn independently from B_0, and the parameter a determines how likely α_i, σ_i^2 is to be a newly drawn value from B_0 rather than take on one of the values from $\alpha_{1:i-1}, \sigma_{1:i-1}^2$.

Chinese restaurant process shares the same characteristics as the Pólya urn process which can be extend to the two-parameter Pitman–Yor process (Pitman and Yor, 1997). Here is the basic idea of Pólya urn process. Imagine that we have an urn with possibly infinitely many colors. Let a (again, the concentration parameter in DP) be the initial number of balls with each color. The urn evolves in discrete time steps - at each step, one ball is sampled uniformly at random and put it back to the urn; The color of the withdrawn ball is observed, and one additional ball with the same color is returned to the urn. This process is then repeated.

Equation (A.3) is also called the Blackwell-MacQueen prediction rule - the conditional distribution of θ_n given previous sampled $\theta_{1:n-1}$ from the Dirichlet process prior. It characterizes the Chinese restaurant process/Pólya urn process and serves as the key component in the Pólya urn Gibbs sampler (Ishwaran and James, 2001).

Prior literature shows the equivalence between Chinese restaurant process/Pólya urn process and aforementioned processes. Blackwell and MacQueen (1973) present the equivalence between Pólya urn process and Dirichlet process. Miller (2019) formally prove that the Chinese restaurant process is equivalent to the stick breaking process.

A.2 Exchangeable Partition Probability Function

A predominant feature of the the exchangeable partition probability function (EPPF) in (2.14) is that it defines a prior distribution over G. To obtain the prior from EPPF, we must first identify all possible group partitions of N units. This problem can be recast as a prototypical "balls and urns" problem: what are the ways of putting N distinguishable balls into N indistinguishable urns if empty urns are allowed?

Example A.1. Consider a simple case in which $N = 3$ and $a = 1$. It is easy to show that there are five ways to group three units. Then the prior distribution over G under Dirichlet process is given by,

$$\Pr(g_1 = g_2 = g_3 = 1) = \frac{\Gamma(1)}{\Gamma(4)}\Gamma(3) = \frac{1}{3},$$

$$\Pr(g_1 = g_2 = 1, g_3 = 2) = \frac{\Gamma(1)}{\Gamma(4)}\Gamma(2)\Gamma(1) = \frac{1}{6},$$

$$\Pr(g_1 = g_3 = 1, g_2 = 2) = \frac{\Gamma(1)}{\Gamma(4)}\Gamma(2)\Gamma(1) = \frac{1}{6},$$

$$\Pr(g_2 = g_3 = 1, g_1 = 2) = \frac{\Gamma(1)}{\Gamma(4)}\Gamma(2)\Gamma(1) = \frac{1}{6},$$

$$\Pr(g_1 = 1, g_2 = 2, g_3 = 3) = \frac{\Gamma(1)}{\Gamma(4)}\Gamma(1)\Gamma(1)\Gamma(1) = \frac{1}{6}.$$

Finding all solutions to the "balls and urns" problem with big N is computationally impossible in general. For a certain number of groups K, the number of ways to assign N unit to K groups is described by the *Stirling number of the second kind*,

$$\mathcal{S}_{N,K} = \frac{1}{K!} \sum_{j=0}^{K} (-1)^j C_K^j (K-j)^N. \tag{A.4}$$

The sum of $\mathcal{S}_{N,K}$ over all possible K, also known as the N-th Bell number, $\mathcal{B}_N = \sum_{K=1}^{N} \mathcal{S}_{N,K}$ describes the number of all possible partitions of N balls. Owing to the rapid growth of the space, listing all feasible partitions becomes computationally impossible. For example from a moderate $N = 12$ to 15, the number of partitions increases from $\mathcal{B}_{12} = 4,213,597$ to $\mathcal{B}_{15} = 1,382,958,545$. Sethuraman (1994) and Pitman (1996) constructively show that group indices/partitions can be drawn from the EPPF for DP using the stick-breaking process defined in (2.8). As a result, the EPPF does not explicitly appear in the posterior analysis in the current setting so long as the priors for the stick lengths are included.

In the example below, we demonstrate how pairwise constraints affect the prior density

of the group partition using Equation (2.4):

$$p(G|\psi, T) \propto p(G) \prod_{i,j} \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right)^{cT_{ij}\delta_{ij}(G)}.$$

(A.5)

Example A.2. Consider again the three-unit case in Example A.1 with $a = 1$. Assume there is a positive-link constraint between units 1 and 2 and that is the only constraint in this example. Then the prior probabilities of the five partitions are adjusted to account for the effect of W_{12}:

$$\Pr(g_1 = g_2 = g_3 = 1) = \frac{1}{3} \cdot \frac{2\exp(4cW_{12})}{\exp(4cW_{12}) + 1} > \frac{1}{3},$$

$$\Pr(g_1 = g_2 = 1, g_3 = 2) = \frac{1}{3} \cdot \frac{\exp(4cW_{12})}{\exp(4cW_{12}) + 1} > \frac{1}{6},$$

$$\Pr(g_1 = g_3 = 1, g_2 = 2) = \frac{1}{3} \cdot \frac{1}{\exp(4cW_{12}) + 1} < \frac{1}{6},$$

$$\Pr(g_2 = g_3 = 1, g_1 = 2) = \frac{1}{3} \cdot \frac{1}{\exp(4cW_{12}) + 1} < \frac{1}{6},$$

$$\Pr(g_1 = 1, g_2 = 2, g_3 = 3) = \frac{1}{3} \cdot \frac{1}{\exp(4cW_{12}) + 1} < \frac{1}{6}.$$

Note that $c > 0$ and $W_{12} > 0$. Comparing to the results in Example A.1, the probabilities of the first two partitions become higher since they all meet the PL constraint between units 1 and 2, while the rest of the partitions violate the constraint and hence the probabilities drop.

A.3 Hierarchical Dirichlet Process

The hierarchical Dirichlet process (HDP) was developed by Teh et al. (2006). The HDP is a nonparametric Bayesian approach to clustering grouped data, with the known group membership. It equips a Dirichlet process for each group of data, with the Dirichlet processes for all groups sharing a base distribution which is itself drawn from a Dirichlet process. This method allows groups to share statistical strength via sharing of clusters across groups. The base distribution being drawn from a Dirichlet process is important, because draws from a Dirichlet process are atomic probability measures, and the atoms will appear in all group-level Dirichlet processes. Since each atom corresponds to a cluster, clusters are shared across all groups.

The HDP is parameterized by a base distribution H that governs the prior distribution over data items, and a number of concentration parameters that govern the prior number of clusters and amount of sharing across groups. Assume that we have J groups of data, each consisting of N_j data points, $y_{j1}, \ldots y_{jN_j}$. The process defines a set of random probability

measures $(B_j)_{j=1}^J$, one for each group. The random probability measure B_j for the j-th group is distributed as a Dirichlet process:

$$B_j | B_0 \sim \text{DP}(\gamma, B_0), \tag{A.6}$$

where γ is the concentration parameter and B_0 is the base distribution shared across all groups. The distribution of the global random probability measure B_0 is given by,

$$B_0 \sim \text{DP}(\alpha_0, H), \tag{A.7}$$

with concentration parameter α_0 and base distribution H.

A hierarchical Dirichlet process can be used as the prior distribution over the parameters for grouped data. For each j, let $(\phi_{ji})_{i=1}^{n_j}$ be i.i.d. random variables distributed as B_j. Each ϕ_{ji} is a parameter corresponding to a single observation y_{ji}. The likelihood is given by,

$$\begin{aligned}
\phi_{ji} | B_j &\sim B_j, \\
y_{ji} | \phi_{ji} &\sim p(\phi_{ji}).
\end{aligned} \tag{A.8}$$

The resulting model above is called a HDP mixture model, with the HDP referring to the hierarchically linked set of Dirichlet processes, and the mixture model referring to the way the Dirichlet processes are related to the data items.

To understand how the HDP implements a clustering model, and how clusters become shared across groups, recall that draws from a Dirichlet process are atomic probability measures with probability one. The base distribution B_0 can be expressed using a stick-breaking representation,

$$B_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}, \tag{A.9}$$

where there are an infinite number of atoms, $\theta_k \sim H$, $k = 1, 2, \ldots$. Each atom is associated with a mass β_k and $\beta = (\beta_i)_{i=1}^{\infty} \sim \text{Stick}(\gamma)$ are mutually independent. Since B_0 is the base distribution for the group specific Dirichlet processes, each B_j has the same atoms as B_0 and can be written in the form,

$$B_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}. \tag{A.10}$$

Let $\pi_j = (\pi_{jk})_{k=1}^{\infty}$. Note that the weights π_j are independent given β (since the B_j are independent given B_0). It can be shown that the connection between the weights π_j and the

global weights β is

$$\pi_j \mid \alpha_0, \beta \sim \mathrm{DP}\left(\alpha_0, \beta\right). \tag{A.11}$$

Thus the set of atoms is shared across all groups, with each group having its own group-specific atom masses. Relating this representation back to the observed data, we see that each data item is described by a mixture model,

$$y_{ji}|B_j \sim \sum_{k=1}^{\infty} \pi_{jk} f\left(\theta_k\right), \tag{A.12}$$

where the atoms θ_k play the role of the mixture component parameters, while the masses π_{jk} play the role of the mixing proportions. As a result, each group of data is modeled using a mixture model, with mixture components shared across all groups and group-specific mixing weights.

Chinese Restaurant Franchise

Teh et al. (2006) have also described the marginal probabilities obtained from integrating over the random measures B_0 and $(B_j)_{j=1}^{J}$. They show that these marginals can be described in terms of a Chinese restaurant franchise (CRF) that is an analog of the Chinese restaurant process.

Recall that ϕ_{ji} are random variables with distribution B_j. In the following discussion, we will let $\theta_1, \ldots, \theta_K$ denote K i.i.d. random variables distributed according to H, and, for each j, we let $\psi_{j1}, \ldots, \psi_{jT_j}$ denote T_j i.i.d. variables distributed according to B_0.

Each ϕ_{ji} is associated with one ψ_{jt}, while each ψ_{jt} (table id) is associated with one θ_k. Let t_{ji} be the index of the ψ_{jt} associated with ϕ_{ji}, and let k_{jt} (dish id) be the index of θ_k associated with ψ_{jt}. Let n_{jt} be the number of ϕ_{ji}'s associated with ψ_{jt}, while m_{jk} is the number of ψ_{jt}'s associated with θ_k. Define $m_k = \sum_j m_{jk}$ as the number of ψ_{jt}'s associated with θ_k over all j. Notice that while the values taken on by the ψ_{jt}'s need not be distinct, they are distributed according to a discrete random probability measure $B_0 \sim \mathrm{DP}(\gamma, H)$, we are denoting them as distinct random variables.

First consider the conditional distribution for ϕ_{ji} given $\phi_{j1}, \ldots, \phi_{ji-1}$ and B_0, where B_j is integrated out, we have,

$$\phi_{ji}|\phi_{j1}, \ldots, \phi_{ji-1}, \alpha_0, G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0, \tag{A.13}$$

This is a mixture, and a draw from this mixture can be obtained by drawing from the terms on the right-hand side with probabilities given by the corresponding mixing proportions. If a term in the first summation is chosen, then we set $\phi_{ji} = \psi_{jt}$ and let $t_{ji} = t$ for the chosen t.

If the second term is chosen, then we increment T_j by one, draw $\psi_{jT_j} \sim B_0$ and set $\phi_{ji} = \psi_{jT_j}$ and $t_{ji} = T_j$.

Now we proceed to integrate out B_0. Notice that B_0 appears only in its role as the distribution of the variables ψ_{jt}. Since B_0 is distributed according to a Dirichlet process, we can integrate it out and writing the conditional distribution of ψ_{jt} directly:

$$\psi_{jt}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k} + \frac{\gamma}{\sum_k m_k + \gamma} H. \tag{A.14}$$

If we draw ψ_{jt} via choosing a term in the summation on the right-hand side of this equation, we set $\psi_{jt} = \theta_k$ and let $k_{jt} = k$ for the chosen k. If the second term is chosen, we increment K by one, draw $\theta_K \sim H$ and set $\psi_{jt} = \theta_K, k_{jt} = K$.

In short, the CRF is comprised of J restaurants with a shared menu across the restaurants. Each restaurant corresponds to an HDP group, and an infinite buffet line of dishes common to all restaurants. The process of seating customers at tables, however, is restaurant specific. Each customer is preassigned to a given restaurant determined by that customer's group j. Upon entering the j th restaurant in the CRF, customer y_{ji} sits at currently occupied tables t_{ji} with probability proportional to the number of currently seated customers, or starts a new table $T_j + 1$ with probability proportional to α. The first customer to sit at a table goes to the buffet line and picks a dish k_{jt} for their table, choosing the dish with probability proportional to the number of times that dish has been picked previously, or ordering a new dish θ_{K+1} with probability proportional to γ. The intuition behind this predictive distribution is that integrating over the global dish probabilities β results in customers making decisions based on the observed popularity of the dishes throughout the entire franchise.

A.4 Random Effects vs. Fixed-Effects

Regarding the connection between Bayesian and frequentists' panel data model, according to Koop (2003), if we impose a

(i) non-hierarchical prior (such as Normal prior without hyperpriors) on the intercept α_{i0}, the resulting panel data model is equivalent to the frequentist fixed-effects model. This is basically a Bayesian linear regression with standard priors on parameters.

(ii) hierarchical prior on the intercept α_{i0}, the resulting panel data model is equivalent to the frequentists' random effects model. A convenient hierarchical prior assumes that, for $i = 1, \ldots, N$,

$$\alpha_{i0} \sim N\left(\mu_\alpha, \sigma_\alpha^2\right).$$

The hierarchical structure of the prior arises if we treat μ_α and V_α as unknown parameters which require their own prior. We assume μ_α and V_α to be independent of one another with

$$\mu_\alpha \sim N\left(\underline{\mu}_\alpha, \underline{\sigma}^2_\alpha\right),$$

and

$$\sigma^2_\alpha \sim IG\left(\underline{\tau}_\alpha, \underline{\nu}_\alpha\right).$$

This is analogous to the random effects model as α_i are essentially assumed to draw from the underlying distribution, and data are used to update our prior on the hyperparameters of the underlying distribution.

The discussion in Koop (2003) is in line with the hierarchical models discussed in Smith (1973). The panel data model equipped with a non-hierarchical prior is a two-stage hierarchical model which results in a fixed effects model, while incorporating a hierarchical prior forms a three-stage hierarchical model that corresponds to a random effects model.

Back to our settings, if the baseline prior for α_{0i} is a DP (DPM) prior, then our proposed nonparametric Bayesian prior is a type of non-hierarchical (hierarchical) prior with latent group structure in intercepts and hence we call our proposed estimator as the constrained grouped fixed (random) effects estimator.

B Priors

Prior on Group-Specific Parameters

$$\left(\alpha_i, \sigma^2_i\right) \sim \sum_{k=1}^{\infty} \pi_k \delta_{\left(\alpha_k, \sigma^2_k\right)} \text{ with } \left(\alpha_k, \sigma^2_k\right) \sim B_0(\phi). \tag{B.1}$$

B_0 is an Independent Normal Inverse-Gamma (INIG) distribution:

$$B_0 := INIG\left(\mu_\alpha, \Sigma_\alpha, \frac{\nu_\sigma}{2}, \frac{\delta_\sigma}{2}\right), \tag{B.2}$$

with a set of hyperparameters $\phi = \left(\mu_\alpha, \Sigma_\alpha, \frac{\nu_\sigma}{2}, \frac{\delta_\sigma}{2}\right) = (0, 1, 6, 5)$.

Prior on Stick Lengths

$$\tilde{\varsigma}_k \sim Beta\left(1, a\right), \tag{B.3}$$

where a is the concentration parameter.

Hyper-prior on Concentration Parameter

$$a \sim Gamma\,(m, n)\,, \tag{B.4}$$

with $(m, n) = (0.4, 10)$.

Prior on Common and Individual Slope Coefficients (if any) Finally, the prior distribution for the common parameter ρ is chosen to be a normal distribution to stay close to the linear regression framework,

$$\rho \sim N(0, \sigma_\rho^2) \text{ with } \sigma_\rho^2 = 1. \tag{B.5}$$

The prior of heterogeneous parameter β_i follows,

$$\beta_i \sim N(0, \Sigma_\beta) \text{ with } \Sigma_\beta = 1 \times \mathbf{I}_p. \tag{B.6}$$

C   Posterior Distributions and Algorithms

C.1   Blocked Gibbs Sampler and Algorithm

Initialization:

(i) Preset the initial number of active groups as $K_0^a = N$.

(ii) Set concentration parameter a to its prior mean.

(iii) In ignorance of group heterogeneity ($K = 1$) and heteroskedasticity, use Anderson and Hsiao (1982) IV approach to get $\hat{\alpha}_{IV}$ and $\hat{\Sigma}_{\alpha, IV}$. These IV estimators serve as the mean and covariance matrix in the related priors.

(iv) Generate K_0^a random sample from the distribution $N(\hat{\alpha}_{IV}, \hat{\Sigma}_{\alpha, IV})$.

(v) Initialize group membership G by using assuming no group structure: $G^{(0)} = [1, 2, ..., N]$.

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 (C.14):

$$\xi_k | a^{(s-1)}, G^{(s-1)} \sim Beta\left(|B_k| + 1, a + \sum_{j=1}^{N} \mathbf{1}(g_j > k) \right),$$

and calculate group probability in accordance to (C.15).

(iii) Group heterogeneity: for $k = 1, 2, ..., K^a$, draw $\alpha_k^{(s)}$ from a normal distribution in (C.12):

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

(iv) Group heteroscedasticity: for $k = 1, 2, ..., K^a$ and $t = 1, 2, ..., T$, draw $\sigma_k^{2(s)}$ from an inverse Gamma distribution in (C.13):

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

(v) Auxiliary variables: for $i = 1, 2, ..., N$, draw u_i from an uniform distribution in (C.18):

$$u_i | \Xi^{(s)}, G^{(s-1)} \sim Unif(0, p_{g_i}^{(s)}).$$

Then calculate u^* according to (C.9).

(vi) DP concentration parameter:

(a) Draw latent variable η from a Beta distribution in (C.16):

$$\eta \sim Beta(a + 1, N)$$

(b) Draw concentration parameter a from a mixture of Gamma distribution in (C.17):

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

and π_a is defined as

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

(vii) 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 $\pi_{\tilde{K}} = \xi_{\tilde{K}} \prod_{j < \tilde{K}} \left(1 - \xi_j\right)$ 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.

(xi) Group membership: for $j = 1, 2, ..., J$ and $k = 1, 2, ..., K^a$, draw g_j from a multinomial distribution in (C.19).

C.2 Random Coefficients Model with Soft Constraints

We present the conditional posterior distributions of parameters in the time-invariant random effects model with heteroscedasticity, positive-link constraints and negative-link constraints, which is the most complicated scenarios. For other models, such as its homoscedastic counterparts, adjustment can be easily made by assuming common error variances.

C.2.1 Derivation

Model:

$$y_{it} = \alpha'_{g_i} x_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N\left(0, \sigma^2_{g_i}\right). \tag{C.1}$$

To facilitate derivation, we stack observations and parameters,

$$\text{Dependent variable: } Y = [y_1, y_2, \ldots, y_N], y_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]',$$
$$\text{Covariates: } X = [x_1, x_2, \ldots, x_N], x_i = [x_{i1}, x_{i2}, \ldots, x_{iT}]',$$
$$\text{Grouped-specific parameters: } \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N],$$
$$\text{Error variance: } \Sigma = \left[\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2\right],$$
$$\text{Stick length: } \Xi = [\xi_1, \xi_2, \ldots],$$
$$\text{Group indices: } G = [g_1, \ldots, g_N],$$
$$\text{Auxiliary variable: } u = [u_1, u_2, ..., u_N],$$
$$\text{Hyper parameters: } \phi = [\mu_\alpha, \Sigma_\alpha, \nu_\sigma, \delta_\sigma].$$

In order to write down the posterior of unknown parameters given a set of pairwise constraints, a probabilistic model of how weights of constraints are obtained must be specified. Inspired by Shental et al. (2003), we have the following assumptions:

Assumption 3. *(Data)*

(i) Data points are first sampled i.i.d from the full probability distribution conditional on G.

(ii) From this sample, pairs of points are randomly chosen according to a uniform distribution. In case both points in a pair belong to the same source a positive-link constraint is formed and a negative-link if formed when they belong to different sources.

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

$$
\begin{aligned}
&p(\alpha, \sigma^2, \Xi, a, G | Y, X, W, \phi) \\
&\propto p(Y|X, \alpha, \sigma^2, G) p(\alpha, \sigma^2 | \phi) p(G|\Xi, W) p(\Xi|a) p(a).
\end{aligned}
\tag{C.2}
$$

All priors have been well-defined except for $p(G|\Xi, W)$ - the prior for group indices G conditional on stick lengths Ξ and the weights of constraints W.

Using the Bayes rule, the modified prior for the group indices is

$$
p(G|\Xi, W) = \frac{p(W|G) p(G|\Xi)}{\sum_{G'} p(W|G') p(G'|\Xi)} \propto p(W|G) p(G|\Xi),
\tag{C.3}
$$

where the sum in the denominator is taken over all possible group partitioning, $p(W|G)$ is the weighting function of the form:

$$
p(W|G) = \prod_{i,j} \exp\left(c W_{ij} \delta_{ij}\right),
$$

and $p(G|\Xi)$ is the density of a categorical distribution with probabilities generated by the stick-breaking process.

From (C.3), the prior of g_i conditional on the group indices of the other $G^{(-i)}$ is

$$
p(g_i|\Xi, W_i, G^{(-i)}) \propto p\left(W_i|G\right) p(g_i|\Xi),
\tag{C.4}
$$

where $W_i = \{W_{ij} | j = 1, ..., N\}$ and

$$
p(W_i|G) = \prod_{j=1}^{N} \exp\left(2c W_{ij} \delta_{ij}\right).
\tag{C.5}
$$

Given the expression of $p(g_i|\Xi, W_i, G^{(-i)})$ and the DP prior specified in Appendix B, the

posterior of unknown objects in the random coefficients model can be written as,

$$p(\alpha, \sigma^2, \Xi, a, G | Y, X, W, \phi)$$

$$\propto p(Y | X, \alpha, \sigma^2, G) p(\alpha, \sigma^2 | \phi) p(G | \Xi, W) p(\Xi | a) p(a)$$

$$\propto \prod_{i=1}^{N} p\left(y_i | x_i, \alpha_{g_i}, \sigma^2_{g_i}\right) \prod_{j=1}^{\infty} p(\alpha_j, \sigma^2_j | \phi) \prod_{j=1}^{\infty} p(\xi_j | a) \prod_{i=1}^{N} p\left(g_i | \Xi, W_i, G^{(-i)}\right) p(a)$$

$$= \left[\prod_{i=1}^{N} p\left(y_i | x_i, \alpha_{g_i}, \sigma^2_{g_i}\right) p\left(g_i | \Xi, W_i, G^{(-i)}\right)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma^2_j | \phi) p(\xi_j | a)\right] p(a)$$

$$= \left[\prod_{i=1}^{N} p\left(y_i | x_i, \alpha_{g_i}, \sigma^2_{g_i}\right) p(W_i | G) p(g_i | \Xi)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma^2_j | \phi) p(\xi_j | a)\right] p(a). \qquad (C.6)$$

In the following derivation and algorithm, we adopt the slice sampler (Walker, 2007) that avoids approximation in Ishwaran and James (2001). 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, $u_i \overset{iid}{\sim} U(0, 1)$. Then the augmented posterior is written as,

$$p(\alpha, \sigma^2, \Xi, a, G, u | Y, X, W, \phi)$$

$$\propto \left[\prod_{i=1}^{N} p\left(y_i | x_i, \alpha_{g_i}, \sigma^2_{g_i}\right) \mathbf{1}(u_i < \pi_{g_i}) p(W_i | G)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma^2_j | \phi) p(\xi_j | a)\right] p(a)$$

$$= \left[\prod_{i=1}^{N} p\left(y_i | x_i, \alpha_{g_i}, \sigma^2_{g_i}\right) p(u_i | \pi_{g_i}) \pi_{g_i} p(W_i | G)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma^2_j | \phi) p(\xi_j | a)\right] p(a), \qquad (C.7)$$

where $\pi_{g_i} = p(g_i | \Xi)$, $p(u_i | \pi_{g_i})$ is a uniform distribution defined on $[0, \pi_{g_i}]$, and $\mathbf{1}(\cdot)$ is the indicator function, which is equal to zero unless the specific condition is met. 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

$$K^* = \min_{k} \left\{ u^* > 1 - \sum_{j=1}^{k} \pi_j \right\}, \qquad (C.8)$$

where

$$u^* = \min_{1 \leqslant i \leqslant N} u_i. \tag{C.9}$$

Such a 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_k, σ_k^2) to K^* as the densities of (α_k, σ_k^2) and equal 0 for $k > K^*$ due to $\mathbf{1}(u_i < \pi_k) = 0$, which will be clear in the subsequent posterior derivation. Intuitively, the latent variable u_i has an effect of "dynamically truncating" the number of groups needed to be sampled.

Next, we define the number of active groups

$$K^a = \max_{1 \leqslant i \leqslant N} g_i. \tag{C.10}$$

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

As the base distribution B_0 is the Independent-Normal-Inverse-Gamma distribution, the prior density of α_i and σ_i^2 are independent.

Conditional posterior of α (grouped coefficients).

$$p(\alpha|\sigma^2, G, Y, X, \phi) \propto \left[\prod_{i=1}^{N} p\left(y_i|x_i, \alpha_{g_i}, \sigma_{g_i}^2\right)\right] \left[\prod_{j=1}^{\infty} p(\alpha_j, \sigma_j^2|\phi)\right].$$

For $k = 1, 2, ..., K^a$, define a set of units that belong to the group k,

$$B_k = \{i|g_i = k, i \in \{1, 2, ..., N\}\}, \tag{C.11}$$

then the posterior density for α_k read as

$$p(\alpha_k|\sigma_k^2, G, Y, X, \phi)$$

$$\propto \left[\prod_{i \in B_k} p(y_i|x_i, \alpha_{g_i}, \sigma_{g_i}^2)\right] p(\alpha_k|\phi)$$

$$\propto \exp\left[-\frac{1}{2\sigma_k^2} \sum_{i \in B_k} (y_i - x_i\alpha_k)' (y_i - x_i\alpha_k)\right] \exp\left[-\frac{1}{2} (\alpha_k - \mu_\alpha)' \Sigma_\alpha^{-1} (\alpha_k - \mu_\alpha)\right]$$

$$\propto \exp\left[-\frac{1}{2} (\alpha_k - \bar{\mu}_{\alpha_k})' \bar{\Sigma}_{\alpha_k}^{-1} (\alpha_k - \bar{\mu}_{\alpha_k})\right].$$

---

[17]See proof in theorem 4.
[18]See proof in theorem 4.

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

$$\alpha_k | \sigma_k^2, G, Y, X, \phi \sim N\left(\bar{\mu}_{\alpha_k}, \bar{\Sigma}_{\alpha_k}\right). \tag{C.12}$$

where

$$\bar{\Sigma}_{\alpha_k} = \left(\Sigma_\alpha^{-1} + \sigma_k^{-2} \sum_{i \in B_k} x_i' x_i\right)^{-1},$$

$$\bar{\mu}_{\alpha_k} = \bar{\Sigma}_{\alpha_k} \left(\Sigma_\alpha^{-1} \mu_\alpha + \sigma_k^{-2} \sum_{i \in B_k} x_i' y_i^\alpha\right),$$

$$y_i^\alpha = y_i - x_i \alpha_{g_i}.$$

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

$$p(\sigma_k^2 | \alpha_k, G, Y, X, \phi) \propto \left[\prod_{i \in B_k} p\left(y_i | x_i, \alpha_{g_i}, \sigma_{g_i}^2\right)\right] p(\sigma_k^2 | \phi).$$

With 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

$$p(\sigma_k^2 | \alpha_k, G, Y, X, \phi)$$

$$\propto \prod_{i \in B_k} \left[\left(\sigma_k^2\right)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma_k^2}(y_i - x_i\alpha_k)'(y_i - x_i\alpha_k)\right)\right] \left(\frac{1}{\sigma_k^2}\right)^{\frac{v_\sigma}{2}+1} \exp\left(-\frac{\delta_\sigma}{2\sigma_k^2}\right)$$

$$= \left(\frac{1}{\sigma_k^2}\right)^{\frac{v_\sigma + T|B_k|}{2}+1} \exp\left[-\frac{\delta_\sigma + \sum_{i \in B_k}(y_i - x_i\alpha_k)'(y_i - x_i\alpha_k)}{2\sigma_k^2}\right].$$

This implies

$$\sigma_k^2 | \alpha_k, G, Y, X, \phi \sim IG\left(\frac{\bar{v}_{\sigma,k}}{2}, \frac{\bar{\delta}_{\sigma,k}}{2}\right), \tag{C.13}$$

where

$$\bar{v}_{\sigma,k} = v_\sigma + T|B_k|,$$

$$\bar{\delta}_{\sigma,kt} = \delta_\sigma + \sum_{i \in B_k}(y_i - x_i\alpha_k)'(y_i - x_i\alpha_k),$$

$$|B_k| = \text{\# of units in group } k.$$

Conditional posterior of Ξ (stick length).

$$p(\Xi|a,G)$$

$$\propto \left[\prod_{i=1}^{N} p(u_i|\pi_{g_i})\pi_{g_i}\right]\left[\prod_{j=1}^{\infty} p(\xi_j|a)\right]$$

$$\propto \left[\prod_{i=1}^{N} p(u_i|\pi_{g_i})\xi_{g_i}\prod_{l<g_i}(1-\xi_l)\right]\left[\prod_{j=1}^{\infty} p(\xi_j|a)\right].$$

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

$$p(\Xi|a,G) \propto \left(\prod_{i\in B_k}\xi_k\right)(1-\xi_k)^{\sum_{j=1}^{N}\mathbf{1}(g_j>k)}(1-\xi_k)^{a-1},$$

$$\propto \xi_k^{|B_k|}(1-\xi_k)^{a+\sum_{j=1}^{N}\mathbf{1}(g_j>k)-1}.$$

where B_K is the set of units that currently belong to group k, see equation (C.11).

Therefore, posterior distribution of ξ_k is

$$\xi_k|a,G \sim Beta\left(|B_k|+1, a+\sum_{j=1}^{N}\mathbf{1}(g_j>k)\right). \tag{C.14}$$

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

$$\pi_k|G,\Xi = \begin{cases} \xi_1, & k=1 \\ \xi_k\prod_{j<k}(1-\xi_j), & k=2,\ldots,K^a \end{cases}. \tag{C.15}$$

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,

$$\eta \sim Beta(a+1, J). \tag{C.16}$$

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

$$p(a|\eta, K^a) = \pi_a Gamma\,(m + K^a, n - \log(\eta)) + (1 - \pi_a)Gamma\,(m + K^a - 1, n - \log(\eta)),$$
(C.17)

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 indices G, it is straightforward to show that the posterior density of u_i is a uniform distribution defined on $(0, \pi_{g_i})$, that is

$$u_i|\Xi, G \sim Unif\,(0, \pi_{g_i}),$$
(C.18)

where $\pi_{g_i} = \xi_{g_i} \prod_{j < g_i}(1 - \xi_j)$. Moreover, it is worth noting that the values for K^* and u^* need to be updated according to equation (C.8) and (C.9) after this step.

Conditional posterior of G (group indices). 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., $G^{(-i)} = G \backslash g_i$. As a result, for $k = 1, 2, ..., K^*$,

$$p\left(g_i = k|y_i, x_i, \alpha_k, \sigma_k^2, G^{(-i)}, u_i\right)$$
$$\propto p\left(y_i|x_i, \alpha_k, \sigma_k^2\right) \mathbf{1}(u_i < \pi_k)p\,(W_i|G)$$
$$= p\left(y_i|x_i, \alpha_k, \sigma_k^2\right) \mathbf{1}(u_i < \pi_k) \prod_{j=1}^{N} \exp\,(2cW_{ij}\delta_{ij}).$$
(C.19)

Finally, we normalize the point mass to get a valid distribution.

D Technical Proofs

D.1 Slice Sampling

Theorem 4. *Suppose that we have a model with posterior as given in Appendix C.2. Given the definition of the number of potential component K^* in (C.8), the minimum of auxiliary variables u^* in (C.9) and the number of active group K in (C.10), we have*

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

(ii) $K < K^*$.

Proof. (i) By definition, $u^* = \min\limits_{1 \leqslant i \leqslant N} u_i$ 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.* $1 - u^* < \sum\limits_{j=1}^{K^*} \pi_j$, then $K \leqslant K^*$.

$\square$

D.2   Connection to Lu and Leen (2004) and Lu and Leen (2007)

In this section, we will first show the close connection between the modified prior in the presence of soft constraints defined in (2.4) and the framework of penalized probabilistic clustering proposed by Lu and Leen (2004) and Lu and Leen (2007). Then we will discuss the properties of the weights W_{ij}.

We start with joint prior odds in (2.4):

$$\prod_{i,j} \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right)^{cT_{ij}\delta_{ij}} = \prod_{i,j} \exp \left[c\delta_{ij} \log \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right)^{T_{ij}} \right]. \tag{D.1}$$

Define the weight as $W_{ij} = \log \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right)^{T_{ij}}$. Then when $T_{ij} = 1$, we have

$$W_{ij} = \log \left(\frac{\psi_{ij}}{1 - \psi_{ij}} \right) \Leftrightarrow \psi_{ij} = \frac{\exp(W_{ij})}{1 + \exp(W_{ij})}. \tag{D.2}$$

When $T_{ij} = -1$, we get

$$W_{ij} = \log \left(\frac{1 - \psi_{ij}}{\psi_{ij}} \right) \Leftrightarrow 1 - \psi_{ij} = \frac{\exp(W_{ij})}{1 + \exp(W_{ij})}. \tag{D.3}$$

Combining (D.2) and (D.3) yields that

$$\frac{\exp\left(W_{ij}\right)}{1+\exp\left(W_{ij}\right)} = \psi_{ij}^{\frac{1}{2}(1+T_{ij})}(1-\psi_{ij})^{\frac{1}{2}(1-T_{ij})}. \tag{D.4}$$

This is exactly the equation (7) in Lu and Leen (2007) with $\gamma_{ij} = \psi_{ij}$ and $L_{ij} = \frac{1}{2}(T_{ij}+1)$, which uniquely defines the expression for the weights associated with each pairwise constraint given γ_{ij} and L_{ij}. Since both L_{ij} and T_{ij} are indicators for the type of constraints, they don't affect the formula for W_{ij}, thus the following formula weights coincides with the one used in Lu and Leen (2007) and the both frameworks converge,

$$W_{ij} = \begin{cases} \log\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right) & \text{if } T_{ij} = 1 \\ \log\left(\frac{1-\psi_{ij}}{\psi_{ij}}\right) & \text{if } T_{ij} = -1 \\ 0 & \text{if } T_{ij} = 0. \end{cases} \tag{D.5}$$

Accordingly, the prior defined in (D.1) can be rewritten in term of W_{ij} as

$$\prod_{i,j}\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right)^{cT_{ij}\delta_{ij}} = \prod_{i,j}\exp\left(cW_{ij}\delta_{ij}\right). \tag{D.6}$$

The weight W_{ij} associated with the constraint between unit i and j as in (D.5) has the following properties:

(a) Unboundedness: $W_{ij} \in (-\infty, \infty)$;

(b) Symmetry: $W_{ij} = W_{ji}$;

(c) Sign reflects constraint's type: If $(i,j) \in \mathcal{P}$ or $L_{ij} = 1$, then $W_{ij} = \log\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right) > 0$; If $(i,j) \in \mathcal{N}$ or $L_{ij} = -1$, then $W_{ij} = \log\left(\frac{1-\psi_{ij}}{\psi_{ij}}\right) < 0$; If (i,j) doesn't involve in any constraint or $L_{ij} = 0$, then $W_{ij} = 0$.

(d) Absolute value reflects constraint's accuracy:

$$\frac{e^{|W_{ij}|}}{1+e^{|W_{ij}|}} = \psi_{ij}.$$

It is straightforward to show that $|W_{ij}|$ is monotonically increasing in ψ_{ij}.

D.3 *Prior Similarity Matrix*

Proof of Theorem 1. Given equation (2.14) and (2.6), the prior probability of unit i and j being in the same group is

$$\Pr(g_i = g_j | W)$$

$$= \sum_{G \in \mathcal{G}_{ij}} \frac{1}{M} p(G) \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right)$$

$$= \sum_{G \in \mathcal{G}_{ij}} \frac{1}{M} \frac{\Gamma(a)}{\Gamma(a+N)} \left[\prod_{k=1}^{K} a\Gamma(|B_k|) \right] \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right)$$

$$= \sum_{G \in \mathcal{G}_{ij}} A(G) \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right) \tag{D.7}$$

where $\mathcal{G}_{ij}$ is the set of all possible group indices that satisfies $g_i = g_j$ and M is the normalization constant in (2.6).

$\mathcal{G}_{ij}$ and $\mathcal{G}_{ik}$ are closed related. It is straightforward to see that the numbers of element in $\mathcal{G}_{ij}$ and $\mathcal{G}_{ik}$ are equal since they are all equal to the number of permutation of other $N - 2$ units. Moreover, as unit j and k are exchangeable, $\mathcal{G}_{ik}$ can be constructed from $\mathcal{G}_{ij}$ by swapping the group index of unit j and k.

As a result, we can find an one-on-one mapping between $\mathcal{G}_{ij}$ and $\mathcal{G}_{ik}$. That is, for any $G \in \mathcal{G}_{ij}$, if we swap the group index of unit j and k, the resulting partition $s_{jk}(G)$ belongs to $\mathcal{G}_{ik}$, and vice versa. As the constant $A(G)$ depends only on the size of partitions, we have $A(G) = A(s_{jk}(G))$.

The properties between $\mathcal{G}_{ij}$ and $\mathcal{G}_{ik}$ enable we to compare each summand in $\Pr(g_i = g_j | W)$ and $\Pr(g_i = g_k | W)$. The difference between these two probabilities is

$$\Pr(g_i = g_j | W) - \Pr(g_i = g_k | W)$$

$$= \sum_{G \in \mathcal{G}_{ij}} A(G) \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right) - \sum_{G \in \mathcal{G}_{ik}} A(G) \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right)$$

$$= \sum_{G \in \mathcal{G}_{ij}} A(G) \exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right) - A(s_{jk}(G)) \exp \left(c \sum_{m,n} W_{mn} \delta'_{mn} \right)$$

$$= \sum_{G \in \mathcal{G}_{ij}} A(G) \left[\exp \left(c \sum_{m,n} W_{mn} \delta_{mn} \right) - \exp \left(c \sum_{m,n} W_{mn} \delta'_{mn} \right) \right]. \tag{D.8}$$

where δ'_{mn} is evaluated at $s_{jk}(G)$.

We can classify a group partitioning G into two cases:

(i) $G = s_{jk}(G)$. This happens when units j and k are assigned to the same group. Swapping them doesn't affect the group partitioning, which indicates that $\sum_{m,n} W_{mn}\delta_{mn} = \sum_{m,n} W_{mn}\delta'_{mn}$.

(ii) $G \neq s_{jk}(G)$. These are the more common cases. We again compare $\sum_{m,n} W_{mn}\delta_{mn}$ with $\sum_{m,n} W_{mn}\delta'_{mn}$. $W_{mn}\delta_{mn}$ and $W_{mn}\delta'_{mn}$ are equal when $m \neq j,k$ and $n \neq j,k$ as these terms remain unchanged regardless of the group indices of units j and k. For $m = j,k$, note that $\delta_{jn} = \delta'_{kn}$ and $\delta_{kn} = \delta'_{jn}$ for all $n = 1, 2, .., N$. Therefore, under the assumption that $W_{jn} = W_{kn}$ for $\forall n$, we have,

$$\sum_{n=1}^{N} W_{jn}\delta_{jn} + \sum_{n=1}^{N} W_{kn}\delta_{kn} = \sum_{n=1}^{N} W_{jn}\delta'_{kn} + \sum_{n=1}^{N} W_{kn}\delta'_{jn} = \sum_{n=1}^{N} W_{kn}\delta'_{kn} + \sum_{n=1}^{N} W_{jn}\delta'_{jn}, \quad \text{(D.9)}$$

and hence

$$\sum_{m,n} W_{mn}\delta_{mn}$$

$$= \sum_{m,n \notin (j,k)} W_{mn}\delta_{mn} + 2\left(\sum_{n=1}^{N} W_{jn}\delta_{jn} + \sum_{n=1}^{N} W_{kn}\delta_{kn}\right)$$

$$= \sum_{m,n \notin (j,k)} W_{mn}\delta'_{mn} + 2\left(\sum_{n=1}^{N} W_{jn}\delta'_{jn} + \sum_{n=1}^{N} W_{kn}\delta'_{kn}\right)$$

$$= \sum_{m,n} W_{mn}\delta'_{mn},$$

where the first and third equalities use facts that $W_{mn} = W_{nm}$, $\delta_{mn} = \delta_{nm}$, and $W_{nn} = 0$ for $\forall n, m$. The second equality follows the result in (D.9).

In both cases, we have $\sum_{m,n} W_{mn}\delta_{mn} = \sum_{m,n} W_{mn}\delta'_{mn}$ for all $G \in \mathcal{G}_{ij}$ and therefore

$$\Pr(g_i = g_j | W) - \Pr(g_i = g_k | W) = 0. \quad \text{(D.10)}$$

$\square$

D.4 PC-KMeans

Proof of Theorem 2. We start with a brief discussion of *PC-KMeans* algorithm (Basu et al., 2004a). Given a set of observations $(y_1, y_2, \ldots, y_N)$, a set of positive-link constraints $\mathcal{P}$, a set of negative-link constraints $\mathcal{N}$, the cost of violating constraints $w = \{w^p_{ij}, w^n_{ij}\}$ and the number of groups K, the *PC-KMeans* algorithm aims to partition the N units into K groups

so as to minimize the following objective function,

$$L(G) = \underbrace{\frac{1}{2} \sum_{k=1}^{K} \sum_{i \in B_k} \|z_i - \mu_k\|^2}_{\text{within-cluster sum of squares}} + \underbrace{\sum_{(i,j) \in \mathcal{P}} \omega_{ij}^m \mathbf{1}\left(g_i \neq g_j\right) + \sum_{(i,j) \in \mathcal{N}} \omega_{ij}^c \mathbf{1}\left(g_i = g_j\right)}_{\text{cost of violation}}, \qquad \text{(D.11)}$$

where μ_k is the centroid of group k, i.e., $\mu_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i$, B_k is the set of units that are assigned to group k, and $|B_k|$ is the size of group k. Equation (D.11) can be rewritten as

$$L(G) = \frac{1}{2} \sum_{i=1}^{N} \|y_i - \mu_{g_i}\|^2 - \sum_{i,j} c W_{ij} \delta_{ij} + Const, \qquad \text{(D.12)}$$

where $Const = c \left(\sum_{(i,j) \in \mathcal{P}} W_{ij} - \sum_{(i,j) \in \mathcal{N}} W_{ij} \right)$ is a constant, c is the scaling constant introduced in (2.4), and

$$W_{ij} = \begin{cases} \frac{\omega_{ij}^m}{2c} & \text{if } (i,j) \in \mathcal{P} \\ -\frac{\omega_{ij}^c}{2c} & \text{if } (i,j) \in \mathcal{N} \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(D.13)}$$

The clustering process includes minimizing the objective function over both group partition G (assignment step) and the model parameters $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$ (update step). Next, we will show that the *PC-KMeans* algorithm is embodied in our proposed Gibbs sampler with soft constraints.

Under assumption (i), we can rewrite the baseline model with a set of variables z_{it} that don't have grouped heterogeneous effects on y_{it},

$$y_{it} = \alpha'_{g_i} x_{it} + \beta'_i z_{it} + \varepsilon_{it} = \alpha_{g_i} + \beta'_i z_{it} + \varepsilon_{it},$$

where the second equality holds due to the assumption of $x_{it} = 1$. β_i can be equal across units, i.e., $\beta_i = \beta$.

Under assumption (ii), we fix the number of groups upfront and thus we don't rely on slice sampling in which K is unknown and determined dynamically. Hereinafter, we focus on posterior distribution without the auxiliary variable u_i. Notice that the indicator function $\mathbf{1}(u_i < \pi_{g_i})$ in the posterior density reduces to π_{g_i}.

Part 1: Assignment Step

Assume we have soft pairwise constraints and weights are specified in (D.13). Under the

assumptions (iii) and (iv), the posterior density of the group membership indicators G is,

$$p(G|\alpha, \beta, \sigma^2, Y, X, Z, W)$$

$$= \frac{1}{Z_S} \prod_{i=1}^{N} \left[p(y_i | \beta_i, \alpha_{g_i}, \sigma^2_{g_i}, x_i, z_i) \pi_{g_i} \right] p(W|G)$$

$$= \frac{1}{Z_S} \prod_{i=1}^{N} p(y_i | \beta_i, \alpha_{g_i}, \sigma^2_{g_i}, x_i, z_i) \pi_{g_i} \prod_{i,j=1}^{N} \exp\left(\frac{cW_{ij}}{\sigma^2} \delta_{ij} \right)$$

$$= \frac{1}{Z_S} \prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{T}{2}} \pi_{g_i} \exp\left[-\frac{1}{2\sigma^2} \|\tilde{y}_i - \alpha_{g_i}\|^2 \right] \prod_{i,j=1}^{N} \exp\left(\frac{cW_{ij}}{\sigma^2} \delta_{ij} \right), \qquad \text{(D.14)}$$

where $\tilde{y}_i = y_i - \beta_i' z_i$, $z_i = [z_{i1}\ z_{i2} \dots z_{iT}]'$ and Z_S is the normalization constant.

Next, we define the optimal group partition G^* that minimizes the objective function of *PC-KMeans* defined in (D.12) with $x_i = \tilde{y}_i$ and $\mu_k = \alpha_k$, that is,

$$G^* \equiv \arg\min_G L(G)$$

$$= \arg\min_G \frac{1}{2} \sum_{i=1}^{N} \|\tilde{y}_i - \alpha_{g_i}\|^2 - \sum_{i,j} cW_{ij}\delta_{ij}. \qquad \text{(D.15)}$$

Now we consider the asymptotic behavior of the posterior probability in (D.14). We will show that as σ^2 goes to 0, the posterior probability of G approaches 0 for all group partitions except for G^*:

$$\lim_{\sigma^2 \to 0} p(G|\rho, \beta, \alpha, \Sigma, Y, X, W) \to \begin{cases} 1 & \text{if } G = G^*; \\ 0 & \text{otherwise.} \end{cases}$$

We start with the log posterior density of G in (D.14),

$$l(G) \equiv \log p(G|\rho, \beta, \alpha, \Sigma, Y, X, W)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \|\tilde{y}_i - \alpha_{g_i}\|^2 + \sum_{i,j=1}^{N} \left(\frac{cW_{ij}}{\sigma^2} \delta_{ij} \right)$$

$$- \frac{NT}{2} \log(2\pi\sigma^2) + \sum_{i=1}^{N} \log(\pi_{g_i}) - \log Z_S. \qquad \text{(D.16)}$$

The difference between two log posterior probabilities evaluated at G^* and any other G

is

$$l(G^*) - l(G)$$

$$= \frac{1}{\sigma^2} \left[\left(\frac{1}{2} \sum_{i=1}^{N} \|\tilde{y}_i - \alpha_{g_i}\|^2 - \sum_{i,j=1}^{N} cW_{ij}\delta_{ij} \right) - \left(\frac{1}{2} \sum_{i=1}^{N} \|\tilde{y}_i - \alpha_{g_i^*}\|^2 - \sum_{i,j=1}^{N} cW_{ij}\delta_{ij}^* \right) \right]$$

$$+ \sum_{i=1}^{N} \left[\log(\pi_{g_i^*}) - \log(\pi_{g_i}) \right]. \tag{D.17}$$

The first term is strictly positive according the definition of G^* in (D.15). For simplicity, we denote the expression within the first square brace as V and $V > 0$. The second term is finite since

$$\left| \sum_{i=1}^{N} \left[\log(\pi_{g_i^*}) - \log(\pi_{g_i}) \right] \right| \leqslant N \left| \max(\log(\pi_j)) - \min(\log(\pi_j)) \right| < +\infty$$

Thus, for any $G \neq G^*$, in the limit as $\sigma^2 \to 0$, we have

$$\lim_{\sigma^2 \to 0} l(G^*) - l(G) = \lim_{\sigma^2 \to 0} \frac{V}{\sigma^2} + \sum_{i=1}^{N} \left[\log(\pi_{g_i^*}) - \log(\pi_{g_i}) \right] = +\infty. \tag{D.18}$$

This indicates that

$$\lim_{\sigma^2 \to 0} \frac{p(G|\alpha, \sigma^2, Y, X, Z, W)}{p(G^*|\alpha, \sigma^2, Y, X, Z, W)} = \lim_{\sigma^2 \to 0} \exp\left[l(G) - l(G^*) \right] = \exp(-\infty) = 0.$$

We take the sum over all possible group partitions and get,

$$\lim_{\sigma^2 \to 0} \frac{\sum_{G'} p(G'|\alpha, \sigma^2, Y, X, Z, W)}{p(G^*|\alpha, \sigma^2, Y, X, Z, W)}$$

$$= \lim_{\sigma^2 \to 0} \frac{\sum_{G' \neq G} p(G'|\alpha, \sigma^2, Y, X, Z, W) + p(G^*|\alpha, \sigma^2, Y, X, Z, W)}{p(G^*|\alpha, \sigma^2, Y, X, Z, W)}$$

$$= 1.$$

Since $\sum_{G'} p(G'|\alpha, \sigma^2, Y, X, Z, W) = 1$, we have

$$\lim_{\sigma^2 \to 0} p(G^*|\alpha, \sigma^2, Y, X, Z, W) = 1. \tag{D.19}$$

Therefore, when $\sigma^2 \to 0$, every posterior draw of G from the proposed Gibbs sampler is the solution to the assignment step of the *PC-KMeans* algorithm, conditional on the posterior draws of other parameters.

Part 2: Update Step

During Gibbs sampling, once we have performed one complete set of Gibbs moves on the group assignments and non-group-specific parameters including β_i and σ^2, we need to sample the α_k conditioned on all assignments and observations.

Let $|B_k|$ be the number of units assigned to group k, then the posterior density for α_k read as

$$p(\alpha_k|\beta, \sigma^2, Y, X, Z) \propto \exp\left[-\frac{1}{2}(\alpha_k - \bar{\mu}_{\alpha_k})'\bar{\Sigma}_{\alpha_k}^{-1}(\alpha_k - \bar{\mu}_{\alpha_k})\right], \tag{D.20}$$

where

$$\bar{\Sigma}_{\alpha_k} = \left(\Sigma_\alpha^{-1} + |B_k|\sigma^{-2}I_T\right)^{-1},$$

$$\bar{\mu}_{\alpha_k} = \bar{\Sigma}_{\alpha_k}\left(\Sigma_\alpha^{-1}\mu_\alpha + \sigma^{-2}\sum_{i\in B_k}\tilde{y}_i\right),$$

$$\tilde{y}_i = y_i - \rho y_{-1,i} - x_i\beta_i.$$

We can see that the mass of the posterior distribution becomes concentrated around the posterior group mean $\bar{\mu}_{\alpha_k}$ as $\sigma^2 \to 0$. Meanwhile, the posterior group mean $\bar{\mu}_{\alpha_k}$ equals the group "sample" mean in the limit:

$$\lim_{\sigma^2\to0}\bar{\mu}_{\alpha_k} = \lim_{\sigma^2\to0}\left(\Sigma_\alpha^{-1} + |B_k|\sigma^{-2}I_T\right)^{-1}\left(\Sigma_\alpha^{-1}\mu_\alpha + \sigma^{-2}\sum_{i\in B_k}\tilde{y}_i\right)$$

$$= \lim_{\sigma^2\to0}\left(\sigma^2\Sigma_\alpha^{-1} + |B_k|I_T\right)^{-1}\left(\sigma^2\Sigma_\alpha^{-1}\mu_\alpha + \sum_{i\in B_k}\tilde{y}_i\right)$$

$$= |B_k|^{-1}\sum_{i\in B_k}\tilde{y}_i.$$

In other words, after we determine the assignments of units to groups, we update the means as the "sample" mean of the units in each group. This is equivalent to the standard *KMeans* cluster update step in general. Of course, we need additional steps to draw β_i and σ^2 before updating group means. □

E Monte Carlo Simulation

In this section, we conducted Monte Carlo simulations to examine the performance of various constrained BGFE estimators under different data generating processes (DGPs) and

prior belief on G. Two sets of DGPs are considered. For the simple DGPs, we introduce various group pattern in the fixed-effects only. The general DGPs, on the other hand, include more covariates with group-specific slope coefficients. Such designs enable us to investigate not only how our proposed estimators perform under various DGPs with specific features, but also the accuracy of estimating the number of groups.

We consider a short-panel environment in which the sample size is $N = 200$ and the time span is $T = 11$. As we focus on one-step ahead forecasts, the last observation of each unit serves as the hold-out sample for evaluation. A similar framework can be applied to H-step ahead forecasts by generating additional H observations. The true number of groups is set to $K_0 = 4$. Given N and K^0, we divide the entire sample into K^0 balanced blocks with N/K_0 units in each block.[19] For each DGP, 1,000 datasets are generated, and we run the block Gibbs samplers for each data set with $M = 5,000$ iterations after a burn-in of 5,000 draws.

E.1 Data Generating Processes

E.1.1 Simple DGPs

We begin with a simple dynamic panel data model with group pattern in the fixed-effects and no covariates or heteroskedasticity.

DGP 1 & 2:

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

where $\rho = 0.7$ and $\varepsilon_{it} \sim N(0,1)$. The distributions of initial values are selected to ensure the simulation paths are stationary. Idiosyncratic error ε_{it} are independent across i and t, and mutually independent. ε_{it} is also independent of all regressors.

We assume α_k has zero mean and takes the form $\alpha_k = m(k - 2.5)$, where m controls the cross-sectional variance of α_i. Two sets of α_k are specified: $m = 1.79$ such that var $(\alpha_k) = 1/4$ in DGP 1 and $m = 0.51$ such that var $(\alpha_k) = 1/50$ in DGP 2, see details in Appendix E.1.2. The difference in α_k between these two DGPs distinguishes their properties. As depicted in Figure E.1, the group pattern is readily apparent in DGP 1. Different groups of units are perfectly divided, and the simulated paths are pretty flat. DGP 2 has a less visible group structure than DGP 1 because the difference between group means of α_k is smaller. The simulated pathways are considerably noisier and fluctuate around the unconditional mean.

---

[19]If N/K_0 is not an integer, we assign $\lfloor N/K_0 \rfloor$ units for group 1,2,..,$K_0 - 1$ and the last group contains the remainder.

Figure E.1: Simulated Paths for Units in Different DGPs

(a) DGP 1: Sharp Group

(b) DGP 2: Noisy Group
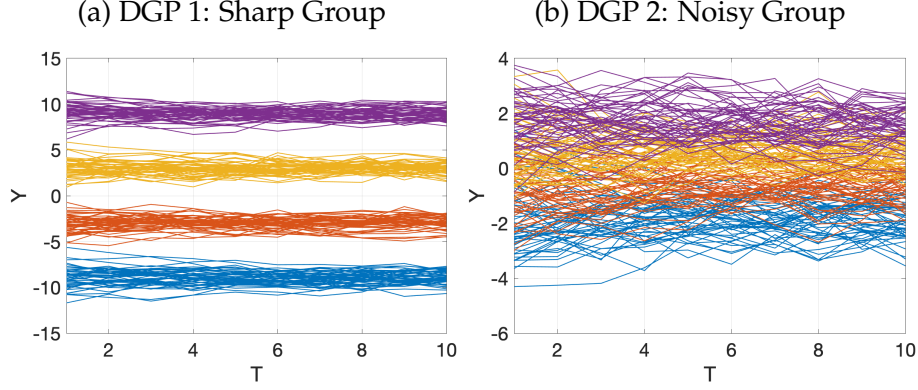
E.1.2 Details of the Simple DGPs

We start with the mean of α_k. Assume the values of α_k for group k take the form of $\alpha_k = m(k - c)$, where c is a shifting constant and m is a scaling constant. With loss of generality, we fix the mean of α_k to 0,

$$\sum_{k=1}^{K_0} \alpha_k = m \sum_{k=1}^{K_0} (k - c) = 0 \tag{E.2}$$

It immediately follows that $c = \frac{K_0+1}{2}$ and

$$\alpha_k = m \left(k - \frac{K_0 + 1}{2} \right), \quad \text{for } k = 1, 2, .., K_0. \tag{E.3}$$

Next, m is the only unknown coefficient in the DGP. To find a reasonable value for m, we connect it to the variance of α_i. As α_i are assume to be identical within a group, the sample variance of α_i is given by

$$V(\alpha_i) = \frac{1}{N} \sum_{i=1}^{N} \alpha_i^2 = \frac{1}{N} \frac{N}{K_0} \sum_{k=1}^{K_0} \alpha_k^2 = \frac{1}{K_0} \sum_{k=1}^{K_0} \alpha_k^2$$

Plugging in the expression of α_k in (E.3), we have

$$V(\alpha_i; m, K_0) = \frac{m^2}{K_0} \sum_{k=1}^{K_0} \left(k - \frac{K_0 + 1}{2} \right)^2. \tag{E.4}$$

To make the DGPs more comparable as more groups are considered, we assume $V(\alpha_i; m, K_0)$ is monotonically increasing in K_0^2, e.g., $V(\alpha_i; m, K_0) = V_0 K_0^2$ for some constant V_0. As a result,

we can deduct the value of m from (E.4),

$$m(K_0, V_0) = \left[\frac{V_0 K_0}{\sum_{k=1}^{K_0} \left(k - \frac{K_0+1}{2} \right)^2} \right]^{\frac{1}{2}}.$$

(E.5)

It is straightforward to find V_0 controls the dispersion of the underlying DGP. A larger V_0 indicate α_k are more separated and hence the group pattern become sharper, and vice versa.

E.1.3   General DGPs

The general DGP is based on the dynamic panel data model specified in (2.1) with an exogenous predictor z_{it} that has common effect for all units. This DGP incorporates group heterogeneity in the fixed-effects, the lagged term $x_{it}^{(1)} = y_{it-1}$ and an exogenous predictor $x_{it}^{(2)}$, as well as error variance $\sigma_{g_i}^2$.

DGP 3:

$$y_{it} = \alpha'_{g_i} x_{it} + \gamma z_{it} + \sigma_{g_i} \varepsilon_{it},$$

(E.6)

where $x_{it} = [1, x_{it}^{(1)}, x_{it}^{(2)}]'$, $\gamma = 1.5$, $y_{i0} \sim N(0,1)$ and $\varepsilon_{it} \sim N(0,1)$. For each i, the initial value is specified to guarantee that the time series $(y_{i0}, y_{i1}, ..., y_{iT})$ is strictly stationary. We assume there are $K^0 = 4$ balanced groups, with the true grouped coefficients summarized in Table (E.1). The AR(1) coefficients represent different degree of persistence persistence. The exogenous variable $x_{it}^{(2)}$ is drawn from $N(0,1)$ and z_{it} is drawn from $Gamma(1,1)$, capped by 10.

Table E.1: True Grouped Coefficients in the General DGP

	$\alpha_{0,k}$	$\alpha_{1,k}$	$\alpha_{2,k}$	σ_k^2
	(FE)	(lagged)	(exo var.)	(variance)
Group 1	-0.15	0.4	0.16	0.500
Group 2	-0.05	0.8	0.14	0.375
Group 3	0.05	0.5	0.12	0.250
Group 4	0.15	0.7	0.10	0.125

E.2   Construction of Soft Pairwise Constraints

Number of constraints: We set the number of constraints N_{PL} and N_{NL} as a function of N and K_0 to facilitate performance comparisons across different settings and to ensure that the

information of constraints does not vanish as N increases. Specifically, N_{PL} and N_{NL} are a predetermined proportion of the total number of correct constraints for each type which are given by,

$$N^*_{PL}(N,K) = KC^2_{N/K} = K\frac{N/K(N/K-1)}{2} = \frac{N(N-K)}{2K},$$ (E.7)

$$N^*_{NL}(N,K) = \left(\frac{N}{K}\right)^2 C^2_K = \left(\frac{N}{K}\right)^2 \frac{K(K-1)}{2} = \frac{N^2(K-1)}{2K}.$$ (E.8)

In the setting with $(N, K_0) = (200, 4)$, we have $N^*_{PL}(200, 4) = 4,900$ and $N^*_{NL}(200, 4) = 15,000$. We choose randomly select 5% of these constraints, leading to $N_{PL} = 245$ and $N_{NL} = 750$.

Type of pairwise constraints: The pairwise constraints are generated randomly. Given the number of PL constraints N_{PL}, each PL constraint is generated by randomly selecting a group and uniformly selecting two units within that group to be positive-linked. Similarly, for each of N_{NL} NL constraints, two unique groups are chosen at random and one unit is randomly selected from each. Regarding the remaining unselected units, we assume they are not restricted and have no prior belief on them.

Accuracy of pairwise constraints: Each constraint is annotated with a level of accuracy ψ generating from a transformed Beta distribution defined on $[0.5, 1]$. We begin by drawing ν from a Beta distribution: if the constraint is correct, $\nu \sim \text{Beta}(3, 2)$ for some $\alpha > 1$; otherwise, $\nu \sim \text{Beta}(2, 3)$. Then the level of confidence is $\psi = \frac{\nu}{2} + 0.5$ so that its domain is $[0.5, 1]$. We derive ψ in this manner to reflect the assumption that an expert should have less certainty in erroneous constraints than in correct ones.

Perturbation in pairwise constraints: To examine the performance with soft constraints under inaccurate prior belief, we artificially add errors to the randomly generated constraints. A fraction e of the constraints are mislabeled – a positive-link would be mislabeled as a negative-link and vice versa. We turn eN_{PL} true PL into NL and eN_{NL} true NL into PL with $e = 20\%$.

All DGPs are equipped with the same set of pairwise constraints, e.g., we only draw pairwise constraints and construct weights once.

E.3 Alternative Estimators

We explore various types of estimators that differ in the prior belief on G.

(i) *BGFE*: The baseline Bayesian grouped fixed-effects (BGFE) estimator are correctly-specified, i.e. assuming that the true model exhibits time-invariant grouped heterogeneity and that variance of error term is constant (varying) across units in the simple (general) DGPs. No prior belief on G is available for this estimator.

(ii) *BGFE-cstr*: The baseline BGFE estimator that takes pairwise constraints into consideration.

(iii) *BGFE-oracle*: This estimator is a variant of the BGFE estimator equipped with *known* true G.

We also evaluate the other Bayesian estimators with different prior assumptions on α_i that don't model group structure.

(iv) *Pooled*: Bayesian pooled estimator views α_i as a common parameter and, consequently, all units have the same prior level of α_i.

(v) *Flat*: flat-prior estimator assumes $p(\alpha_i) \propto 1$. There is no pooling across units in this case and α_i's are individually estimated using their own history. This also amounts to sampling from a posterior whose mode is the MLE estimate.

E.4 Posterior Predictive Densities and Performance Evaluation

E.4.1 Posterior Predictive Densities

Given S posterior draws, the posterior predictive distribution estimated from the MCMC draws is

$$\hat{p}(y_{iT+1}|Y, X) = \frac{1}{S} \sum_{j=1}^{S} \left[\sum_{k=1}^{K^{(j)}(G)} \mathbf{1}(g_i = k) p\left(y_{iT+1}|Y, X, \Theta^{(j)}\right) \right]. \tag{E.9}$$

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

E.4.2 Point Forecasts

The optimal posterior forecast under quadratic loss function is obtain by minimizing the posterior risk, with is the posterior mean. Conditional on posterior draws of parameters, the mean forecast can be approximated by the Monte Carlo averaging,

$$\hat{y}_{i,T+1|T} \approx \frac{1}{S} \sum_{j=1}^{S} \hat{y}_{iT+1|T}^{(j)} = \frac{1}{S} \sum_{j=1}^{S} \hat{\alpha}_{g_i}^{(j)\prime} x_{iT+1}, \tag{E.10}$$

and the RMSFE across units is given by

$$RMSFE_{T+1} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{i,T+1} - \hat{y}_{i,T+1})^2}. \tag{E.11}$$

E.4.3 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 $1 - \alpha$. In other words, it requires that the probability of $y_{iT+1} \in CS_{iT+1}$ conditional on having observed the history Y be at least $1 - \alpha$, i.e.,

$$P(y_{iT+1} \in CS_{iT+1}) \geqslant 1 - \alpha, \quad \text{for all } i, \tag{E.12}$$

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,

$$Cov_{T+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(y_{iT+1} \in CS_{iT+1}\right), \tag{E.13}$$

and (2) the average length of the sets C_{iT+1},

$$AvgL_{T+1} = \frac{1}{N} \sum_{i=1}^{N} \left(\delta_i^u - \delta_i^l\right). \tag{E.14}$$

E.4.4 Density Forecasts

To compare the performance of density forecasts for various estimators, we examine the continuous ranked probability score (Matheson and Winkler, 1976; Hersbach, 2000) across units. The continuous ranked probability score (CRPS) is frequently used to assess the respective accuracy of two probabilistic forecasting models. It is a quadratic measure of the difference between the predictive cumulative distribution function, $F_i^{T+1|T}(y)$, and the empirical CDF of the observation with the formula as follows,

$$\begin{aligned} CRPS_{T+1} &= \frac{1}{N} \sum_{i=1}^{N} CRPS(F_i^{T+1|T}, y_{iT+1}) \\ &= \frac{1}{N} \sum_{i=1}^{N} \int_0^\infty \left[F_i^{T+1|T}(y) - \mathbf{1}\left(y_{iT+1} \leqslant y\right)\right]^2 dy, \end{aligned} \tag{E.15}$$

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

In practice, the true predictive cumulative distribution function $F_i^{T+1|T}(y)$ or the PIT of y_{iT+1} is not available. We approximate it via the empirical distribution function for each unit

based on the posterior draws from the predictive density,

$$\hat{F}_i^{T+1|T}(y) = \frac{1}{S} \sum_{j=1}^{S} \mathbf{1}\left(y_{iT+1|T}^{(j)} \leq y\right),$$ (E.16)

Based on sorted posterior draws $\tilde{y}_{iT+1}^{(j)}$, we can calculate CRPS using the below representation by Laio and Tamea (2007),

$$CRPS\left(\hat{F}_i^{T+1|T}, y_{iT+1}\right) = \frac{2}{S^2} \sum_{j=1}^{S} \left(\tilde{y}_{iT+1|T}^{(j)} - y_{iT+1}\right)\left(\mathbf{1}\left\{y_{iT+1} < \tilde{y}_{iT+1|T}^{(j)}\right\} s - i + \frac{1}{2}\right).$$ (E.17)

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

$$LPS_{T+1} = -\frac{1}{N} \sum_{i=1}^{N} \ln \int p\left(y_{iT+1}|Y, X, \Theta\right) p(\Theta|Y, X) d\Theta,$$ (E.18)

where the expectation can be approximated using posterior draws,

$$\int p\left(y_{iT+1}|Y, X, \Theta\right) p(\Theta|Y, X) d\Theta \approx \frac{1}{S} \sum_{j=1}^{S} p\left(y_{iT+1}|Y, X, \Theta^{(j)}\right).$$ (E.19)

E.5 Simulation Results

E.5.1 Simple Dynamics Panel Data

To evaluate the advantage of pooling units into groups, we report the RMSE, bias, standard deviation, average length of 95% credible set, and frequentist coverage of the posterior estimate of ρ across Monte Carlo repetitions. For the fixed effects α, we only present the average bias as it may not be of importance for most empirical study.

The comparison across alternative estimators is shown in Table E.2. In DGP 1,the BGFE-cstr and BGFE estimators are equally accurate as the oracle estimator. This is not surprising because the units are well-separated by design, and the data provide sufficient information for the BGFE estimator to determine the group pattern. In this situation, prior knowledge of G or the true group indices has quite marginal influence. The pooled estimator, on the other hand, erroneously pools all groups together, resulting in inaccurate estimates of α_i and ρ. Despite the fact that the flat estimator treats units separately, it is still inferior to the BGFE-type estimators. This is because it cannot utilize cross-sectional information to estimates parameters in this short panel and hence bears much larger bias.

In DGP 2, where the group pattern is less apparent, the BGFE-cstr estimator is arguably the most accurate. In contrast to the standard BGFE estimator, it uses cross-sectional data and pairwise constraints to determine the group pattern. These properties substantially reduce the biases of $\hat{\beta}$ and $\hat{\alpha}_i$, enabling the BGFE-cstr estimator to outperform the unconstrained estimator by a significant margin and to perform comparable to the oracle estimator. Remember that we manually add 20% incorrect constraints into the prior knowledge. Despite the presence of these misspecified constraints, the BGFE-cstr estimator is still able to extract relevant information from constraints in order to enhance the overall performance. The BGFE estimator, however, is unable to correctly reconstruct the group structure due to the noisy data, which results in the algorithm improperly grouping the units and hence generating inaccurate estimates.

Table E.2: Monte Carlo: Estimates, Soft Constraint

			$\hat{\rho}$				$\hat{\alpha}_i$	Group
		RMSE	Bias	Std	AvgL	Cov	Bias	Avg K
DGP 1	BGFE-oracle	0.0104	0.0037	0.0072	0.0276	0.92	0.0371	4
	BGFE-cstr	0.0102	0.0030	0.0072	0.0282	0.94	0.0369	4.92
	BGFE	0.0103	0.0037	0.0071	0.0274	0.92	0.0377	4.4
	Pooled	0.3543	0.3543	0.0032	0.0125	0	1.7889	-
	Flat	0.1713	0.1711	0.0073	0.0283	0	0.8668	-
DGP 2	BGFE-oracle	0.0186	0.0030	0.0137	0.0527	0.95	0.0235	4
	BGFE-cstr	0.0202	0.0058	0.0143	0.0557	0.93	0.0373	5.06
	BGFE	0.0546	0.0443	0.0212	0.0809	0.66	0.1357	4.78
	Pooled	0.2920	0.2919	0.0077	0.0298	0	0.5060	-
	Flat	0.1170	0.0834	0.0131	0.0509	0.14	0.2344	-

Table E.3 provides a summary of the prediction performance of each estimator. In general, the conclusions of the one-step-ahead forecast agree with those of the estimation. In DGP 1, the performance of the three BGFE estimators are quite similar, followed by the flat and pooled estimators. In DGP 2, the BGFE-cstr estimator, which utilizes prior belief on G, beats the other feasible estimators in point, set, and density forecast and is comparable to the oracle estimator.

E.5.2   General Panel Data

As the number of parameters increases for DGP 3, we present the RMSE and absolute bias of $\alpha_{g_i} = [\alpha_{1,g_i} \ \alpha_{2,g_i} \ \alpha_{3,g_i}]'$ and γ, as well as metrics for point and density prediction. In addition, all BGFE estimators now account for heteroskedasticity because the cross-sectional variance in DGP 3 is informative to group structure. As a result, we have the *BGFE-he-oracle*, *BGFE-he-cstr* and *BGFE-he* estimators in this exercise, where "*he*" denotes heteroskedasticity. For comparison, we also offer the *BGFE-ho-cstr* estimator, which assumes homoskedasticity, and *Flat-he* estimator, which is the heteroskedastic flat estimator.

Table E.3: Monte Carlo: Forecast, Soft Constraint

		Point Forecast			Set Forecast		Density Forecast	
		RMSFE	Error	Std	AvgL	Cov	LPS	CRPS
	BGFE-oracle	0.4989	0.0001	0.4989	1.9627	0.95	0.7254	0.2818
	BGFE-cstr	0.4991	0.0004	0.4990	1.9666	0.95	0.7256	0.2819
DGP 1	BGFE	0.4990	0.0001	0.4990	1.9616	0.95	0.7255	0.2818
	Pooled	0.6401	0.0006	0.6404	3.0114	0.98	1.0064	0.3657
	Flat	0.5620	0.0003	0.5622	2.4265	0.97	0.8544	0.3184
	BGFE-oracle	0.4990	0.0001	0.4989	1.9629	0.95	0.7254	0.2819
	BGFE-cstr	0.5021	0.0001	0.5021	1.9790	0.95	0.7314	0.2836
DGP 2	BGFE	0.5186	0.0002	0.5187	2.0546	0.95	0.7633	0.2930
	Pooled	0.5396	0.0005	0.5395	2.2444	0.96	0.8079	0.3052
	Flat	0.5286	0.0002	0.5287	2.1165	0.95	0.7841	0.2987

Table E.4: Results for Estimation, Point Forecast and Estimated K

	Estimates								Forecast		Group	
	$R(\hat{\alpha}_0)$	$B(\hat{\alpha}_0)$	$R(\hat{\alpha}_1)$	$B(\hat{\alpha}_1)$	$R(\hat{\alpha}_2)$	$B(\hat{\alpha}_2)$	$R(\hat{\gamma})$	$B(\hat{\gamma})$	RMSFE	LPS	AvgK	PctK
Flat-he	0.258	0.199	0.131	0.098	0.200	0.149	0.026	0.014	0.667	1.108	-	-
BGFE-he-oracle	0.126	0.137	0.092	0.101	0.119	0.132	0.569	0.602	0.840	0.706	4	1
BGFE-he-cstr	0.171	0.164	0.290	0.179	0.125	0.135	0.566	0.608	0.847	0.716	4.087	0.914
BGFE-ho-cstr	0.218	0.198	0.328	0.220	0.140	0.144	0.625	0.671	0.850	0.769	4.342	0.682
BGFE-he	0.303	0.317	0.560	0.482	0.137	0.147	0.601	0.640	0.871	0.756	3.575	0.534
Pooled	0.444	0.503	1.262	1.527	0.131	0.148	0.734	0.805	0.993	0.910	-	-

Notes: The first line gives the levels of the each metrics based on the Flat-he estimator, which is the benchmark model, and the following lines in the columns head "Estimates" and "Forecast" present ratios of the respective method relative those based on the flat-he estimator. In the columns head "Group", we show the average of number of groups (AvgK) and the percentage of iterations that the posterior sampler selects K_0 (PctK) averaged over 1,000 runs of algorithm. $R(\cdot)$ is RMSE of the posterior mean estimator. $B(\cdot)$ is the absolute bias of the posterior mean estimator.

Table E.4 presents the relative performance of estimation and forecasting for DGP 3. The benchmark model is *Flat-he*. Several findings arise. First, the gain from incorporating pairwise constraints is evident. It reduces the RMSE and bias for all parameters and improve both point and density forecast, when comparing BGFE-he-cstr to BGFE-he. The percentage of the Gibbs sampler that visits the true number of groups K_0 grows considerably from 53.4% to 91.4%. Even when pairwise constraints are taken into account, this percentage is just 68.2% if heteroskedasticity is ignored. Second, when we include prior belief on G, the improvement in α_{1,g_i}, the AR coefficient, is the greatest among all three grouped coefficients with a bias reduction of more than 60%. This also suggests that the AR coefficient may be more sensitive to the estimated group structure. Thirdly, BGFE-he-cstr and BGFE-ho-cstr have comparable RMSFE values, but BGFE-he-cstr has a significantly lower LPS, showing that modeling heteroskedasticity in the current setting is favorable for the density forecast. The empirical results below also confirm this finding. Lastly, all BGFE-type estimators generate similar estimates for the exogenous variables that don't have group effects on y_{it} as the

improvement in α_{2,g_i} and γ are marginal when prior belief on group is included or when true group is imposed.

E.5.3   Computational Time

We compare the running times of each estimator using the simulation with the general DGPs. For each estimator, we run the block Gibbs samplers with 5,000 iterations after a burn-in of 5,000 draws. We calculate the running time by averaging over 1,000 repetitions. All programs, including this exercise, the simulations above, and the empirical analysis in Section 4 were executed in Matlab 2020a running on a server with Intel Skylake CPUs and 32 GB of RAM.

Table E.5 shows the running time in seconds. It suggests that modeling group structure does not considerably increase computational load. The running time of the BGFE-he estimator is comparable to that of the Flat estimator when the number of observations is small. As T increases, the BGFE-he estimators take slightly more time to finish than the flat estimator. The BGFE-he-cstr estimator requires approximately 20% more time to complete the calculations than the BGFE-he estimator. This is because BGFE-he-cstr adapts pairwise constraints and has additional term to evaluate in the posterior distribution of the group indices. Notice that we fix the scaling constant c in this experiment. If we find the optimal c using grid search as proposed in Section 4.1.2, the running time of BGFE-he-cstr will be m times as great as what is shown below, where m is the size of the grid.

Table E.5: Running Time in Seconds

N	T	BGFE-he-cstr	BGFE-he	Flat	Pooled
	5	74.9	62.1	1.2	61.7
100	10	82.5	69.8	1.3	64.7
	20	86.1	75.7	1.3	62.4
	5	145.0	111.9	1.2	114.3
200	10	146.2	118.8	1.3	113.0
	20	159.4	133.2	1.5	117.0

It is possible to speed the algorithm up especially when we select the optimal value for c. Wang and Dunson (2011) propose the sequential updating and greedy search (SUGS) algorithm, which essentially simplifies the algorithm 1 with approximation. The SUGS algorithm allows fast yet accurate approximate Bayes inferences under DP priors with just a single cycle of simple calculations for each unit, while also producing marginal likelihood estimates to be used in selecting the constant c. This will be left for future study.

F   Data Description

F.1   Inflation of the U.S. CPI Sub-Indices

The seasonally adjusted series of CPI for All Urban Consumers (CPI-U) for subcategories at all display level are obtained from the BLS Online Databases.[20] The raw data contains 318 series, which are recorded on a monthly basis and spanned the period from January 1947 to August 2022. Notice that the raw dataset doesn't include an indicator for the expenditure categories. We manually merge the raw dataset with the table of content of CPI entry level items[21] by entry level item (ELI) code, the series description, and universal classification codes (UCC), if necessary.[22]

Series can enter and exit the sample. The BLS discontinued and launched series on a regular basis owing to changes in source data and methodology, for example, see the Post for the updates on series since 2017. The measure of certain subcategories was impacted by the Pandemic and hence missing. Since the Pandemic, the related activities and venues (sports events, bars, schools) were canceled and close temporarily, such as admission to sporting events (SS62032), distilled spirits away from home (SS20053), food at elementary and secondary schools (SSFV031A), etc. We chose to not impute the missing values since there was no clear benchmark to compare with, especially given the depressed economic conditions.

The CPI-U consists of eight major expenditure categories (1) Apparel; (2) Education and Communication; (3) Food and beverages; (4) Housing; (5) Medical Care; (6) Recreation; (7) Transportation; (8) Other goods and services. Each major category contains multiple sub-categories, resulting in a hierarchy of categories with increasing specificity. BLS provides a detailed table[23] that records the series code, series name, and display level. We resort to the display level to build the tree structure of the CPI sub-indices and eliminate those parent nodes, as illustrated in Figure F.1.

---

[20]https://data.bls.gov/PDQWeb/cu

[21]https://www.bls.gov/cpi/additional-resources/entry-level-item-descriptions.xlsx

[22]Some series are labeled by UCC rather than ELI. The concordance provided by the BLS can be found here: https://www.bls.gov/cpi/additional-resources/ce-cpi-concordance.htm.

[23]https://download.bls.gov/pub/time.series/cu/cu.item.

Figure F.1: Hierarchical Structure of CPI: Eliminating Parent Nodes
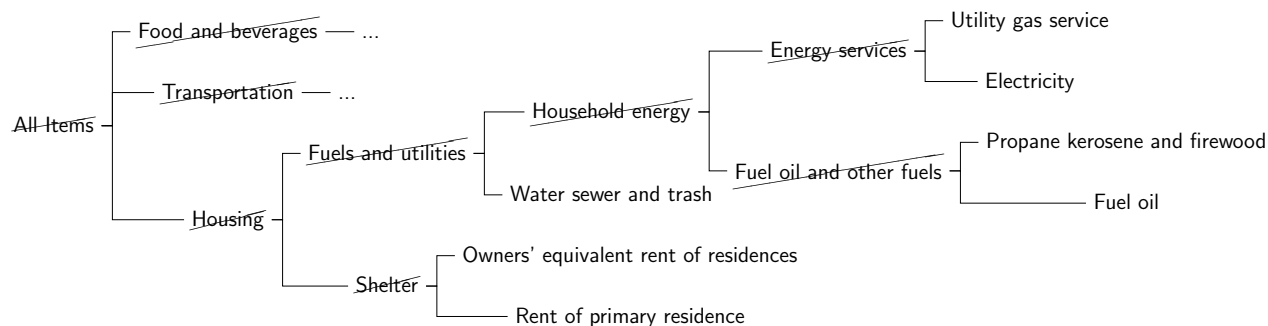
Because all CPI data is available on a monthly basis, we use the unemployment gap as labor market slack measures in the Phillips curve model. We use the seasonally adjusted unemployment rate[24] from FRED and construct the "gap" measures using the Hamilton filter (Hamilton, 2018). The Hamilton filter has two parameters: number of lags p and number of lookahead periods h. We follow Hamilton's suggestion and set $h = 24$ and $p = 12$, or an AR(12) process, additionally lagged by 24 lookahead periods for the monthly time series.

F.2   Income and Democracy

All data in this section are taken from the replication files of BM.[25] The data set contains a balanced panel of 89 countries and 7 periods at a five-year interval over 1970-2000. The main measure of democracy is the Freedom House Political Rights Index. A country receives the highest score if political rights come closest to the ideals suggested by a checklist of questions, beginning with whether there are free and fair elections, whether those who are elected rule, whether there are competitive parties or other political groupings, whether the opposition plays an important role and has actual power, and whether minority groups have reasonable self-government or can participate in the government through informal consensus. See more details in Acemoglu et al. (2008), Section 1.

Table F.1 contains descriptive statistics for the main variables. The sample period is 1970–2000, and each observation corresponds to five-year intervals. The table shows these statistics for all countries and also for high- and low-income countries, split according to the median of the countries' averaged income. The comparison of high- and low-income countries in the medium and lower panels reveals the pattern that richer countries tend to be more democratic.

---

[24]https://fred.stlouisfed.org/series/UNRATE
[25]https://www.dropbox.com/s/ssjabvc2hxa5791/Bonhomme_Manresa_codes.zip?dl=0

Table F.1: Summary Statistics for the Democracy Data Set

	Mean	Median	S.E.	Min	Max
Full Sample					
Democracy index	0.5535	0.5000	0.3727	0	1.0000
GDP per capita (in log)	8.2534	8.2444	1.0763	5.7739	10.4450
High-Income					
Democracy index	0.7852	1.0000	0.2934	0	1.0000
GDP per capita (in log)	9.1490	9.1975	0.6079	7.4970	10.4450
Low-Income					
Democracy index	0.3247	0.1667	0.2973	0	1.0000
GDP per capita (in log)	7.3576	7.3208	0.6051	5.7739	8.81969

G   Additional Empirical Results

G.1   Inflation of the U.S. CPI Sub-Indices

Figure G.1 shows the posterior mean of the number of active groups (red) along with total number of available CPI sub-indices (dark blue). Note that we choose not to show the credible set or the distribution of K because the distributions are concentrated around a single number in most samples.

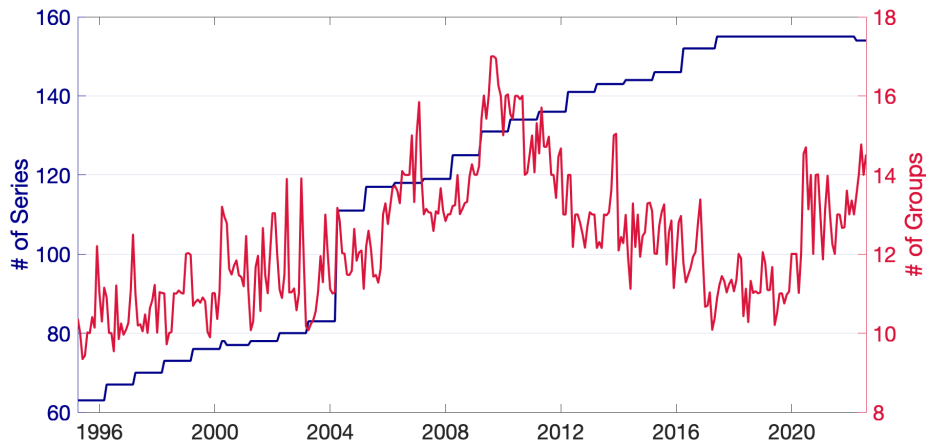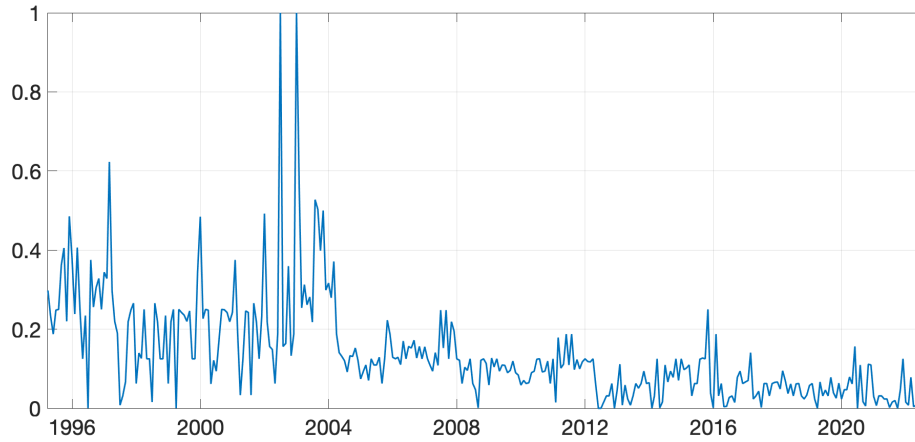Figure G.1: Number of Active Groups, BGFE-he-cstr

Figure G.2 shows the selected scaling constant c over time.

Table G.1 and G.2 present the ratio of RMSE between each estimator and AR-he for each

Figure G.2: Scaling Constant, BGFE-he-cstr

expenditure category in five periods: 1995-1999, 2000-2004, 2005-2009, 2010-2014, 2015-2019, 2020-2022. Similar results for the difference of LPS are shown in Table G.3 and G.4.

G.1.1 Network Visualization of Posterior Similarity Matrix

In our empirical work, we estimate the posterior similarity matrix (PSM) among 154 series for the last sample (August 2022). Presenting and examining $154 \times 154 = 23,716$ estimates pairwise posterior probabilities in PSM would be thoroughly uninformative. Hence we characterize the estimated PSM graphically as network graphs, which contain node names, node color, and link size (one per link since the network is undirected).

- *Node name* shows the item names of CPI sub-indices.
- *Node color* indicates the group structure used in the prior, e.g, expenditure category.
- *Link size* represents the pairwise probabilities in the PSM.

We use the *qgraph* package in R for network visualization. Node locations are determined by a modified version of a force-embedded algorithm proposed by Fruchterman and Reingold (1991). Figure G.3 show the full-sample CPI sub-index network graphs.

Table G.1: Relative RMSE by Expenditure Category and Period

	Full	95 - 99	00 - 04	05 - 09	10 - 14	15 - 19	20 - 22
Average of All Series							
BGFE-he-cstr	0.97	0.97	0.97	0.99	0.96	0.97	0.94
BGFE-he	0.98	0.98	0.98	1.00	0.97	0.98	0.95
BGFE-ho	1.00	1.04	0.98	1.00	0.98	0.99	1.00
AR-he-PC	1.01	1.02	1.03	0.99	1.00	1.01	1.04
Pooled	1.00	1.04	0.98	1.03	1.01	0.99	0.94
Category 1: Apparel							
BGFE-he-cstr	0.97	0.98	0.99	0.97	0.96	0.98	0.91
BGFE-he	0.97	0.98	0.99	0.98	0.97	0.98	0.92
BGFE-ho	0.98	0.99	1.00	0.98	0.98	0.99	0.93
AR-he-PC	1.02	1.02	1.00	1.03	1.01	1.01	1.08
Pooled	0.99	1.00	0.97	0.99	1.03	1.03	0.90
Category 2: Education and Communication							
BGFE-he-cstr	0.96	0.93	0.99	0.95	0.94	0.97	0.93
BGFE-he	0.95	0.95	1.00	0.94	0.95	0.97	0.92
BGFE-ho	1.08	2.18	1.02	1.00	1.00	0.97	0.91
AR-he-PC	1.01	1.03	0.99	1.00	0.95	1.01	1.08
Pooled	1.16	2.24	1.21	1.05	1.12	1.03	0.97
Category 3: Food and Beverages							
BGFE-he-cstr	0.98	0.99	0.99	1.00	0.97	0.98	0.95
BGFE-he	0.98	0.99	1.00	1.00	0.97	0.98	0.95
BGFE-ho	0.99	1.03	0.99	0.97	0.97	0.98	0.95
AR-he-PC	1.00	1.01	1.02	0.99	1.00	1.01	0.98
Pooled	1.01	1.03	1.01	1.05	1.01	0.99	0.93
Category 4: Housing							
BGFE-he-cstr	0.97	0.96	0.95	1.03	0.94	0.96	0.96
BGFE-he	0.96	0.97	0.93	1.03	0.94	0.96	0.96
BGFE-ho	0.98	1.14	0.93	1.02	0.96	0.97	0.95
AR-he-PC	1.03	1.03	1.04	1.04	0.99	1.01	1.07
Pooled	0.99	1.14	0.92	1.06	0.97	0.97	0.94

Notes: Benchmark model = AR-he.

Table G.2: Relative RMSE by Expenditure Category and Period, *cont.*

	Full	95 - 99	00 - 04	05 - 09	10 - 14	15 - 19	20 - 22
Category 5: Medical Care							
BGFE-he-cstr	0.95	0.93	0.98	0.95	1.01	0.96	0.87
BGFE-he	0.95	0.93	0.98	0.95	1.01	0.96	0.85
BGFE-ho	1.06	1.32	1.15	1.07	1.09	0.99	0.86
AR-he-PC	1.03	1.04	0.99	1.02	1.03	1.02	1.07
Pooled	1.15	1.30	1.38	1.23	1.15	1.04	0.92
Category 6: Recreation							
BGFE-he-cstr	0.97	0.97	0.99	0.99	0.98	0.96	0.94
BGFE-he	0.97	0.97	1.00	0.99	0.99	0.96	0.92
BGFE-ho	1.02	1.18	1.09	1.05	1.00	0.97	0.94
AR-he-PC	1.03	1.03	1.03	1.02	1.01	1.01	1.08
Pooled	1.12	1.17	1.24	1.31	1.17	0.99	0.96
Category 7: Transportation							
BGFE-he-cstr	0.99	0.97	1.01	0.99	0.99	1.00	0.96
BGFE-he	0.99	0.97	1.01	0.98	0.99	1.00	0.96
BGFE-ho	1.03	1.07	1.02	1.00	1.01	1.02	1.07
AR-he-PC	1.02	1.04	1.05	0.97	1.01	0.99	1.07
Pooled	0.99	1.07	1.02	0.99	0.97	0.96	0.97
Category 8: Other Goods and Services							
BGFE-he-cstr	0.97	0.98	0.99	1.02	0.89	0.96	0.93
BGFE-he	0.95	0.94	0.95	1.05	0.89	0.97	0.91
BGFE-ho	0.96	0.99	0.97	0.99	0.89	0.97	0.91
AR-he-PC	1.02	1.01	1.02	1.02	0.97	1.00	1.11
Pooled	0.98	0.99	1.04	0.99	0.89	0.99	0.94

Notes: Benchmark model = AR-he.

Table G.3: Relative LPS, by Expenditure Category and Period

|	Full	95 - 99	00 - 04	05 - 09	10 - 14	15 - 19	20 - 22
Average of All Series							
BGFE-he-cstr	−0.08	−0.08	−0.07	−0.08	−0.09	−0.07	−0.08
BGFE-he	−0.06	−0.06	−0.05	−0.06	−0.07	−0.05	−0.08
BGFE-ho	0.64	0.77	0.72	0.77	0.48	0.53	0.45
AR-he-PC	0.01	0.02	0.02	0.01	0.01	0.01	0.01
Pooled OLS	0.66	0.84	0.73	0.79	0.53	0.52	0.44
Category 1: Apparel							
BGFE-he-cstr	−0.05	−0.03	−0.02	−0.05	−0.05	−0.02	−0.14
BGFE-he	−0.04	−0.02	−0.02	−0.04	−0.05	−0.01	−0.14
BGFE-ho	0.10	0.25	0.07	0.07	0.11	0.06	0.03
AR-he-PC	0.03	0.03	0.01	0.03	0.02	0.01	0.09
Pooled	0.12	0.25	0.07	0.10	0.18	0.11	−0.05
Category 2: Education and Communication							
BGFE-he-cstr	−0.12	−0.06	−0.07	−0.19	−0.17	−0.10	−0.11
BGFE-he	−0.13	−0.08	−0.11	−0.20	−0.16	−0.09	−0.12
BGFE-ho	0.92	1.44	1.04	0.87	0.91	0.56	0.56
AR-he-PC	0.01	0.00	0.01	0.02	0.00	0.01	0.07
Pooled	0.98	1.45	1.09	0.95	0.99	0.62	0.62
Category 3: Food and Beverages							
BGFE-he-cstr	−0.04	−0.03	−0.02	−0.06	−0.05	−0.05	−0.07
BGFE-he	−0.04	−0.04	−0.02	−0.05	−0.05	−0.05	−0.08
BGFE-ho	0.37	0.66	0.36	0.34	0.27	0.39	0.03
AR-he-PC	0.00	0.01	0.03	−0.01	0.01	0.00	−0.08
Pooled	0.44	0.84	0.41	0.44	0.33	0.41	0.03
Category 4: Housing							
BGFE-he-cstr	−0.12	−0.13	−0.14	−0.10	−0.11	−0.12	−0.07
BGFE-he	−0.11	−0.12	−0.13	−0.10	−0.11	−0.12	−0.07
BGFE-ho	0.81	0.96	1.13	0.83	0.60	0.60	0.70
AR-he-PC	0.02	0.00	0.03	0.04	0.01	0.01	0.06
Pooled	0.82	0.97	1.13	0.88	0.63	0.60	0.68

Notes: Benchmark model = AR-he.

Table G.4: Relative LPS by Expenditure Category and Period, *cont.*

	Full	95 - 99	00 - 04	05 - 09	10 - 14	15 - 19	20 - 22
Category 5: Medical Care							
BGFE-he-cstr	−0.15	−0.29	−0.14	−0.14	−0.16	−0.05	−0.09
BGFE-he	−0.15	−0.29	−0.13	−0.14	−0.16	−0.05	−0.09
BGFE-ho	1.16	1.64	1.34	1.11	1.14	0.78	0.78
AR-he-PC	0.02	0.02	0.01	0.03	0.02	0.01	0.04
Pooled	1.21	1.64	1.39	1.20	1.20	0.84	0.84
Category 6: Recreation							
BGFE-he-cstr	−0.04	−0.04	−0.05	−0.02	−0.02	−0.05	−0.05
BGFE-he	−0.03	−0.03	−0.05	−0.01	−0.01	−0.04	−0.07
BGFE-ho	0.55	0.91	0.79	0.58	0.49	0.25	0.12
AR-he-PC	0.02	0.03	0.03	0.02	0.01	0.01	0.06
Pooled	0.61	0.91	0.84	0.70	0.58	0.29	0.13
Category 7: Transportation							
BGFE-he-cstr	−0.04	−0.10	−0.03	−0.05	−0.07	0.00	0.00
BGFE-he	−0.03	−0.10	0.00	−0.04	−0.06	0.00	0.00
BGFE-ho	1.59	0.98	1.56	2.83	0.79	1.51	2.11
AR-he-PC	0.01	0.03	0.03	−0.02	−0.01	0.00	0.07
Pooled	1.42	0.98	1.41	2.38	0.75	1.24	1.98
Category 8: Other Goods and Services							
BGFE-he-cstr	−0.03	0.22	−0.09	−0.07	−0.14	−0.06	−0.05
BGFE-he	−0.03	0.25	−0.09	−0.06	−0.13	−0.06	−0.06
BGFE-ho	0.64	0.59	0.66	0.69	0.83	0.53	0.49
AR-he-PC	0.01	0.02	0.00	0.03	0.00	0.01	0.04
Pooled	0.69	0.61	0.71	0.77	0.89	0.57	0.53

Notes: Benchmark model = AR-he.

Figure G.3: Individual CPI Sub-Index Network Graph based on Posterior Similarity Matrix, August 2022
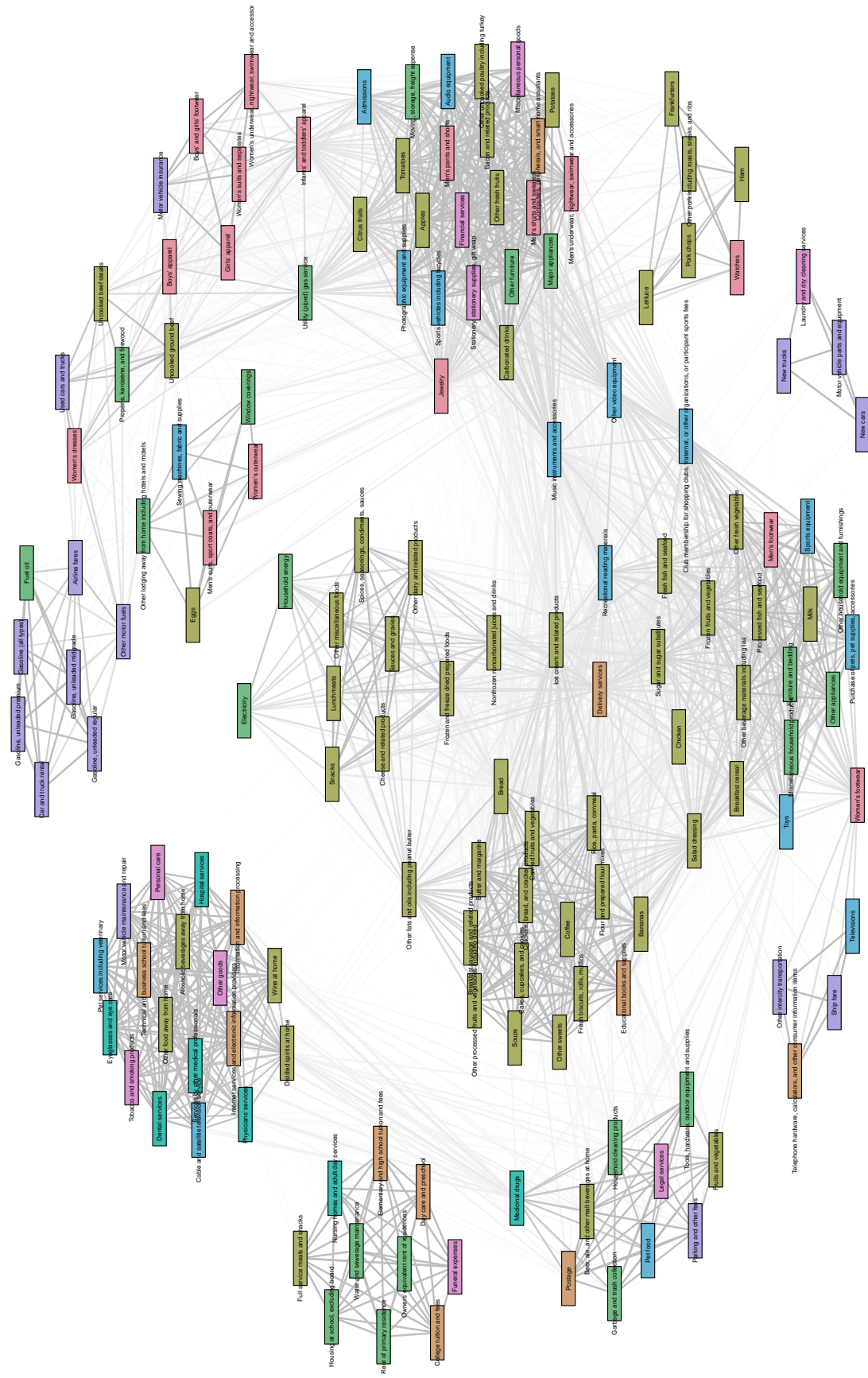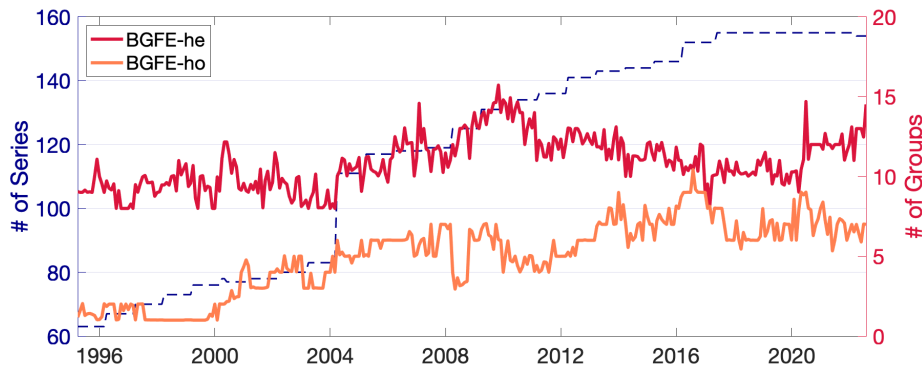
G.1.2 Heteroskedasticity vs. Homoskedasticity

We conclude by examining how grouped heteroskedasticity impacts forecast accuracy and why this is important. For illustrative purposes, we focus on the two BGFE estimators, BGFE-he and BGFE-ho, that do not involve pairwise constraints.

A distinguishing characteristic between BGFE-he and BGFE-ho is the estimated number of groups. Figure G.4 depicts the number of groups over samples. BGFE-he estimator forms 9 groups for the beginning of the sample, and increase it during the Great Recession and the Pandemic. However, the estimated number of groups for BGFE-ho is rather low in the 1990s, and progressively increases to around seven by the end of the sample. It is noticeable that when heteroskedasticity is allowed, there are more groups than when it is not. This is intuitive. Two groups can be expected to have comparable estimates of grouped fixed-effects and slope coefficients, but vastly different error variances. As a result, allowing for heteroskedasticity would result in a more refined group structure and increase the overall number of groups.

Figure G.4: Number of Groups

As seen in Figure 6 and 7, the grouped heteroskedasticity doesn't improve the point forecast but the density forecast. Figure G.5 depicts a clear perspective of it and demonstrates the performance of point and density forecasts through time. In panel (a), we observe that the ratio of RMSE is generally around one over the whole sample, meaning that heteroskedasticity cannot improve the point forecast in general. In panel (b), the difference in LPS is consistently negative. This demonstrates that the improved density prediction performance is not a fluke and that enabling heteroskedasticity improves the density forecast regardless of sample. This is actually in line with the simulation results presented in Table E.4.

Density forecasts vary substantially across categories. We pick three typical sub-categories and plot their posterior predictive densities of August 2022 in Figure G.6. The vertical dashed black lines represent the actual values. Several insights emerge while comparing these three subcategories. First, BGFE-he and BGFE-ho provide comparable posterior means for all three subcategories - the posterior predictive densities concentrate around the similar

Figure G.5: Results of BGFE-he vs. BGFE-ho

(a) RMSE$_{he}$ / RMSE$_{ho}$

(b) LPS$_{he}$ - LPS$_{ho}$
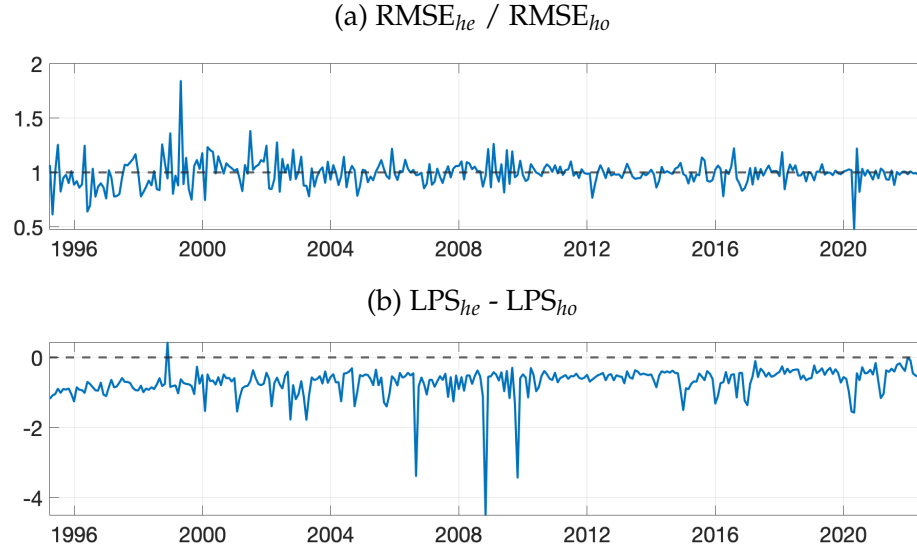
price levels. This explains why BGFE-he and BGFE-ho have comparable results in point forecasting. Second, BGFE-he reveals different predictive variance. As the rolling sample size is set to 4 years, all observations throughout the Pandemic are included, and it is anticipated that the price levels of elementary and high school (basic education) tuition and fees, major appliances, and airline prices would respond differently to the shock. Intuitively, education tuition and fees should not fluctuate as much as other prices, while airline fares have been strongly influenced by the fluctuating oil prices since the beginning of the Pandemic. Consequently, accounting for heteroskedasticity successfully captures this characteristic, such that college tuition and fees have a smaller predictive variance than that of BGFE-ho, but airline fares have a greater predictive variance.

Combing these two observations together reveals why BGFE-he has a better density forecast: the capacity to optimally cluster units according to the error variance and accommodate heteroskedasticity. For elementary and high tuition and fees, providing that both BGFE-he and BGFE-ho yields accurate posterior mean, BGFE-he yields much lower predictive variance, decreasing the LPS dramatically. Both BGFE-he and BGFE-ho underestimate the inflation rate for airline fares, but BGFE-he subtly creates a greater predicted variance to account for the wild probable shift in this sub-category and hence reduces the LPS significantly. Major appliances is an example to show that BGFE-he and BGFE-ho generate comparable density forecasts for some sub-categories.

Figure G.6: Predictive Posteriors for Selected Series: BGFE-he vs. BGFE-ho

(a) Basic Edu. Tuition and Fees

(b) Major Appliances

(c) Airline Fares
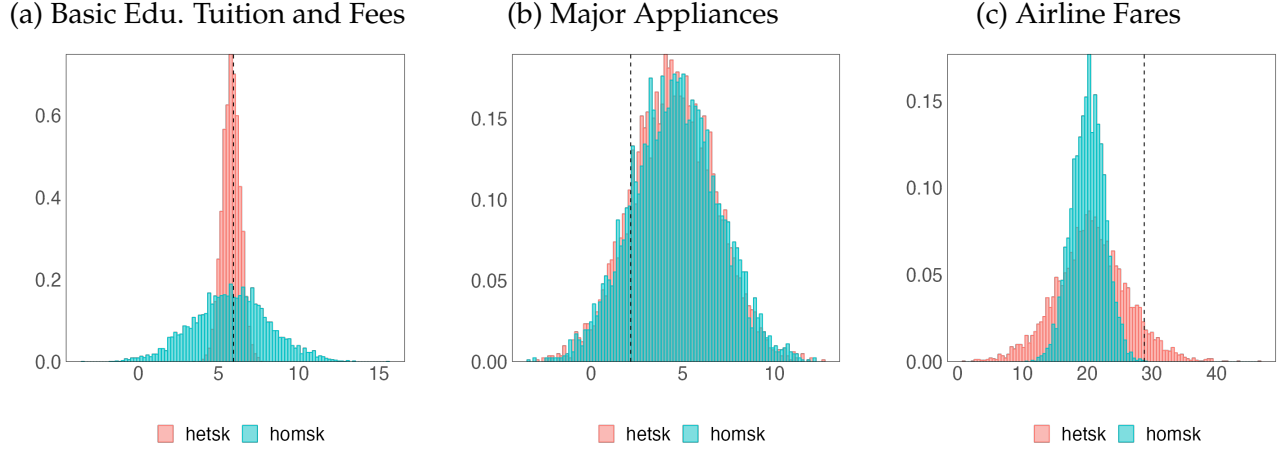
G.2 Income and Democracy

G.2.1 Results of Specification 1

We start our analysis with the specification 1 in (4.15). Table G.5 demonstrates the posterior probability of the number of groups utilizing various estimators. Notably, the BGFE-ho in this specification is identical to the primary model in BM, allowing us to evaluate the optimal number of groups. BGFE-ho creates 8 groups in all posterior draws, which is consistent with BM's conclusion of using BIC: the upper bound of the true number of groups is 10. Despite the fact that BM is unable to validate the ideal number of groups for their study, our BGFE-ho estimator provides an accurate estimate of it. Intriguingly, accounting for heteroskedasticity drastically reduces the number of groups, with BGFE-he identifying three groups in 92.9% of posterior draws. Adding pairwise constraints based on geographic information increase the number groups. Two-third of posterior draws from BGFE-he-cstr generate 5 group.

Table G.5: Probability for the number of groups

	BGFE-he-cstr	BGFE-he	BGFE-ho
$Pr(K < 3)$	0.000	0.000	0.000
$Pr(K = 3)$	0.000	**0.929**	0.000
$Pr(K = 4)$	0.344	0.071	0.000
$Pr(K = 5)$	**0.656**	0.000	0.000
$Pr(K > 5)$	0.000	0.000	**1.000**

The marginal data density (MDD) of each estimators in Table G.6 provides some insight on different models. Even while BGFE-ho produces eight groups and has a tendency to overfit, its MDD is the lowest of the three estimators. BGFE-he with fewer groups is superior to BGFE-ho with higher MDD. BGFE-he-cstr has the highest MDD because the pairwise

constraints give direction on grouping and identify the ideal group structure, which BGFE-he cannot uncover without our prior knowledge.

Table G.6: Marginal Data Density

BGFE-he-cstr	BGFE-he	BGFE-ho
425.690	381.218	368.918

We focus on the BGFE-he-cstr estimator and use the approach outlined in Section 3.2 to identify the unique group partitioning $\widehat{G}$. The left panel of Figure G.7 presents the world map colored by $\widehat{G}$, while the right panel present the group-specific averages of democracy index over time. The estimated group structure $\widehat{G}$ features four distinct groups, which is coincident to the choice of BM. As described in BM, we refer to groups 1-4 as the "high-democracy", "low-democracy", "early transition", and "late transition" group, respectively. With the exception of the "early transition" group that is slight at odd with the counterpart in BM, the group-specific averages of the democracy index for all other groups are relatively similar to those in BM. Notice that BM manually sets the number of groups to four, but we discover that four is the optimal number. Consequently, by employing model specification 1 and accounting for heteroskedasticity, we find the support for BM's main results.

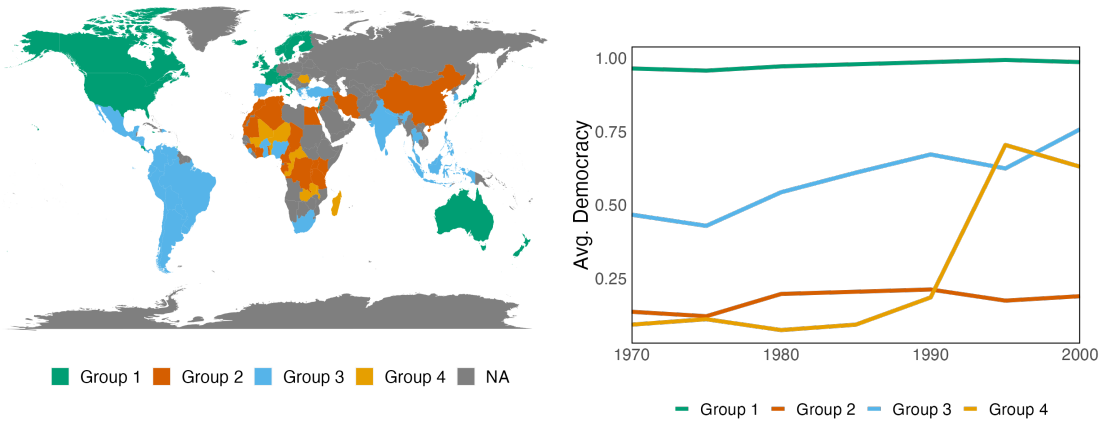Figure G.7: Point Estimation of Group Partitioning and Average Democracy

Table G.7 shows the posterior mean and 90% credible set for each coefficient, with G fixing at the point estimate $\widehat{G}$. Comparing to the pooled OLS, $\hat{\rho}$ and $\hat{\beta}$ once we incorporate the group-specific time patterns. The results are essentially consistent with the conclusion in BM: there is modest persistence and a positive effect of income on democracy, but the cumulative income effect $\beta/(1-\rho) = 0.08$ is quantitatively small.

Table G.7: Coefficient estimates across groups

	Lagged democracy (ρ)		Lagged Income (β)	
	Coef.	Cred. Set	Coef.	Cred. Set
BGFE-he-cstr	0.499	[0.438, 0.558]	0.040	[0.027, 0.053]
Pooled OLS	0.665	[0.616, 0.718]	0.082	[0.065, 0.100]

G.2.2 Network Visualization of Posterior Similarity Matrix

Figure G.8 and G.9 show the full-sample country network graphs, for specification 1 and 2 respectively. Node color reflects the geographic information used to construct the group structure in the prior.

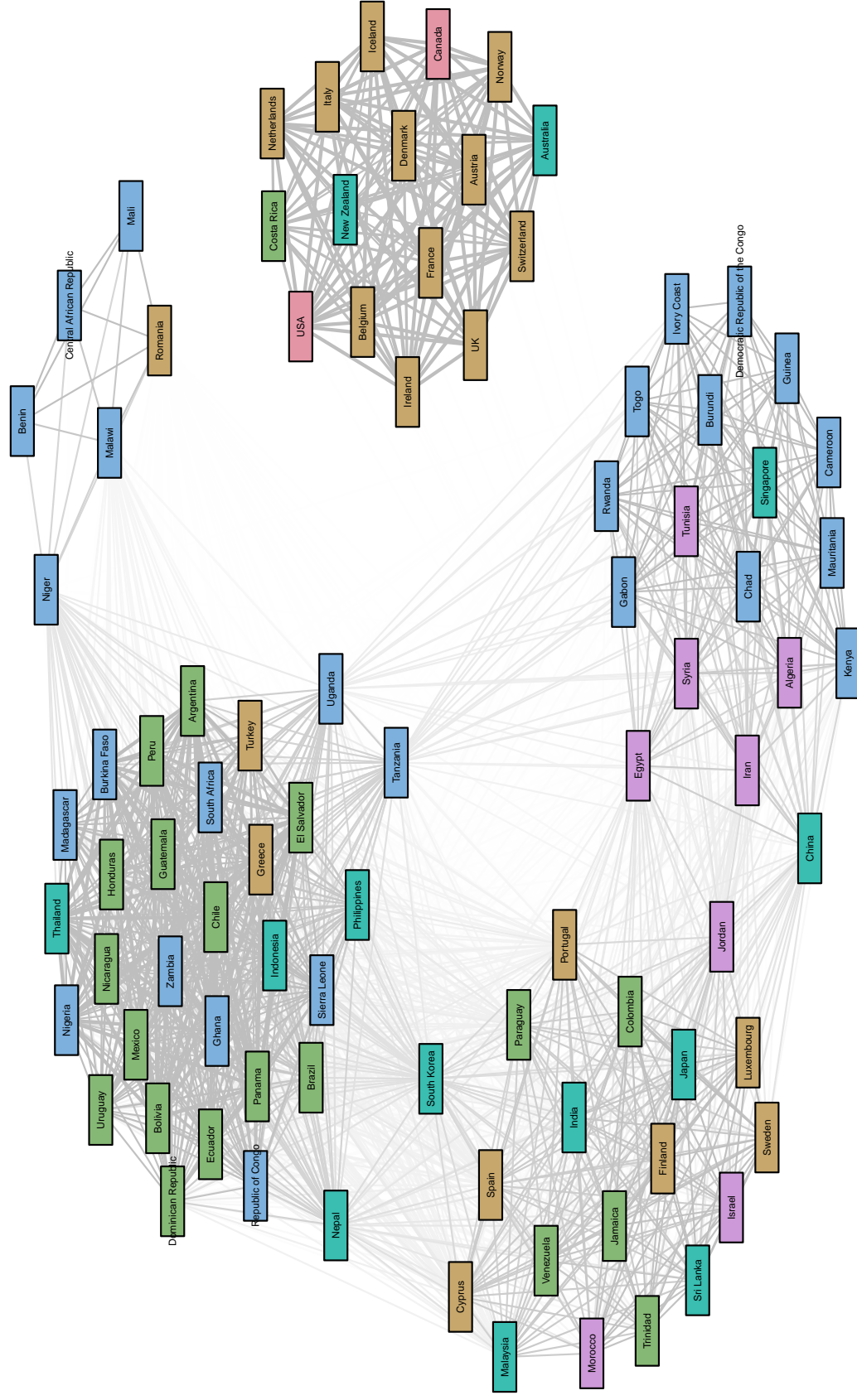Figure G.8: Individual Country Network Graph based on Posterior Similarity Matrix, Specification 1

Figure G.9: Individual Country Network Graph based on Posterior Similarity Matrix, Specification 2