Quasi-Bayesian Inference for Grouped Panels*

Job Market Paper - Latest Version Here

Jiaming Huang

December 30, 2023

Abstract

This paper considers the problem of conducting inference in panel data models with latent group structures. I introduce a quasi-Bayesian framework that combines general classes of loss functions and priors for joint inference on the latent group structures, including group-level parameters and group assignments. Theoretically, I establish consistency of the proposed framework and derive posterior contraction rates for the quasi-Bayesian posterior distribution. Simulation results demonstrate significant improvements in bias and coverage for group-level parameters compared to existing methods, especially when group assignments cannot be precisely estimated. Using the quasi-Bayesian clustering approach, I revisit the heterogeneous income risks of households and identify two previously undetected groups. The first experiences income increases in response to higher unemployment rates, and the second suffers substantial income losses despite being wealthy. These findings provide new insights into the shock amplification channels in heterogeneous agent models.

Keywords: grouped heterogeneity, clustering, quasi-Bayes, income risks

JEL: C11, C14, C33, E24, E32, H24

*Universitat Pompeu Fabra and Barcelona School of Economics: jiaming.huang@upf.edu.

I would like to thank my advisor Geert Mesters for his guidance and support. For helpful comments and discussions, I thank Stéphane Bonhomme, Barbara Rossi, Kirill Evdokimov, Katerina Petrova, Davide Debortoli, Wendun Wang, and the participants of the 3rd International Econometrics PhD Conference, UPF internal Econometrics seminar, and UPF Macroeconomics Lunch seminar. I gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation (FPI Grant PRE2020-092551). All errors are my own. The `Julia` package for implementing the quasi-Bayesian estimator is available on my webpage.

1 Introduction

Modeling the heterogeneous behavior of economic agents has been an active research area in economics; see Mian et al. (2013) and Postel-Vinay and Robin (2002) for prominent examples in macro and microeconomics. For theoretical, interpretability, or tractability reasons, researchers often impose that heterogeneity can be captured by a discrete partitioning of entities, i.e., a group structure. For example, in the study of monetary transmission, households are grouped based on financial constraints (Kaplan et al., 2018), home ownership (Cloyne et al., 2020), or skill levels (Dolado et al., 2021). Alternatively, in labor economics, individuals are classified according to race (Bils, 1985), age (Dustmann et al., 2017), or geographical location (Monte et al., 2018).

While pre-defined groupings can be convenient, economic theory often imposes a *latent* group structure. For instance, at the center of the debate over aggregate shock amplification are the hand-to-mouth households—those with a high marginal propensity to consume (MPC). However, who are the hand-to-mouth households has remained an open question (Aguiar et al., 2020). This has motivated the development of more agnostic approaches for determining group structures, in which agents are placed into groups using data-driven methods (e.g., Bonhomme and Manresa, 2015). Such methods can be used to answer a variety of questions: Is there any heterogeneity in the population? Which entities belong to which groups? How do different groups behave differently? And so on.

To answer these questions, existing approaches typically employ a two-step procedure: Group assignments are estimated using methods such as k-means (Zhang et al., 2019) or penalized estimation (Su et al., 2016). Conditional on the estimated group structure, group-level parameters are estimated using methods such as OLS (Cytrynbaum, 2020) or GMM (Cheng et al., 2019). While the order is logical, estimation errors compound across steps and can thus lead to selection bias and underestimated standard errors, especially when group assignments are difficult to determine (Leeb and Pötscher, 2005).

To circumvent such propagation of errors, I develop a quasi-Bayesian methodology for grouped panels that allows for joint inference on the latent group structure. The framework provides a straightforward way to quantify the uncertainty associated with the latent group structure through posterior sampling (Wade and Ghahramani, 2018; Rigon et al., 2023). The resulting confidence sets, in stark contrast to conventional ones, explicitly account for parameter uncertainty in the estimated group structure. Moreover, compared with conventional Bayesian clustering, the proposed quasi-Bayesian framework is more robust to model misspecification, due to the use of general loss functions instead of the exact likelihood. This is crucial as Bayesian clustering can suffer from biases in the presence of misspecification

(Guha et al., 2021).

The proposed quasi-Bayesian framework comprises three ingredients. First, a criterion or loss function that identifies the group structure in the population. Popular examples considered include the least squares loss (Bonhomme and Manresa, 2015), the GMM criterion (Huang, 2021), and the quantile loss (Zhang et al., 2019). Second, a prior that assigns nonzero probability to the true parameters in the latent group structure. Examples include the finite mixtures prior (e.g., Miller and Harrison, 2018) and the graphical Laplacian prior (e.g., Kim and Gao, 2020). These priors enable consistent estimation while allowing for efficient sampling algorithms. Third, a learning rate parameter that controls the bias and variance of the quasi-posterior distributions. A higher learning rate puts more weight on the loss component, reducing variance but increasing the risk of selection bias. Conversely, a lower learning rate puts more weight on the prior component, enhancing robustness to misspecification but potentially increasing variance. In practice, the learning rate is calibrated using bootstrap methods to improve inference.

Theoretically, this paper establishes the first frequentist guarantees for a general class of quasi-Bayesian clustering models. Specifically, I establish consistency and derive contraction rates for the quasi-Bayesian posterior distribution under a set of high-level conditions, encompassing flexible classes of criterion functions and priors. The general results quantify the tradeoffs between model specification, data requirement, and prior knowledge, which can be used to discipline the primitive conditions in clustering. I then provide primitive conditions for two popular estimators, the M-estimator and the GMM estimator, under which posterior consistency holds, and derive the corresponding posterior contraction rates. Importantly, similar strategies could be applied in future work to examine other loss functions or priors proposed in the Bayesian clustering literature, for which classical contraction results can be difficult to derive (e.g., Duan and Dunson, 2021).

The theoretical results build on two main observations. First, the quasi-posterior distribution is dominated by the loss component under regularity conditions. For instance, the prior must put sufficient mass around the true parameters, leading the posterior draws to concentrate around the *sample* loss minimizer. Second, as the sample loss converges to its population counterpart, the posterior distribution concentrates around the *population* loss minimizer. If the identification condition holds—meaning that the true parameters of interest are the population loss minimizer—this guarantees the concentration of the posterior draws around the true parameters. Consequently, the posterior contraction rate is determined by (i) the identification condition, (ii) the uniform convergence of the sample loss, and (iii) the prior mass condition. This provides an alternative approach, in contrast to the classical posterior convergence result by Ghosal and van der Vaart (2007), for establishing

contraction rates through the lens of empirical risk minimization.

In practice, the quasi-Bayesian framework can be implemented with a computationally efficient blocked Gibbs sampler where we iterate between drawing group assignments and group-level parameters. In particular, each block can easily incorporate efficient MCMC samplers to improve computational performance.[1] For instance, Metropolis-Hastings algorithms facilitate sampling in non-conjugate models (Chernozhukov and Hong, 2003), and Particle Gibbs methods can improve mixing rates (Bouchard-Cote et al., 2017). Further, the number of latent groups is often unknown. The quasi-Bayesian approach accounts for this uncertainty by specifying a prior on the group number, and I show the posterior ratio consistency for the resulting quasi-posterior. This implies that we can simply select the number of groups using the posterior mode, thereby avoiding the need to evaluate intractable integrals as typical in conventional methods (e.g., Kass and Raftery, 1995).

Compared with existing approaches, the advantages of the quasi-Bayesian framework are threefold. First, uncertainty in the estimated group structure can be readily measured in the proposed framework, whereas it remains a challenging task using frequentist methods.[2] Second, the framework enables coherent incorporation of estimation uncertainty in the group structure when conducting inference on the group-level parameters. This allows researchers to construct confidence sets that are free from the selection bias. Finally, the framework avoids specifying the exact likelihood of the data, which is often intractable due to the presence of heteroskedasticity and autocorrelation of unknown form. Moreover, with the learning rate calibration, the quasi-Bayesian approach is more robust to (likelihood) misspecification than conventional Bayesian clustering methods.

I illustrate the advantages of the proposed method in a large-scale simulation study. First, conditional on the correct specification of the group number, the coverage probabilities of quasi-Bayesian confidence sets remain close to the nominal level even with low signal-to-noise ratios and heteroskedasticity (Stock and Watson, 2008). In contrast, conventional confidence sets as constructed in Bonhomme and Manresa (2015) or Su et al. (2016) severely undercover in the presence of misclassification errors. Second, when the data is informative, e.g., when the time series dimension is large, quasi-Bayesian confidence sets perform competitively as the conventional ones in terms of both the coverage and the length of confidence intervals. Moreover, the root mean squared error of the quasi-Bayesian estimator follows closely the frequentist counterpart. Third, the quasi-Bayesian posterior is able to measure the uncertainty in the number of groups, and the posterior mode converges to the true group

[1] See, for example, Robert et al. (2018) for a recent review on scalable MCMC methods.

[2] While a frequentist mixture model accommodates uncertainty in group assignments, it is unable to quantify uncertainty in the number of groups. Further discussion on the motivation for Bayesian clustering can be found in Bishop and Svensen (2012).

number as the sample size increases.

With the quasi-Bayesian clustering approach, I revisit the literature on cyclical income risks using a biennial panel on household income from the PSID from 1999 to 2009. Specifically, I examine the heterogeneity in the elasticity of log earnings to the unemployment rate. While previous studies classify households according to demographic proxy variables (e.g., Busch et al., 2022; Patterson, 2023), I directly estimate the latent group structure based on the heterogeneity in earnings elasticities.

The proposed method reveals three latent groups with significant heterogeneity in income-unemployment elasticities. Notably, I identify one cluster of wealthy households with substantial asset holdings that experience earnings increases in response to higher unemployment rates, akin to the findings of Guvenen et al. (2014). However, I also document a group of rich households that suffer considerable income losses during recessions, *despite not being liquidity constrained*. This discovery challenges the conventional wisdom that poor, constrained households are most vulnerable to economic downturns, thereby casting doubt on shock amplification mechanisms in heterogeneous agent models. Moreover, no single demographic indicator can fully capture the documented group differences, highlighting the importance of *latent* group heterogeneity. Crucially, the revealed group patterns are also missed by popular k-means clustering methods (Bonhomme and Manresa, 2015). In summary, the empirical application underscores the advantages of the proposed method for flexibly detecting heterogeneous cyclical income dynamics, while simultaneously improving estimation and inference on group-level parameters through joint modeling.

The remainder of this paper is organized as follows. I continue this introduction by relating the quasi-Bayesian to the existing literature. Section 2 provides an illustrative example of the setup and contrasts the existing approaches with the proposed method through a minimal simulation study. I then formally presents the quasi-Bayesian framework in Section 3 for which the asymptotic properties are developed in Section 4. Section 5 evaluates the finite sample performance of the quasi-Bayesian approach and Section 6 presents the empirical application. Section 7 concludes. All proofs as well as additional simulation and empirical results are relegated to the Appendix.

1.1 Related Literature

This paper relates to several strands of literature.

First, this paper adds to the extensive literature on recovering latent group heterogeneity in panel data models. Bonhomme and Manresa (2015) apply the k-means algorithm to recover the group structure in linear models and show consistency and asymptotic normality for

the resulting estimator, which is further extended to factor models by Ando and Bai (2017), quantile models by Zhang et al. (2019), nonlinear GMM models by Cheng et al. (2019), and general M-estimation settings by Liu et al. (2020). Similar results have also been established among the class of penalization-based estimators, including those with L1-penalty (Su et al., 2016) and SCAD penalty (Wang et al., 2018), in settings with cross-sectional dependence (Su and Ju, 2018) or cointegration (Huang et al., 2020). However, these methods may suffer from severe size distortion in the presence of non-negligible misclassification errors — a point well appreciated in the clustering literature (Lu and Zhou, 2016). In such cases, e.g., when T is modest or small, clustering estimators are only guaranteed to converge to pseudo-true parameters, which may differ from the true parameters (Pollard, 1981, 1982). The proposed quasi-Bayesian approach generalizes previous methods, akin to Park and Casella (2008)'s extension of Lasso. Specifically, the maximum a posteriori estimate is shown to be equivalent to the frequentist estimate under certain conditions. Moreover, the proposed framework improves upon past techniques in several aspects. To begin with, it accounts for uncertainty in group structure through posterior samples, which remains a challenging task within frequentist framework. Additionally, it produces more robust confidence sets, even when the group structure cannot be precisely estimated.

Second, this paper contributes to the long-standing literature on Bayesian clustering. Within the standard Bayesian framework, the literature has evolved by developing prior classes that induce a group structure, including for example the product partition (PP) prior (Quintana and Iglesias, 2003) and the mixture of finite mixture (MFM) prior (Miller and Harrison, 2018), and by developing more efficient sampling procedures, such as the reversible jump MCMC (Richardson and Green, 1997) and Gibbs sampling (Ishwaran and James, 2001). Given the prior choice, the standard Bayesian approach proceeds by specifying the exact likelihood, which has been applied in various economic applications (Kim and Wang, 2019; Ren et al., 2022; Zhang, 2023). However, the full data distribution may sometimes be intractable or simply unavailable. In light of this, this paper joins recent advances in generalized Bayesian clustering literature where the exact likelihood is replaced by some loss function; see for example Duan and Dunson (2021) using pairwise distances as the loss and Rigon et al. (2023) using within-cluster distances as the loss. Although flexible, specifying only a loss function leads to a misspecified likelihood, under which classical results on the posterior contraction such as Nguyen (2013) no longer hold (Guha et al., 2021).[3] It is then unclear what assumptions are required to provide frequentist guarantees. This paper fills this

[3]Since the pseudo-likelihood can be view as a special class of loss functions, this paper is also remotely related to the literature on the asymptotic properties of Bayesian posteriors under misspecification (e.g., Kleijn and van der Vaart, 2012).

gap by providing the first consistency and posterior contraction results in the quasi-Bayesian clustering framework, for general classes of loss functions and priors.

Third, this paper contributes to the theoretical literature on quasi-Bayesian inference. The pioneering works by Chernozhukov and Hong (2003) and Kim (2002) establish the asymptotic properties of the quasi-Bayesian estimator. More recently, theoretical properties of the quasi-Bayesian framework have been studied in high-dimensional problems (Atchadé, 2017; Petrova, 2019; Yano and Kato, 2020), and nonparametric settings (Kato, 2013; Syring and Martin, 2020). For finite-dimensional models, inference has typically relied on the local asymptotic normality condition (LAN), based on which Bernstein-von Mises theorem can be derived to establish the posterior contraction rate (e.g., Chernozhukov and Hong, 2003). For high-dimensional or infinite-dimensional models, posterior contraction results typically require the uniformly consistent tests (UCT) assumption (e.g., Ghosal and van der Vaart, 2017, Theorem 6.16), under which convergence in weak topology can be derived.[4] Following Miller (2021), the current paper considers a different route, by exploiting the close connection between the quasi-Bayesian approach and empirical risk minimization — a fact that is well established to demonstrate risk performance of the quasi-Bayesian estimator (Jiang and Tanner, 2008; Grunwald and Mehta, 2020; Alquier, 2023).

Fourth, this paper also speaks to the emerging empirical literature exploiting latent group heterogeneity in panel data models. Early contributions typically sort economic entities based on ex ante characteristics (e.g., Guvenen et al., 2014). More recently, clustering approaches have found applications across various fields of economics, such as labor economics (Bonhomme et al., 2019; Gregory et al., 2021; Abbott and Gallipoli, 2022), political economy (Gratton et al., 2021; Acemoglu et al., 2019), and macroeconomics (Chen et al., 2019). The proposed framework serves as an external validation to the group structure identified by these studies. This is particularly useful given that the panel data setups considered often feature short time series but large cross-sections, where misclassification errors are likely.

Finally, this paper contributes to the extensive literature on heterogeneous income risks (Low et al., 2010; Guvenen et al., 2014). Since the influential work by Storesletten et al. (2004), estimating heterogeneous income cyclicalities has become crucial, as the cyclicalities constitute an important set of moments in structural macro models (Auclert, 2019; Patterson, 2023). Moreover, cyclical income patterns are closely related to inequality and social insurance design (Blundell et al., 2016). This paper adds to the literature by employing an agnostic approach to recover the latent group structure in income-unemployment elasticities. Additionally, the revealed group patterns cannot be explained by popular demographic

[4]One notable exception is Syring and Martin (2023) where they prove posterior concentration rates for sub-exponential type loss functions. Our framework includes such losses as a special case.

indicators (e.g., Busch et al., 2022; Figueiredo, 2022; Almgren et al., 2022), underscoring the importance of unobserved heterogeneity in understanding income dynamics.

2 Illustrative example

In this section, I illustrate the quasi-Bayesian framework informally for a canonical linear panel data model with grouped coefficients, and compare its performance to some of the currently available methods.

Consider

$$y_{it} = x_{it}^\top \beta_{\gamma_i} + \epsilon_{it} \ , \qquad t = 1, \dots, T, \quad i = 1, \dots, N \ , \tag{1}$$

where y_{it} is the scalar outcome, x_{it} is a $d \times 1$ vector of covariates, β_{γ_i} is the grouped parameter of interest, and ϵ_{it} is the error term. Specifically, the slope parameter is indexed by the group membership indicator γ_i so that when $\gamma_i = \gamma_j = g$, units i and j share the same coefficients β_g, with G the number of groups and $\gamma_i \in \{1, \dots, G\}$. The above model therefore strikes a balance between the fully pooled model with homogeneous coefficients ($\beta_{\gamma_i} = \beta, \forall i$) and the fully heterogeneous model with distinct unit-specific coefficients β_i. Due to its simplicity and flexibility, model (1) has been widely used in the empirical finance and macro literature (e.g., Chen et al., 2019; Huang et al., 2023).

For now, the group number G is assumed known, and the goal is to recover the group-level coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_G) \in \mathcal{B}^G \subset \mathbb{R}^{d \times G}$ and the group assignment vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_N)^\top \in \Gamma^G \subset [G]^N$. The estimation of G is discussed in the general methodology section 3.

2.1 Existing methods

Conventional methods start by minimizing the following objective function:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta} \in \mathcal{B}^G, \gamma \in \Gamma^G}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}^\top \beta_{\gamma_i})^2 + \lambda \text{Pen}(\boldsymbol{\beta}) \tag{2}$$

where $\text{Pen}(\boldsymbol{\beta})$ is a penalty function that can induce group sparsity, and λ is a tuning parameter. Including the penalty term is not necessary and in the absence thereof (2) can be solved by an iterative two-step procedure: (i) given the group-level parameters $\boldsymbol{\beta}$ assign units to groups so that the least squares objective function is minimized and (ii) given the group assignment $\boldsymbol{\gamma}$ estimate group-level parameters using least squares.

Existing works typically proceed as if the assignment vector $\boldsymbol{\gamma}$ is perfectly recovered by

$\hat{\boldsymbol{\gamma}}$, and derive the asymptotic distribution of group-level parameters $\hat{\beta}_g$ given $\hat{\boldsymbol{\gamma}}$. Denoting the true grouped parameter by β_g^0 , the estimation errors of $\hat{\beta}_g$ can then be decomposed as

$$\hat{\beta}_g - \beta_g^0 = \left(\sum_{i,t} \mathbf{1}\{\hat{\gamma}_i = g\} x_{it} x_{it}^\top \right)^{-1} \left(\sum_{i,t} \mathbf{1}\{\hat{\gamma}_i = g\} \left[x_{it} x_{it}^\top (\beta_{\gamma_i^0}^0 - \beta_g^0) + x_{it} \epsilon_{it} \right] \right) . \tag{3}$$

Under the assumption that $\hat{\boldsymbol{\gamma}}$ converges to $\boldsymbol{\gamma}^0$ sufficiently fast, it implies that (e.g. Bonhomme and Manresa, 2015, Corollary S2)

$$\sqrt{N_g T}(\hat{\beta}_g - \beta_g^0) \xrightarrow{d} N(0, V_g) + o_p(1), \quad g = 1, \ldots, G \tag{4}$$

where V_g is the standard sandwich-form asymptotic variance.

This two-step procedure has some limitations, due to its inherent difficulty of conducting inference on the group assignment $\hat{\boldsymbol{\gamma}}$.[5] The first limitation is its inability to evaluate the validity of pre-defined grouping criteria, thereby restricting the applicability of the clustering approach. To address this, model-based clustering methods such as finite mixture models and Bayesian clustering have been proposed. However, these approaches rely on a *correctly specified* generative model for the data, making estimates sensitive to model misspecification.

A second limitation concerns the potential bias and under-coverage for the group-level parameters $\hat{\boldsymbol{\beta}}$. As decomposition (3) indicates, the estimated cluster does not center at the true parameter β_g^0, unless the group assignments are perfectly recovered $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^0$. Additionally, inference based on (4) is prone to the selection bias (Tibshirani et al., 2016), due to the dependence of $\hat{\gamma}_i$ on $\{x_{it}\epsilon_{it}\}_{t=1}^T$. Correction of such bias turns out to be challenging beyond the homoskedastic setup (Gao et al., 2022).

2.2 Quasi-Bayesian clustering

To overcome these limitations, I introduce a quasi-Bayesian framework that allows for joint inference on both the group structure and the group-level parameters while preserving robustness against model misspecification.

Specifically, the quasi-Bayesian framework combines the loss function (2) with a prior $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ on the group structure through a learning rate ψ. The resulting quasi-posterior

[5] A notable exception is Dzemski and Okui (2021) where the authors construct confidence sets for group membership by inverting unit-specific tests. It would be interesting to see how the resulting confidence sets can be used to improve inference for group-level parameters.

density $\pi_{NT}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is defined as

$$\pi_{NT}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\exp\left[-\psi \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}^{\top} \beta_{\gamma_i})^2\right] \pi(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\int \exp\left[-\psi \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}^{\top} \beta_{\gamma_i})^2\right] \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \, \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\gamma}} \,. \tag{5}$$

The quasi-posterior density replaces the likelihood function in the standard Bayesian posterior with an exponential loss. This substitution offers two key advantages. First, the loss function avoids the need to fully specify the data distribution, and researchers can concentrate on the parameters of interest. Second, the learning rate ψ enhances model robustness, by trading off between prior and data evidence. For instance, a smaller ψ puts more weight on the prior, thereby improving robustness to model misspecification.

Next, I illustrate the framework with concrete examples of the prior and the learning rate. A natural choice is a finite mixture prior on the group structure $\boldsymbol{\gamma}$, and an independent normal prior on the group-level parameters $\boldsymbol{\beta}$ (e.g., Diebolt and Robert, 1994), such that

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}) \quad &= \pi(\boldsymbol{\gamma}|\eta)\pi(\eta) \textstyle\prod_{g=1}^{G} \pi(\beta_g) \\ \boldsymbol{\eta} = (\eta_1, \ldots, \eta_G) \quad &\sim \text{Dirichlet}(\alpha_{\text{dir}}), \qquad \alpha_{\text{dir}} \geq 1 \\ \gamma_i|\boldsymbol{\eta} \quad &\sim \text{Categorical}(\eta_1, \ldots, \eta_G), \quad i = 1, \ldots, N \\ \beta_g \quad &\sim N(\mu, \Sigma), \ \Sigma \text{ p.d.} \qquad g = 1, \ldots, G \end{aligned} \,. \tag{6}$$

With the above prior, the quasi-posterior distribution (5) can be sampled efficiently using a blocked Gibbs sampler presented in Algorithm 0. Specifically, the algorithm iterates between two blocks: (i) given the group-level parameters $\boldsymbol{\beta}$, sample group assignments by the Pólya urn scheme (Pitman, 1996) and (ii) given the group assignment $\boldsymbol{\gamma}$, sample group-level parameters from a multivariate normal distribution.

Notice that Algorithm 0 closely resembles the iterative procedure discussed in the previous section. Consider first the group assignment step. The crucial difference is that the group assignments are updated *probabilistically* rather than deterministically. To see why this is important, denote the set of units in group g as $\mathcal{C}_g$, and $\mathcal{C}_{g,-i}$ indicates unit i is removed from $\mathcal{C}_g$. The posterior odds ratio is given by

$$\frac{\pi_{NT}(\gamma_i = j|\boldsymbol{\beta})}{\pi_{NT}(\gamma_i = k|\boldsymbol{\beta})} = \exp\left[-\psi \sum_{t} \left((y_{it} - x_{it}^{\top}\beta_j)^2 - (y_{it} - x_{it}^{\top}\beta_k)^2\right) + \ln \frac{|\mathcal{C}_{j,-i}| + \alpha_{\text{dir}}}{|\mathcal{C}_{k,-i}| + \alpha_{\text{dir}}}\right] , \tag{7}$$

which is of order $\exp\left[O(1) - \psi O(T)\right]$ when the group sizes are of similar scales. In the limiting case where $T \to \infty$, the posterior odds ratio is dominated by the excess loss and thus the quasi-Bayesian maximum-a-posteriori (MAP) estimator reduces to the k-means

estimator. In finite sample, however, (7) is non-zero and can be large when the sample size is small or when the sample loss is flat. This allows the quasi-Bayesian estimator to explore alternative model spaces (group assignments) and thereby escape local optima. Further, the estimation of grouped parameters (9) takes into account the sampling uncertainty of β_g. As a result, the posterior draws of β_g automatically incorporate the uncertainty in both model specification and parameter estimation. By averaging grouped parameters across different group assignments, the posterior distribution may perform better than estimates conditional on a particular assignment.

Algorithm 0 Quasi-Bayes Clustering: Illustrative example

The mth MCMC iteration computes:

1: **Sampling Assignment**. Given $\boldsymbol{\beta}^{(m-1)}$, draw for $i = 1, \ldots, N$

$$\gamma_i^{(m)} \sim \pi_{NT}(\gamma_i = g|\boldsymbol{\beta}^{(m-1)}) \propto \exp\big[-\psi \sum_t (y_{it} - x_{it}^\top \beta_g^{(m-1)})^2\big](|\mathcal{C}_{g,-i}^{(m-1)}| + \alpha_{\text{dir}}) \ . \tag{8}$$

2: **Sampling Grouped Parameters**. Given $\gamma^{(m)}$, draw for $g = 1, \ldots, G$

$$\beta_g^{(m)} \sim N(\tilde{\mu}, \tilde{\Sigma}), \quad \tilde{\mu} = \tilde{\Sigma}^{-1}\left[2\psi Y_g X_g + \mu^\top \Sigma^{-1}\right], \quad \tilde{\Sigma}^{-1} = 2\psi X_g^\top X_g + \Sigma^{-1} \tag{9}$$

where $Y_g = (y_i)_{i \in \mathcal{C}_g^{(m)}}$ and $X_g = (x_i)_{i \in \mathcal{C}_g^{(m)}}$ are the stacked data assigned to group g.

A well-known issue with the proposed framework is that inference based on the quasi-posterior distribution generally leads to incorrect coverage of the parameters of interest. This is because the loss component can be viewed as a misspecified likelihood, which violates the generalized information equality and yields an incorrect asymptotic variance matrix (Chernozhukov and Hong, 2003; Müller, 2013).

Nevertheless, the learning rate ψ can be calibrated to improve the coverage. Intuitively, the posterior odds (7) collapse to the prior odds when $\psi \to 0$, reflecting total disregard for the observed data. The resulting posterior distribution is as dispersed as the prior. In contrast, the quasi-posterior reduces to a dirac mass at the k-means estimate when $\psi \to \infty$. The learning rate ψ therefore controls the dispersion of the posterior distribution, and thereby the coverage of confidence sets. In practice, this is achieved by updating the learning rate in the spirit of Syring and Martin (2019). Specifically, given an initial guess $\psi^{(0)}$, researchers update the learning rate based on the gap between the desired coverage level and the empirical coverage computed using bootstrap. When the empirical coverage exceeds the desired level, the learning rate is increased—reflecting a higher weight on the loss component—and vice

versa. The algorithm terminates when the empirical coverage is within a small tolerance of the desired level. Further details are given in Section 3.3.

It is interesting to contrast the above approach with existing bootstrap proposals to improve inference on $\hat{\beta}_g$. For instance, Bonhomme and Manresa (2015) propose to estimate the asymptotic variance V_g in (4) by cross-sectional bootstrap. However, this alternative is unlikely to improve inference for two reasons. First, decomposition (3) indicates that the estimator $\hat{\beta}_g$ is biased whereas the bootstrap proposal focuses only on the variance estimation. Second, parameter estimation in each bootstrapped sample is subject to selection bias, potentially leading to overly small asymptotic variance estimates (Neufeld et al., 2022). Therefore, it is unclear how these bootstrapped estimates improve inference.

Finally, the above discussions extend to more general loss functions and priors, outlined in Section 3. Computationally, Algorithm 0 can be easily adapted by augmenting (8) and (9) with additional steps to update the parameters. For example, when the prior is non-conjugate, researchers can sample β_g from the full conditional posterior with a Metropolis-Hastings-within-Gibbs algorithm. Alternatively, it is well known that the sequential allocation of γ_i in (8) may lead to slow mixing. More advanced partition samplers such as the Particle-Gibbs Split-Merge sampler may be applied (Bouchard-Cote et al., 2017).

2.3 Small simulation study

I conclude this section with a small scale simulation study where I compare the performance of the quasi-Bayesian clustering approach with the k-means estimator (Lin and Ng, 2012). Specifically, I simulate data from (1) where $x_{i,t} \overset{\text{i.i.d.}}{\sim} N(0, I_2)$ and I_2 is a 2×2 identity matrix. There are three latent groups with group-level coefficients $\beta_1 = (0.4, 1.6)^\top, \beta_2 = (1.0, 1.0)^\top$ and $\beta_3 = (1.6, 0.4)$. I consider homoskedastic errors $\epsilon_{it} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ where the error variance σ^2 varies across designs to reflect different signal-to-noise ratios. Finally, I consider different sample sizes with ($N = 100$, $T = 20$) and ($N = 100$, $T = 10$). The quasi-Bayesian estimator uses the finite-mixture and normal prior (6), with hyperparameter $\alpha_{\text{dir}} = 1.0$ and uninformative Gaussian prior $\beta_g \overset{\text{i.i.d.}}{\sim} N(0, 100 I_2)$. The learning rate ψ is estimated using the bootstrap calibration approach described in Section 3.3. Throughout, I assume that the true number of groups is known. The simulation setup resembles the design of Su et al. (2016).

Besides the estimation and inference of the group-level parameters, it is also of interest to study the performance for the estimation of average effects. In particular, the population average effects of model (1) is given by $\text{AE} = \mathbb{E}[\beta_{\gamma_i}]$, which can be estimated by $\widehat{\text{AE}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\beta}_{\hat{\gamma}_i}$. For the quasi-Bayesian approach, the posterior average effects in the spirit of Bonhomme and Weidner (2022) can be obtained by the MAP estimator $(\hat{\boldsymbol{\beta}}^{\text{MAP}}, \hat{\boldsymbol{\gamma}}^{\text{MAP}})$ that

maximizes the quasi-posterior (5) among the MCMC draws. Further, the quasi-Bayesian approach also facilitates easy construction of the confidence sets for the average effects using posterior quantiles.

TABLE 1: PERFORMANCE METRICS: ILLUSTRATIVE EXAMPLE

Sample Size	**Metrics**	$\sigma^2 = 1.0$		$\sigma^2 = 2.0$	
		QB	KM	QB	KM
$N = 100,\ T = 10$	AC	79.96	82.89	54.20	58.65
	BR	36.75	20.68	51.18	17.71
	RMSE	30.19	21.90	80.08	90.25
	Coverage	96.56	74.39	97.44	34.33
	RMSE (AE)	3.37	3.46	7.15	7.43
	Coverage (AE)	98.00	-	97.50	-
$N = 100,\ T = 20$	AC	94.53	94.64	64.56	71.61
	BR	24.47	19.07	42.26	16.68
	RMSE	16.26	10.80	53.82	48.94
	Coverage	96.06	89.00	98.06	52.50
	RMSE (AE)	2.38	2.39	4.91	4.98
	Coverage (AE)	96.33	-	97.50	-

Note: This table reports the classification accuracy (AC), the confidence bands ratios with respect to the unit-level least-squares estimator (BR), the RMSE of the group-level parameters, the RMSE of the average effects, and the coverage probabilities for the quasi-Bayes estimator and the k-means estimator respectively. To evaluate accuracy and bands ratio, maximum-a-posteriori (MAP) estimator $\hat{\gamma}^{\text{MAP}}$ is used; RMSE is computed using the posterior mode of the quasi-posterior, and coverage rates are calculated as the 2.5% and 97.5% quantiles of the posterior draws. The coverage rates for k-means estimator is calculated using confidence intervals with cluster-robust standard errors as in Stock and Watson (2008). AC, BR and Coverage are in percentage terms, and RMSE are multiplied by 100.

Table 1 reports the results. Several patterns stand out. First, when the sample size is large and the data is informative, the k-means estimator accurately recovers the latent group assignment. However, the classification accuracy severely deteriorates when the time series dimension is limited ($T = 10$) or when the data is noisy ($\sigma^2 = 2$). Consistent with the analysis of (7), the classification accuracy of the quasi-Bayesian MAP estimator is strictly lower than the k-means estimator, even more so when the data is uninformative. This is because in

those scenarios the quasi-Bayesian approach puts more weight on the uninformative prior component.

Second, the average lengths of the confidence intervals of the quasi-Bayesian estimator are inflated when compared to the k-means estimator, but they remain competitive when compared with unit-level confidence intervals. Again, this results from the fact that the quasi-Bayesian estimator explores the model space. Interestingly, the RMSE of the posterior mode outperforms the k-means counterparts in some cases, reflecting the merits of averaging different group structures when estimation uncertainty is high.

Third, it is striking that the coverage probabilities of the k-means confidence sets fall below the nominal 95% level in all four designs considered, with coverage rates as low as 34.3%. In stark contrast, the quasi-Bayesian confidence sets remain above the nominal level throughout. Fourth, the superior performance of the quasi-Bayesian estimator also holds when estimating the average effects: the quasi-Bayesian approach not only delivers lower RMSE than k-means clustering, but also provides confidence intervals for the posterior average effects with good empirical coverage.

3 General methodology

This section introduces a general quasi-Bayesian framework for clustering in panel data models. Section 3.1 presents the model and examples of loss functions and priors that may be included. Section 3.2 provides the implementation details including the selection of the number of groups, and Section 3.3 discusses the learning rate selection.

3.1 Modeling framework

We observe $p \times 1$ vectors $w_{i,t}$ for individual units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. Each unit i is associated with a $d \times 1$ vector of parameters β_i. We assume that there is a partition of the data into G groups that are indexed by the group membership indicator $\gamma_i \in \{1, \ldots, G\}$. The parameters β_i are constant within each group but vary across groups, i.e., $\beta_i = \beta_{\gamma_i}$ and $\beta_i \neq \beta_j$ if $\gamma_i \neq \gamma_j$. The goal is to recover both the group membership $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^\top$ and the group-level parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_G)$, from which researchers may calculate the object of interest $\zeta = f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ as a functional of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Examples 1-2 are two popular choices.[6] For notational brevity, we collect the set of parameters by $\theta = (\boldsymbol{\beta}, \boldsymbol{\gamma})$.

We consider a loss function $l_{iT}(\beta_{\gamma_i}) = l(w_i, \beta_{\gamma_i})$ where $w_i = (w_{i,1}, \ldots, w_{i,T})^\top$ is the data

[6]Since the group labels are not identifiable, we require $f(\cdot, \cdot)$ to be invariant under relabeling of the groups.

matrix for unit i. The true parameter θ^0 is defined as the unique minimizer of the population loss function

$$\theta^0 = \arg\min_{\theta} L_N(\theta), \quad L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[l_{iT}(\beta_{\gamma_i})\right] \ . \tag{10}$$

Motivated by the population problem (10), a popular frequentist extremum estimator is

$$\hat{\theta} = \arg\min_{\theta} L_{NT}(\theta), \quad L_{NT}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l_{iT}(\beta_{\gamma_i}) \ . \tag{11}$$

Examples 3-5 illustrate the specification of loss functions for popular estimators.

To model the latent group structure, we consider a class of group-sparsity inducing prior $\pi(\theta)$ for θ. Examples 6 and 7 are some popular choices that are compatible with the current framework.

The quasi-Bayesian clustering framework combines the criterion function in (11) and the prior $\pi(\theta)$, through a learning rate $\psi > 0$, leading to the quasi-posterior distribution

$$\Pi_{NT}(A) = \frac{\int_A \exp\left[-NT\psi L_{NT}(\theta)\right]\pi(\theta)\mathrm{d}\theta}{\int_\Theta \exp\left[-NT\psi L_{NT}(\theta)\right]\pi(\theta)\mathrm{d}\theta} \ . \tag{12}$$

Several comments are in order. First, the loss function formulation (10) encompasses a large class of panel data models in the literature, among them popular examples are linear models with exogenous covariates (Bonhomme and Manresa, 2015; Cytrynbaum, 2020), moment condition models (Fernández-Val and Lee, 2013; Cheng et al., 2019), binary choice models (Su et al., 2016; Liu et al., 2020), and censored models (Hahn and Kuersteiner, 2011; Wang and Su, 2021).

Second, in principle $L_N(\theta)$ can be any arbitrary loss function which identifies the parameters of interest in population. When $NT\psi L_{NT}(\theta)$ is the negative log-likelihood of the data, (12) reduces to standard Bayesian posterior for clustering (e.g., Ren et al., 2022; Smith, 2022; Zhang, 2023). When $NT\psi L_{NT}(\theta)$ disagrees with the exact likelihood, as it is often the case, the quasi-posterior distribution still provides a coherent way of updating the prior beliefs given the data (Bissiri et al., 2016). In particular, the quasi-posterior distribution Π_{NT} solves the following minimization problem (e.g., Zhang, 2006, Proposition 5.1)

$$\inf_{\tilde{\Pi}} \left\{ \int NTL_{NT}(\theta)\tilde{\Pi}(\mathrm{d}\theta) + \psi^{-1}\mathrm{KL}(\tilde{\Pi}|\Pi) \right\} \tag{13}$$

where $\mathrm{KL}(\tilde{\Pi}|\Pi)$ is the KullbackLeibler divergence between $\tilde{\Pi}$ and the prior Π, and the infimum is taken over all $\tilde{\Pi}$ that are absolutely continuous with respect to Π. Therefore,

the quasi-Bayesian posterior is the solution of the (integrated) empirical risk minimization, while penalizing the deviation from the prior.

Overall, the proposed framework resembles the recent quasi-Bayesian clustering approaches (e.g., Duan and Dunson, 2021; Rigon et al., 2023; Natarajan et al., 2023). However, the choice of loss functions often lacks theoretical justification in existing quasi-Bayesian clustering, and asymptotic properties of the posterior distributions are unknown. In contrast, I consider loss functions of which the parameters of interest are the extremum estimators, and study large sample asymptotics of the posterior distributions, building on the recent literature on panel data models with latent group structure (Bonhomme and Manresa, 2015; Liu et al., 2020; Cytrynbaum, 2020).

Examples: parameters, loss functions and priors

Example 1 (Ordered group-level parameters (Frühwirth-Schnatter, 2006))**.** A natural object of interest is the collection of group-level parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_G)$. However, matrix $\boldsymbol{\beta}$ is not invariant to relabeling of groups. To ensure identification, we may impose an inequality constraint such that

$$\zeta = (\beta_{\sigma(1)}, \ldots, \beta_{\sigma(G)}) \tag{14}$$

where $\sigma\colon [G] \to [G]$ is a permutation such that $\beta_{\sigma(1)} \leq \ldots \leq \beta_{\sigma(G)}$. Further, we may consider various statistics based on the ordered group-level parameters such as the maximal effects $\zeta = \max\{\beta_1, \ldots, \beta_G\}$. △

Example 2 (Average effects (Bonhomme and Weidner, 2022))**.** When the aggregate effects are of interest, we can calculate the average effects across different groups by $\zeta = \frac{1}{N}\sum_{i=1}^{N} \beta_{\gamma_i}$. △

Example 3 (Linear model with grouped heteroskedasticity (Aguilar and Boot, 2022))**.** We can modify the linear model (2) to accommodate grouped variance structure

$$l_{iT}(w_i, \beta_{\gamma_i}) = \frac{1}{T}\sum_{t=1}^{T}\left[\sigma_{\gamma_i} + \sigma_{\gamma_i}^{-1}(y_{it} - x_{it}^{\top}\tilde{\beta}_{\gamma_i})^2\right] \tag{15}$$

where $\beta_{\gamma_i} = (\tilde{\beta}_{\gamma_i}^{\top}, \sigma_{\gamma_i})^{\top}$, and $\sigma_{\gamma_i} \in \mathbb{R}_+$ is the square root of error variance for unit i in group γ_i. It is well known that when x_{it} includes time dummy variables, the exact likelihood of the data is subject to singularity problem (Hamilton, 1991). The loss function (15) however remains well defined and can be used to estimate the group-level parameters. △

Example 4 (Moment model (Hansen, 1982))**.** Consider a model with moment conditions $\mathbb{E}[m(w_{i,t};\beta^0_{\gamma_i^0})]=0$. We can set

$$l_{iT}(w_i,\beta_{\gamma_i})=\left(\frac{1}{T}\sum_{t=1}^{T}m(w_{i,t};\beta_{\gamma_i})\right)^{\top}\hat{\Omega}_i\left(\frac{1}{T}\sum_{t=1}^{T}m(w_{i,t};\beta_{\gamma_i})\right) \tag{16}$$

where $\hat{\Omega}_i$ is a weighting matrix. For example, we may set $\hat{\Omega}_i=\hat{V}(\beta_{\gamma_i})^{-1}$ where $\hat{V}_{i,h}$ is the Newey and West (1987) variance estimate of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}m(w_{i,t};\beta_{\gamma_i})$.[7] The loss (16) has wide applications in economics and finance, including for example models with instrumental variables, and rational expectations models. △

Example 5 (Binary choice model (Arellano and Carrasco, 2003))**.** Let $F(\cdot)$ be the conditional CDF of standard normal (logistic) distribution, the probit (logit) model is specified via

$$-l_{iT}(w_i,\beta_{\gamma_i})=\frac{1}{T}\sum_{t=1}^{T}\left[y_{it}\ln F(y_{it}-x_{it}^{\top}\beta_{\gamma_i})+(1-y_{it})\ln(1-F(y_{it}-x_{it}^{\top}\beta_{\gamma_i}))\right]\ . \tag{17}$$

This class of models is widely used in empirical applications where we observe binary outcome y_{it}, including for example models of labor supply decisions and portfolio choices. △

Example 6 (Diffuse Partition Prior (Gao et al., 2020))**.** When researchers have little prior knowledge of the latent group structure, a natural choice is a diffuse prior

$$\begin{aligned} G\sim&\ \pi(G)\propto\exp\left[-CN\log G\right]\\ \boldsymbol{\gamma}\sim&\ \pi(\boldsymbol{\gamma})=\tfrac{1}{|\Gamma_G|}\end{aligned} \tag{18}$$

where $|\Gamma_G|$ is the set of group assignments that gives G groups. Notice that Γ_G can also incorporate parameter constraints. For example, Gao et al. (2020) require $\boldsymbol{\gamma}$ to be drawn uniformly from assignments such that group-level covariates are of full column rank. Alternatively, one may require the partitions to satisfy minimal group size requirement: $\Gamma_G=\{\gamma\colon|\mathcal{C}_g|\geq\underline{n}\ \forall g\}$. △

Example 7 (Graph Structured Sparisty Priors (Kim and Gao, 2020))**.** The group structure $\theta=(\boldsymbol{\beta},\boldsymbol{\gamma})$ can be equivalently re-parameterized as a graph (V,E) with $V=[N]$ being the

[7] Notice that the GMM criterion induced by (16) is different from the fully pooled criterion, i.e., $L_{NT}(\theta)=\left(\sum_i\sum_t m(w_{it};\beta_{\gamma_i})\right)^{\top}\hat{\Omega}\left(\sum_i\sum_t m(w_{it};\beta_{\gamma_i})\right)$. Although (16) generally leads to inefficient estimates (Su et al., 2016; Huang, 2021), it is necessary to identify the latent group structure. For example, in the context of structural break estimation, Hall et al. (2012) shows that the pooled GMM criterion is not uniquely minimized at the true break points.

nodes and $E \subset \{(i,j)\colon 1 \leq i < j \leq N\}$ the edges, such that

$$\begin{aligned} \pi(\theta|\lambda) &\propto \prod_{(i,j)\in E} \exp\left[-\tfrac{\|\beta_i-\beta_j\|^2}{2(\nu_0\lambda_{ij}+\nu_1(1-\lambda_{ij}))}\right] \\ \pi(\lambda|\eta) &\propto \prod_{(i,j)\in E} \eta^{\lambda_{ij}}(1-\eta)^{1-\lambda_{ij}} \\ \eta &\sim \text{Beta}(a,b) \end{aligned} \tag{19}$$

where ν_0 is a small scalar and ν_1 a large scalar. When $\lambda_{ij} = 1$, the above prior encourages a smaller $\|\beta_i - \beta_j\|$, similiar to methods that penalize pairwise parameter differences (Yang et al., 2019; Mehrabani, 2023). Moreover, the above prior also admits an alternative parameterization where each unit coefficient β_i shrinks towards unknown cluster center μ_j

$$\pi(\theta|\lambda) \propto \prod_{i=1}^{N}\prod_{j=1}^{G} \exp\left[-\frac{\|\beta_i-\mu_j\|}{2(\nu_0\lambda_{ij}+\nu_1(1-\lambda_{ij}))}\right] \tag{20}$$

which corresponds to the classifier-Lasso estimator Su et al. (2016). The alternative formulation explicitly takes the number of G as input, and thus requires an additional MH step to update the number of groups. △

3.2 Implementation details

We are now ready to describe the implementation details of the quasi-Bayesian clustering algorithm. Our goal is to efficiently sample from the quasi-posterior (12), which gives:

$$\pi_{NT}(\boldsymbol{\beta},\boldsymbol{\gamma}) \propto \exp[-T\psi\textstyle\sum_i l_{iT}(\beta_{\gamma_i})]\pi(\boldsymbol{\beta},\boldsymbol{\gamma}) \ . \tag{21}$$

As is suggested in the previous section, it is straightforward to sample from $\pi_{NT}(\boldsymbol{\beta},\boldsymbol{\gamma})$ with a blocked Gibbs sampler, described as follows:

Algorithm 1 Generic Quasi-Bayes clustering: Blocked Gibbs

Data $\{w_{it}\}$, initial parameters $\boldsymbol{\gamma}^0$ and $\boldsymbol{\beta}^{(0)}$, number of MCMC draws M.

1: **for** `m=1,...,M` **do**
2: **Sample group assignments.** $\boldsymbol{\gamma}^{(m)} \sim \pi_{NT}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(m-1)})$.
3: **Sample group-level parameters.** $\boldsymbol{\beta}^{(m)} \sim \pi_{NT}(\boldsymbol{\beta}|\boldsymbol{\gamma}^{(m)})$.
4: **end for**

Below I first present a canonical full conditional Gibbs sampler for updating the group assignment $\boldsymbol{\gamma}$ with mixture-of-finite-mixture (MFM) prior (6). Overall, Algorithm 2 is a

variant of the "Algorithm 8" of Neal (2000), extended to the setup with mixture of finite mixture priors (Miller and Harrison, 2018) and (generally non-conjugate) loss functions.

The conditional posterior density contains two components. The first component evaluates the exponential loss function, which measures how well parameter β fits the data. The second component comes from the MFM prior on partition, similar to the classical Chinese Restaurant Process (CRP). When the number of groups G is fixed, the prior component depends only on the group size $\mathcal{C}_{g,-i}$ and the hyper-parameter α for Dirichlet process, which assigns higher probability to the assignment to a large group. When the number of groups is estimated, the MFM prior imposes a penalty term $V_{N,G+1}/V_{N,G}$, which shrinks to zero as G increases and thus is able to control the number of groups (Miller and Harrison, 2018). Furthermore, compared with the illustrative example, (8), here the algorithm introduces H auxiliary variables as suggested by Neal (2000). Intuitively, whenever a new group assignment is drawn, the loss component is evaluated at a new group-level parameter β drawn from its prior distribution. In case of diffuse prior, the probability of generating a new group is low, leading to slow convergence of the MCMC chain. By drawing H additional β, the algorithm increases the probability of generating new assignments and thus facilitates the sampler to explore the model space.

Another noticeable feature is that the number of groups G is generated by a deterministic mapping from the sampled assignment vector $\boldsymbol{\gamma}$. This is one of the main benefit of the MFM prior, as researchers are free from manually searching over the varying dimensional parameter space, i.e., the number of groups G. In contrast, classical MFM prior involves an additional Metropolis-Hastings (MH) block to determine the number of groups (Richardson and Green, 1997), which is computationally intensive.

Algorithm 2 Mixture-of-finite-mixture: Pólya urn scheme

Data $\{w_{it}\}$, prior $\pi(\beta)$, penalty $V_{N,G} = \sum_{K=1}^{\infty} \frac{K_{(G)}}{(\alpha K)^{(N)}} \pi(G|\lambda)$.

1: **for** `i=1,...,N` **do**

2: Let $G = |\mathcal{C}_{-i}|$ be the number of groups when unit i is excluded. Sample γ_i from

$$\pi_{NT}\left(\gamma_i \middle| \gamma_{-i}, \boldsymbol{\beta}\right) \propto \begin{cases} \exp\left[-T\psi l_{iT}(\beta_g)\right]\left(|\mathcal{C}_{g,-i}| + \alpha\right) & g \in \{1, \ldots, G\} \\ \exp\left[-T\psi l_{iT}(\beta_g)\right] \frac{V_{N,G+1}}{V_{N,G}} \frac{\alpha}{H} & g \in \{G+1, \ldots, G+H\} \end{cases}. \tag{22}$$

Whenever $g > G$, a new β_g is sampled from its prior.

3: **end for**

Note: Here $x^{(b)} = x \cdots (x+b-1)$, $x^{(0)} = 1$ and $x_{(b)} = x \cdots (x-b+1)$, $x_{(0)} = 1$. If $x < b$, set $x_{(b)} = 0$.

Next we would like to sample from the conditional posteriors for grouped parameters β_g.

Given the general class of loss functions, it is typically difficult to find conjugate priors as in Section 2.2. Instead, we can always supplement the Gibbs sampler with a Metropolis step (Chernozhukov and Hong, 2003). Below I give an example for efficient parameter updates. For more general cases see Robert et al. (2018) and the reference therein.

Algorithm 3 updates group-level parameters with the Robust Adaptive Metropolis (RAM) sampler (Vihola, 2012). Compared with classical Metropolis-Hastings algorithm, the RAM sampler is both more robust and more efficient. For one thing, it avoids using the covariance matrix among posterior samples, which may not exist or positive definite in practice. For another, the algorithm remains adaptive since the scaling matrix $S^{(m)}$ is data dependent.

Algorithm 3 Sampling β_g: Robust Adaptive Metropolis (RAM)

Input: Proposal density $q(\cdot)$ for β, s_1 lower diagonal matrix with positive diagonal elements, $\{\eta_n\}$ step size sequence converging to 0, α^* target acceptance rate.

1: Draw proposal $\tilde{\beta} = \beta^{(m)} + S^{(m-1)}U$, where $U \sim q$.

2: Accept proposal draw with acceptance ratio $\alpha^{(m)} = \min\{1, \pi(\tilde{\beta})/\pi(\beta)\}$.

3: Update the lower-diagonal matrix $S^{(m)}$ such that

$$S^{(m)}S^{(m)\top} = S^{(m-1)}\left(I + \eta_n(\alpha_n - \alpha^*)\frac{U^{(m)}U^{(m)\top}}{\|U^{(m)}\|^2}\right)S^{(m-1)\top} \tag{23}$$

Selecting the number of groups G

Several methods have been proposed for selecting the number of groups (G) within a Bayesian clustering framework. Popular examples include the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), variants of the Bayesian information criterion (BIC) (Drton and Plummer, 2017), and the Bayes factor (Kass and Raftery, 1995). However, establishing consistency of these criteria is generally difficult in Bayesian mixture models, especially in the presence of model misspecification (Fúquene et al., 2019).

Given these challenges, we opt for a simple approach of choosing the group number based on the posterior mode averaged across MCMC iterations, for which I derive posterior ratio consistency similar to Narisetty et al. (2019). The posterior mode strikes a reasonable balance between uncertainty quantification and computational tractability for our application, in contrast with marginal likelihood-based information criteria. Nonetheless, estimating the number of groups G remains an active area of research in Bayesian clustering, and integrating more advanced model selection methods could be a promising direction for future work. To summarize, we select the number of groups using the posterior mode for its simplicity,

flexibility, and adequate performance for our modeling needs. More complex Bayesian model selection techniques may further improve performance but introduce additional complexity.

3.3 Learning rate selection

This section discusses the selection of learning rate parameter ψ in practice. Suppose we are interested in some functional of the parameters $\zeta = f(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and would like to construct confidence sets of ζ. A natural choice is to use quantiles of $\{\zeta^{(m)}\}_{m=1}^{M}$ for M posterior samples from the quasi-posterior distribution $\pi_{NT}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. As is discussed in Section 2.2, the learning rate ψ controls directly the dispersion of the posterior distribution—through its ability to explore alternative model spaces in the assignment step (8) and the sampling uncertainty in the parameter updating step (9)—and thereby the width of the confidence sets. In particular, as the learning rate decreases, the posterior gets more tilted toward the prior distribution, leading to more conservative confidence sets.

Therefore, a heuristic choice updates the learning rate as follows (Syring and Martin, 2019)

$$\psi^{(j+1)} \leftarrow \psi^{(j)} + (j+1)^{-a} \left(\mathbb{P}_{\psi^{(j)}} - (1-\alpha) \right) \tag{24}$$

where $j = 0, 1, \ldots$ is the temperature of the updating scheme and a is a cooling rate which facilitates the convergence of the updating scheme.[8] $\mathbb{P}_{\psi^{(j)}}$ is the *population* coverage probability of the confidence sets of interest, which depends on the current learning rate $\psi^{(j)}$. Intuitively, when the coverage probability exceeds the desired nominal level $(1-\alpha)$, the posterior distribution is too dispersed, and we would like to increase the learning rate so as to reflect a higher weight on the data. Therefore, whenever $\mathbb{P}_{\psi^{(j)}} - (1-\alpha) > 0$, the learning rate is increased and vice versa.

Although the updating scheme (24) is attractive, it requires knowledge of the population coverage probability $\mathbb{P}_{\psi^{(j)}}$, which in turn requires knowledge of the true parameter ζ^0 and the true probability distribution of the data. Syring and Martin (2019) propose to replace both objects through bootstrap.[9]

Specifically, for each bootstrapped sample, we apply Algorithm 1 to obtain an estimate, e.g., the posterior mode, of the parameter of interest ζ. The unknown population parameter ζ^0 is then approximated by the bootstrap average estimator. In a similar vein, although the true data distribution $\mathbb{P}_{\psi^{(j)}}$ is unknown and thus cannot be used to evaluate coverage

[8] In practice, I follow Syring and Martin (2019) and set $a = -0.51$. Alternative choices affect the speed of the convergence but the results are robust to different cooling rates.

[9] Since the time series dimension T is relatively short in the current setup, I use cross-sectional resampling bootstrap as in Kapetanios (2008). Alternatively we may consider cross-sectional dependence bootstrap proposed by Gonçalves and Perron (2020).

probability, we can approximate it using the average coverage rates across bootstrapped samples.

Algorithm 4 Learning Rate Calibration

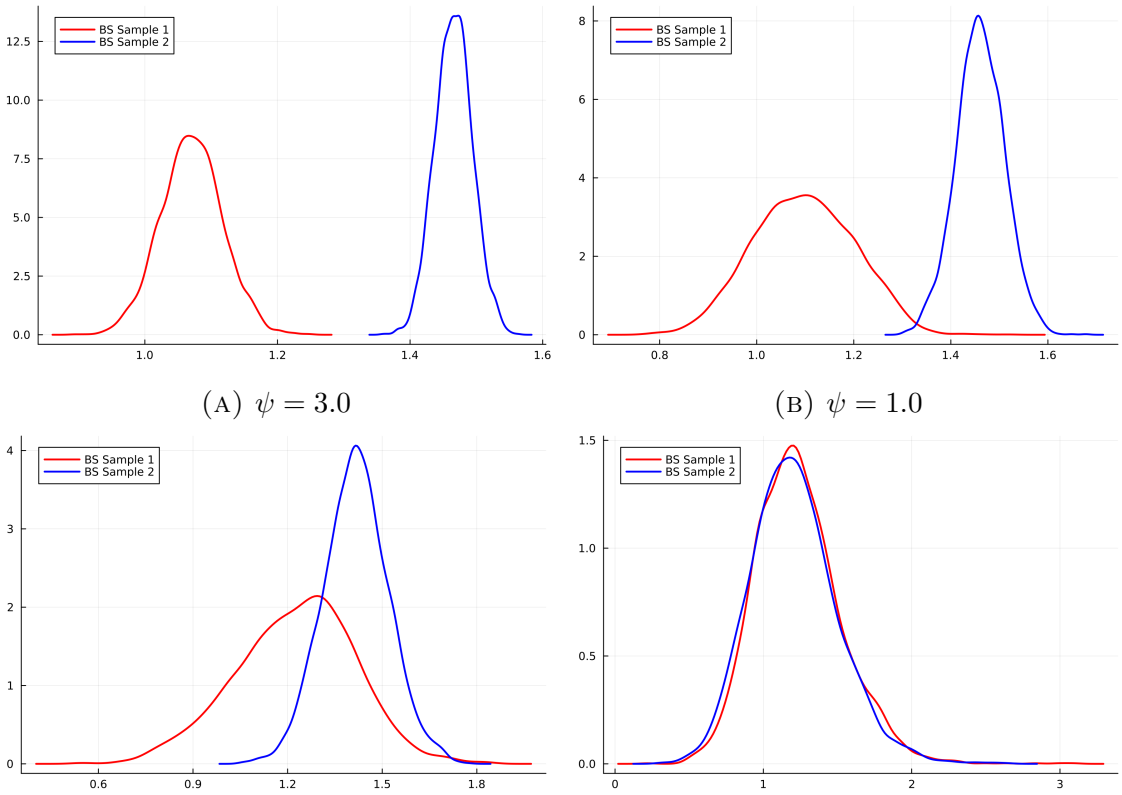
For a given learning rate $\psi^{(j)}$, do the following until $|\hat{\mathbb{P}}_{\psi^{(j)}} - (1-\alpha)| < \frac{1}{B}$

1: **Bootstrap Posterior**. For $b = 1, \ldots, B$, resample with replacement $\{w_{it}^{(b)}\}_{i=1}^{N}$ from the data, and sample from posterior distribution $\pi_{NT}^{*(b)}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \psi^{(j)})$ using Algorithm 1.

2: **Compute Empirical Coverage**. Construct confidence sets $\text{CS}^{(b)}$ and point estimate $\zeta^{(b)}$ from $\pi_{NT}^{*(b)}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \psi^{(j)})$, where ζ is a function of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Compute the empirical coverage rate $\hat{\mathbb{P}}_{\psi} = \frac{1}{B}\sum_b \mathbf{1}\left\{\overline{\zeta} \in \text{CS}^{(b)}\right\}$ where $\overline{\zeta} = \frac{1}{B}\sum_b \zeta^{(b)}$.

3: **Update Learning Rate**. For a given threshold α, update the learning rate by $\psi^{(j+1)} = \psi^{(j)} + (j+1)^{-a}(\hat{\mathbb{P}}_{\psi^{(j)}} - (1-\alpha))$.

Interestingly, the above calibration procedure can be viewed as a way to enhance algorithmic stability, a property that is recently exploited for valid post-selection inference (Zrnic and Jordan, 2023). As an illustration, Figure 1 presents the posterior densities of a group-level parameter for two bootstrapped samples, with varying learning rate. Panel 1(a) shows the case with relatively large ψ, under which each bootstrapped sample provides distinct point estimates. As a by-product, the empirical coverage of the confidence sets will be far below the desired level. As the learning rate decreases, the posterior densities of the two bootstrapped samples start to overlap, indicating that the algorithm produces more and more stable estimates. In the extreme case when ψ is close to zero, Panel 1(d) shows that the posterior densities fully overlap, and the resulting posterior modes are identical. Although the point estimates are still consistent, the confidence sets will be overly conservative. Therefore, similar to Zrnic and Jordan (2023), the learning rate calibration yields a sequence of confidence intervals with *tunable width*, reflecting the bias and variance tradeoffs.



(A) $\psi = 3.0$ (B) $\psi = 1.0$

(C) $\psi = 0.25$ (D) $\psi = 0.05$

FIGURE 1: LEARNING RATE CALIBRATION

Note: The figure presents variations of the posterior densities of a group-level parameter for two bootstrapped samples, with varying learning rate. The original data is generated as in (1) with $N = 100$ and $T = 10$, and the bootstrapped samples are generated by cross-sectional resampling. The true parameter value is $\beta = 1.6$.

Apart from the calibration procedure outlined above, alternative methods have been proposed to select the learning rate. For example, Lyddon et al. (2019) propose to calibrate ψ to match the asymptotic Fisher information, and Fasiolo et al. (2021) chooses learning rate so that the asymptotic variance of the parameters of interest is close to the sandwich-form covariance (Müller, 2013). Although both methods are more computationally efficient than the current approach, they are developed for finite dimensional models. It would be interesting to extend these approaches to the current framework in the future.

4 Asymptotic properties

This section establishes frequentist guarantees for the quasi-Bayesian clustering framework. We start with a general consistency and posterior contraction rate result under a set of high-level conditions. Section 4.1 and 4.2 provide primitive conditions for popular estimation methods, including generalized method of moments (GMM) and M-estimation, under which the general result applies. Before we move on, let us introduce some notation. For any parameters $\theta = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\tilde{\theta} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$, define the following (pseudo) metrics:

$$d_{MS}(\theta, \tilde{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \|\beta_{\gamma_i} - \tilde{\beta}_{\tilde{\gamma}_i}\|^2, \quad d_M(\theta, \tilde{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{\gamma_i \neq \gamma_i^0\right\} \tag{25}$$

$$d_H(\theta, \tilde{\theta}) = \max\left\{ \max_{g \in \{1,\ldots,G_2\}} \min_{\tilde{g} \in \{1,\ldots,G_1\}} \|\tilde{\beta}_{\tilde{g}} - \beta_g\|, \max_{\tilde{g} \in \{1,\ldots,G_1\}} \min_{g \in \{1,\ldots,G_2\}} \|\tilde{\beta}_{\tilde{g}} - \beta_g\| \right\} \tag{26}$$

where $\|\cdot\|$ denotes the Euclidean norm. Similar to the notation of posterior density $\pi_{NT}(\cdot)$, the posterior distribution of θ is denoted by $\Pi_{NT}(\cdot)$.

Below we give general results on the asymptotic properties of the quasi-posterior distribution (12). To better understand the asymptotic behavior of the quasi-posterior distribution, it is helpful to rewrite the posterior as the following:

$$\Pi_{NT}(A) = \frac{\int_A \exp\left[-NT\psi\left(L_{NT}(\theta) - L_N(\theta) + L_N(\theta) - L_N(\theta^0)\right)\right] \pi(\theta) \mathrm{d}\theta}{\int_\Theta \exp\left[-NT\psi\left(L_{NT}(\theta) - L_N(\theta) + L_N(\theta) - L_N(\theta^0)\right)\right] \pi(\theta) \mathrm{d}\theta} \tag{27}$$

where $A = \{d(\theta, \theta^0) \leq \epsilon\} \subset \Theta$ is the neighborhood of the true parameter θ^0 defined by some metric $d(\cdot, \cdot)$. Decomposition (27) suggests that for the quasi-posterior to concentrate around A, we would need three assumptions: (1) to separate the true parameter from its neighborhood values, i.e., lower bounding $L_N(\theta) - L_N(\theta^0)$; (2) to control the approximation errors $L_{NT}(\theta) - L_N(\theta)$ resulting from using the sample loss; and (3) to discipline the prior knowledge $\pi(\theta)$ so that the denominator does not vanish too fast. The following assumptions formalize the intuition.

Assumption 1. *Let $d(\cdot, \cdot)$ be some (pseudo) metric on $\Theta \times \Theta$. We assume that the following conditions hold:*

A. (Identification). We have $\inf_{\theta:\, d(\theta,\theta^0)>\epsilon} L_N(\theta) - L_N(\theta^0) > \tilde{\chi}(\epsilon)$.

B. (Uniform convergence). $\sup_{\theta \in \Theta} \left|L_{NT}(\theta) - L_N(\theta)\right| = o_p(1)$.

C. (Prior mass). $\Pi\left(\{\theta \colon L_N(\theta) - L_N(\theta^0) \leq \epsilon\}\right) \geq \tilde{c}_{NT}(\epsilon)$ for some non-stochastic sequence $\tilde{c}_{NT}$ possibly depending on the sample size.

Overall, Assumption 1 is similar to the conditions required in Miller (2021, Theorem 3) and Syring and Martin (2023, Section 3). Specifically, Assumption 1.A requires that the population loss is uniquely minimized at the true parameters θ^0. This assumption is standard in the literature on grouped panel data models; see for example Assumption A1(ii) in Su et al. (2016). Assumption 1.B requires uniform convergence of the sample loss function to its population counterpart. This assumption is relatively mild and can be verified under various primitive conditions as in Newey (1991). Finally, Assumption 1.C imposes a prior mass condition in the spirit of condition (2.4) in Ghosal et al. (2000). Intuitively, such conditions requires that the prior distribution does not shrink too fast around the neighborhood $\{\theta\colon L_N(\theta) - L_N(\theta^0) \leq \epsilon\}$ of the true parameters, which guarantees that the denominator of the quasi-posterior distribution (12) is lower bounded. In practice, such prior mass condition may be hard to verify. Therefore, we often supplement it with some continuity condition on the loss, which translates the control over excess loss $L_N(\theta) - L_N(\theta^0)$ to the control over the distance $d(\theta, \theta^0)$; see for example Assumption 4.D below.

With the above assumptions, we have the following result on the consistency of the quasi-posterior distribution.

Theorem 1. Under Assumption 1, we have for any $\psi > 0$ and for some $\delta > 0$

$$\mathbb{E}_0 \Pi_{NT}\left(\left\{\theta\colon d(\theta, \theta^0) > \epsilon\right\}\right) \leq \frac{\exp\left[-NT\psi\left(\tilde{\chi}(\epsilon) - o(1) - \delta\right)\right]}{\tilde{c}_{NT}(\delta)} \,. \tag{28}$$

The above theorem provides a unified framework to understand the consistency of the quasi-posterior distribution. As is clear, the quasi-posterior is consistent only when

$$\tilde{\chi}(\epsilon) - o(1) - \delta + \frac{\ln \tilde{c}_{NT}(\delta)}{NT\psi} > 0 \,. \tag{29}$$

This effectively reflects the tradeoff between the signal strength, the regularity of the loss function and the data, and the prior distribution. Specifically, the first term $\tilde{\chi}(\epsilon)$ comes from the identification condition and measures the strength of the signal. The second term arises from the approximation of the sample loss function to the population loss, which under uniform convergence condition is $o(1)$. The third term comes from the restriction that the excess loss is controlled $L_N(\theta) - L_N(\theta^0) \leq \delta$. Finally, the fourth term is the prior mass condition, which measures the complexity of the estimation problem; see for example discussions in Grunwald and Mehta (2020).

Notice that (29) trivially holds when $\tilde{\chi}(\epsilon) = O(1)$. In this case, we can always select an arbitrarily small δ such that $\tilde{\chi}(\epsilon) - \delta - o(1) > 0$. However, recall that by construction $\tilde{c}_{NT}(\delta) \leq 1$. Imposing a small δ essentially inflates the fourth term $\ln \tilde{c}_{NT}(\delta)$, which reflects

the tradeoff between prior informativeness and the regularity of the loss function. In the case of strong signal $\tilde{\chi}(\epsilon) = O(1)$, this tradeoff does not matter, since $\frac{\ln \tilde{c}_{NT}(\delta)}{NT\psi}$ is $o(1)$. However, such tradeoff can be important in two cases. First, when we consider posterior contraction rates, ϵ is no longer some *fixed* constant but some vanishing sequence $\epsilon_{NT} \to 0$. In this case, $\tilde{\chi}(\epsilon)$ has to shrink at a slower rate than $\frac{\ln \tilde{c}_{NT}(\delta)}{NT\psi}$, which serves as an effective bound on the posterior contraction rate. Second, in some cases, the parameter θ may not be strongly identified, where $\chi(\epsilon)$ is vanishing even with fixed $\epsilon > 0$. In this case, the prior mass condition $\tilde{c}_{NT}(\delta)$ plays a crucial role in ensuring the consistency of the quasi-posterior distribution. Moreover, the lower bound $\tilde{\chi}(\epsilon)$ depends on the choice of the distance metric $d(\cdot, \cdot)$, which can affect the consistency of the quasi-posterior distribution. In the next subsection, I discuss the choice of distance metric in more details.

A natural follow up question is how fast the quasi-posterior distribution contracts around the true parameter. Specifically, we aim to find a sequence $\epsilon_{NT} \to 0$ such that

$$\mathbb{E}_0 \Pi_{NT} \left(\left\{ \theta \colon d(\theta, \theta^0) > \epsilon_{NT} \right\} \right) \to 0 \ . \tag{30}$$

To this extent, I follow the recent literature that establishes posterior contraction rates by exploiting the close connection between quasi-Bayesian approach and empirical risk minimization (Grunwald and Mehta, 2020; Syring and Martin, 2023). In particular, the PAC-Bayesian literature has long studied the risk performance of the quasi-Bayesian framework, by deriving various useful PAC-Bayes inequalities (e.g., Alquier, 2023). Moreover, such inequalities are often "model-free" and thus facilitates adaptation to different setup. Below I state a concentration rate result based on the risk bound in Jiang and Tanner (2008).

Assumption 2. *Let $d(\cdot, \cdot)$ be some (pseudo) metric on $\Theta \times \Theta$. We assume that the following conditions hold:*

A. (Identification). We have $\{\theta \colon d(\theta, \theta^0) \geq \epsilon_{NT}\} \subseteq \{\theta \colon L_N(\theta) - L_N(\theta^0) \geq a(\epsilon_{NT})\}$.

B. (Uniform convergence). $\mathbb{P}_0 \left(\sup_{\theta \in \Theta} \left| L_{NT}(\theta) - L_N(\theta) \right| \geq \frac{a(\epsilon_{NT})}{5} \right) = b_{NT}$.

C. (Smoothness). There exists some constant $0 < \tilde{c}_M < \infty$ such that $\left| L_N(\theta) - L_N(\tilde{\theta}) \right| \leq \tilde{c}_M d(\theta, \tilde{\theta})$ for any $\theta, \tilde{\theta} \in \Theta$.

D. (Prior mass). $\Pi \left(\{\theta \colon d(\theta, \theta^0) \leq \epsilon_{NT}\} \right) \geq c(\epsilon_{NT})$ for some non-stochastic sequence c_{NT} possibly depending on the sample size.

Overall, the above assumptions are similar to Assumption 1, except that the dependence between convergence rates are explicitly assumed. Specifically, Assumption 2.A is implied by

primitive conditions in Theorem 3 of Miller (2021). Assumption 2.B is a stronger version of 1.B, as a_{NT} is now a vanishing sequence depending on ϵ_{NT}.The smoothness condition remains the same and is stated for completeness. Finally, the prior mass condition is imposed on the distance metric $d(\cdot,\cdot)$ instead of the loss function. Given the assumptions, we have

Theorem 2. Under Assumption 2, we have

$$\mathbb{E}_0\Pi_{NT}\left(\left\{\theta\colon d(\theta,\theta^0)\geq\epsilon_{NT}\right\}\right)\leq b_{NT}+\frac{\exp\left[-\frac{2}{5}NT\psi a(\epsilon_{NT})\right]}{c(a(\epsilon_{NT})/5c_M)} \tag{31}$$

The above theorem provides a succinct summary of the driving forces behind the convergence rate. First, as ϵ_{NT} converges to 0, we would expect $a(\epsilon_{NT})$ also converging to zero, which leads to a larger b_{NT}. Therefore, the uniform convergence rate of the loss function serves as the first constraint of the posterior convergence rate. Second, the convergence rate is further constrained by the prior mass condition. On the one hand, the function $c(\cdot)$ can possibly depend on the data dimension N and T. On the other hand, as $a(\epsilon_{NT})$ gets smaller, the prior mass also shrinks towards zero. The rate at which ϵ_{NT} converges to zero depends on the specific prior and uniform convergence rate Similar patterns are already revealed in (29).

Finally, it is often difficult to determine the number of groups G in practice. Below, I provide a result on the posterior *ratio* consistency, under a slightly stronger assumption of prior.

Assumption 3. *For $k\in\{1,2,\ldots\}$, assume that*

A. $\Pi(G=k)>0$.

B. $\Pi(\{\mathcal{C}\colon|\mathcal{C}_g|/N>0\})=1$.

C. $\Pi\left(\beta_g=\beta_l|G=k\right)=0$ *for* $1\leq g<l\leq k$.

D. $\inf_{\|\beta-\beta^0_{\gamma_i^0}\|\geq\epsilon}\mathbb{E}[l_{iT}(\beta)-l_{iT}(\beta^0_{\gamma_i^0})]>\check{\chi}(\epsilon)>0$.

The above assumptions are similar to Condition 2.2 in Miller (2023). The first part assumes that the prior assigns positive mass on the (true) number of groups G, which is a necessary condition for the prior mass condition 1.C. The second assumption imposes that the group-sizes are non-negligible.[10]. The third assumption requires that the group-level parameters are distinct.[11] In practice, it is imposed by augmenting the priors with the

[10]In terms of Bayesian mixture model, it is equivalent to assume that mixture weigths $\eta_1,\ldots,\eta_G>0$ (Miller, 2023)

[11]A probabilistic approach for such constraint is the repulsive prior with penalty on similar group-level parameters; see for example Natarajan et al. (2023).

constraints that $\beta_g \neq \beta_l$ for $g \neq l$. For example, in case of the MFM-normal priors as in (6), Assumption 3.C is satisfied if the prior is modified to[12]

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\gamma}|\eta)\pi(\eta) \left(\prod_{g=1}^{G} \pi(\beta_g) \right) \mathbf{1}\left\{ \boldsymbol{\beta} \colon \beta_g \neq \beta_l, 1 \leq g < l \leq G \right\} . \tag{32}$$

The final assumption is the identification condition, imposed at the individual-level loss, which facilitates the comparison of pooled loss under various partitions. Overall, the assumption imposes identifiability constraints on the model, which greatly simplifies the asymptotic analysis of the number of groups. In the absence of such constraint, overfitting G leads to degenerate group structure, where either group-level parameters are allowed to coincide across groups, or group shares are vanishing as the sample size increases, both make the interpretation of the latent group structure difficult. Given the above assumptions, we have the following result on the posterior ratio consistency.

Theorem 3. Under Assumption 1 and 3.C, we have for some $0 < c < 1$,

$$\max_{G \neq G^0} \frac{\pi_{NT}(G)}{\pi_{NT}(G^0)} \leq \frac{\exp\left[-NT\psi\left(c\check{\chi}(\epsilon) - o(1) - \delta\right)\right]}{\tilde{c}_{NT}(\delta)} . \tag{33}$$

The above theorem serves as a posterior ratio consistency result, comparable to Theorem 3 in Duan et al. (2023), but weaker than the strong selection consistency where $\sum_G \frac{\pi_{NT}(G)}{\pi_{NT}(G^0)} \xrightarrow{p} 0$ (Narisetty et al., 2019). It is clear that the posterior ratio converges to zero under conditions similar to (29). The theorem implies that the mode of the quasi-posterior distribution is highest in the true number of groups G^0, and thus the posterior mode can be used to select the number of groups.

In the following subsections, I provide primitive conditions for popular estimators, including generalized method of moments (GMM) and M-estimation, under which the general result applies.

4.1 M-estimation

Consider the following M-estimation problem in grouped panels (Liu et al., 2020; Wang and Su, 2021):

$$L_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} h(w_{it}; \beta_{\gamma_i}) \tag{34}$$

[12] See Kim and Gao (2020) for more examples of structural constraints on the group-level parameters.

where $-h(w_{it};\beta_{\gamma_i})$ is the logarithm of the (pseudo) likelihood function as in example 5. In this section, we apply the general results to the above loss with the mixture of finite mixture prior on the group structure, and a normal prior on the group level parameters (6). We first impose a set of regularity conditions as follows:

Assumption 4 (M-estimation)**.**

A. $\beta_g^0 \in \mathcal{B}$ *for all* $g = 1, \ldots, G^0$ *where* $\mathcal{B}$ *is a convex compact subset of* $\mathbb{R}^d$*.*

B. $\{w_{it}\}_{t=1,\ldots,T}$ *are independent across* i*. For each* i*, it is stationary strong mixing with mixing coefficient* α_i*, and* $\alpha \equiv \max_i \alpha_i$ *satisfies* $\alpha(\tau) \leq c_\alpha \rho^\tau$ *for some* $c_\alpha > 0$ *and* $\rho \in (0,1)$*.*

C. For any $\epsilon > 0$*, we have*

$$\min_i \left[\inf_{\|\beta - \beta^0_{\gamma_i^0}\|^2 > \epsilon} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[h(w_{it};\beta) - h(w_{it};\beta^0_{\gamma_i^0}) \right] \right] = \chi(\epsilon) > 0 \ . \tag{35}$$

D. There exists a non-negative function $M(\cdot)$ *such that* $\sup_{\beta\in\mathcal{B}} |h(w;\beta)| \leq M(w)$*, and* $|h(w;\beta) - h(w;\tilde{\beta})| \leq M(w)\|\beta - \tilde{\beta}\|$ *for all* $\beta, \tilde{\beta} \in \mathcal{B}$*. Moreover,* $\sup_i \mathbb{E}\big|M(w_{it})\big|^q < c_M$ *for some* $c_M < \infty$ *and* $q \geq 6$*.*

E. Assume that $N^2 = O(T^{q/2-1})$ *where* $q \geq 6$ *is the same constant in 4.D.*

F. G^0 *is fixed and* $\min_{g\neq l} \|\beta_g^0 - \beta_l^0\| > 0$ *for all* $g, l \in \{1, \ldots, G^0\}$*.*

G. For all $g \in \{1, \ldots, G^0\}$*,* $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\gamma_i^0 = g\} > \underline{\eta} > 0$*.*

H. Assume that the finite mixture prior on the group structure and the normal prior on the group-level parameters are given by (6)*.*

Overall, Assumption 4 is similar to Assumption A1 in Su et al. (2016), which is standard in the grouped panel literature (e.g., Bonhomme and Manresa, 2015; Liu et al., 2020; Huang, 2021). In particular, 4.A assumes that the parameter space $\mathcal{B}$ is compact and thus the diameter $\text{diam}(\mathcal{B}) < \infty$. 4.F requires that the group-level parameters are well-separated, and 4.G imposes that the true group sizes are non-negligible. Assumption 4.B requires the data to be strong mixing with geometric mixing rate, which facilitates the use of Bernstein type inequality (e.g., Merlevède et al., 2011). Assumption 4.D is a smoothness condition on the loss function, and requires the envelope function $M(w_{it})$ to have finite sixth moment. Finally, Assumption 4.E allows the time series dimension to grow slower than the cross-sectional dimension. Notice that Assumptions 4.B-4.E may be relaxed. For example, if

we impose a faster decay rate in 4.B, e.g., exponential decay such as Assumption 2(c) in Bonhomme and Manresa (2015), then we could allow the cross-sectional dimension to be much larger than the time series dimension. The crucial point is that the tradeoffs among different conditions embodied in (29) is preserved. With the above assumptions, we have the following result.

Theorem 4 (Consistency)**.** Under Assumption 4 and known G, we have for any $\epsilon > 0$,

$$\Pi_{NT}\left(\left\{\theta : d(\theta, \theta^0) > \epsilon\right\}\right) \xrightarrow{P_0} 0 \tag{36}$$

as N, T go to infinity, for $d(\cdot, \cdot)$ being d_{MS}, d_M and d_H defined in (25) and (26).

Theorem 4 states that the quasi-posterior distribution concentrates on the true parameter values, in terms of the average parameter estimation errors, the average misclassification errors, and the Hausdorff distance metrics. Moreover, when the metrics considered are cross-sectional averages (d_{MS} and d_M), we do not need to assume the knowledge of the true group number. In fact, even when considering the convergence of the Hausdorff distance d_H, correct knowledge of G is not a necessary condition. For example, when the diffuse prior (6) is used with constraints on the minimal group size such as Assumption 4.G, the convergence of d_H is almost identical to the counterpart for d_{MS}.

Next we would like to establish the posterior contraction rates of the quasi-Bayesian posteriors. Next we apply Theorem 2 to study the posterior contraction rate of the distance metrics in Theorem 4. Specifically,

Theorem 5. Under Assumption 4 and assume the knowledge of G, we have for $\epsilon_{NT} = O(T^{-1})$

$$\Pi_{NT}\left(\left\{\theta : d(\theta, \theta^0) > \epsilon_{NT}\right\}\right) \xrightarrow{P_0} 0 \tag{37}$$

as N, T go to infinity, for $d(\cdot, \cdot)$ being d_{MS}, d_M and d_H defined in (25) and (26).

Theorem 5 recovers the results in Su et al. (2016, Theorem 2.1). This is not surprising, since my Assumption 4 is comparable to their Assumption A1.

4.2 GMM-type estimators

We now consider another popular class of loss functions, the generalized method of moments (GMM) criterion. For completeness, we restate the GMM criterion as follows:

$$L_{NT}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} m(w_{i,t}; \beta_{\gamma_i})\right)^{\top} \hat{\Omega}_i \left(\frac{1}{T}\sum_{t=1}^{T} m(w_{i,t}; \beta_{\gamma_i})\right) . \tag{38}$$

Assumption 5 (GMM)**.**

A. $\beta_g^0 \in \mathcal{B}$ *for all* $g = 1, \ldots, G^0$ *where* $\mathcal{B}$ *is a convex compact subset of* $\mathbb{R}^d$*.*

B. For all i*,* $\{w_{i,t} \colon 1 \leqslant t \leqslant T\}$ *is a strictly stationary* α*-mixing sequence of random vectors with mixing coefficients* $a_i(\cdot)$ *such that* $a(\cdot) = \max_i a_i(\cdot)$ *satisfies* $a(\tau) \leqslant c_\alpha \rho^\tau$ *for some* $c_\alpha > 0$ *and* $\rho \in (0, 1)$*.*

C. We have $\mathbb{E}m(w_{it}; \beta^0_{\gamma_i^0}) = 0$ *for some* $\beta^0_{\gamma_i^0} \in \operatorname{int}(\mathcal{B})$*. Moreover, For any* $\epsilon > 0$*, we have* $\min_i \inf_{\|\beta - \beta^0_{\gamma_i^0}\|^2 > \epsilon} \left\|\mathbb{E}m(w_{it}; \beta_{\gamma_i})\right\| > \chi(\epsilon) > 0$*.*

D. There exists a non-negative function $M(\cdot)$ *such that* $\|m(w_{it}; \beta)\| \leq M(w_{it})$ *for all* $\beta \in \mathcal{B}$*, and* $\|m(w_{it}; \beta) - m(w_{it}; \tilde{\beta})\| \leq M(w_{it})\|\beta - \tilde{\beta}\|$ *for all* $\beta, \tilde{\beta} \in \mathcal{B}$*. Moreover,* $\sup_i \mathbb{E}\left|M(w_{it})\right|^q < c_M$ *for some* $c_M < \infty$ *and* $q \geq 6$*.*

E. There exists a deterministic sequence of symmetric positive definite matrices $\{\hat{\Omega}_i\}_{i=1}^N$ *such that for any* $\nu > 0$ *we have* $\sup_i \|\hat{\Omega}_i - \Omega_i\| = o_p(N^{-1})$*.*

F. Assume that $N^2 = O(T^{q/2-1})$ *where* $q \geq 6$ *is the same constant in 5.D.*

G. G^0 *is fixed and* $\min_{g \neq l} \|\beta_g^0 - \beta_l^0\| > 0$ *for all* $g, l \in \{1, \ldots, G^0\}$*.*

H. For all $g \in \{1, \ldots, G^0\}$*,* $\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\gamma_i^0 = g\} > \underline{\eta} > 0$*.*

Overall, Assumption 5 is similar to conditions in Fernández-Val and Lee (2013) and Cheng et al. (2019). In fact, since the GMM objective function (38) is of similar form to the M-estimation objection (34), the assumptions are also similar to the ones in the previous section. The main difference is that in the identification assumption 5.C, the population loss is zero, as required by the validity of moment conditions. Moreover, the weighting matrix in 5.E also introduces an additional complication, as the uniform convergence rate generally depends on the choice of weighting matrix.

With the above assumption, we have the following result.

Theorem 6 (Consistency)**.** Under Assumption 5 and known G, we have for any $\epsilon > 0$,

$$\Pi_{NT}\left(\left\{\theta : d(\theta, \theta^0) > \epsilon\right\}\right) \xrightarrow{P_0} 0 \tag{39}$$

as N, T go to infinity, for $d(\cdot, \cdot)$ being d_{MS}, d_M and d_H defined in (25) and (26).

5 Simulation study

This section investigates the finite-sample performance of the quasi-Bayesian clustering method. Section 5.1 describes the simulation design and Section 5.2 presents the simulation results with fixed number of groups.

5.1 Simulation design

We consider first the class of linear panel data models.

DGP 1: Linear panel data model with homoskedastic errors. The data are generated from the following linear panel data model:

$$y_{it} = x_{it}\beta^0_{\gamma_i^0} + \mu_i^0 + \epsilon_{it} \tag{40}$$

where the covariates $x_{it} \in \mathbb{R}^2$ are generated from standard normal and $\mu_i^0 \overset{\text{i.i.d.}}{\sim} N(0,1)$ are the individual fixed effects. Group-level parameters are set to $\beta_1^0 = (0.4, 1.6)^\top, \beta_2^0 = (1,1)^\top, \beta_3^0 = (1.6, 0.4)$ when there are three groups ($G = 3$), and $\beta_1^0 = (0.4, 1.6)^\top, \beta_2^0 = (1,1)^\top$ when there are two groups ($G = 2$).

DGP 2: Linear panel data model with heterogeneous errors. The data are generated from a linear panel data model as in (40), except that the error terms are heterogeneously distributed (Patton and Weller, 2022). Specifically, for each unit i, e_{it} is randomly drawn from $N(0,1)$, Exp(2), $Unif(-3,3)$, $\chi^2(4)$ and $t(5)$ with equal probability, standardized to have mean zero and unit variance.

In each of the designs, I simulate 10,000 MCMC draws with 1,000 burn-in for 300 replications. For the learning rate selection, I calibrate it with 100 bootstrapped samples.

5.2 Known number of groups

This section reports the results where the true number of groups is known a priori. In particular, I set the prior $\pi(G^0|\lambda) = 1$ and zero otherwise. As a benchmark, I also report the results of k-means clustering following Bonhomme and Manresa (2015). For both methods, I calculate the RMSE of the estimated group-level parameters, and compute the coverage rates of confidence intervals.

Table 2-Table 3 report the results. Overall, the patterns are similar to the results in Section 2, demonstrating that the quasi-Bayesian framework performs well in more complicated setup.

Noticeably, the quasi-Bayesian confidence sets seem overly conservative in some cases.

This is because the posterior draws are subject to label switching problem (Stephens, 2000), a well known issues in the class of mixture models. When the draws highly overlap, as is the case when the data is noisy, deterministic post-processing algorithms may fail to correctly re-label the group-level parameters. In this case, the posterior distributions of group-level parameters are contaminated by outlier draws from other groups, thereby inflating the confidence sets.

TABLE 2: HOMOSKEDASTIC DESIGN ($G = 2$)

Sample Size	**Metrics**	$\sigma^2 = 1.0$		$\sigma^2 = 2.0$	
		QB	KM	QB	KM
$N = 100,\ T = 10$	AC	86.77	87.81	61.23	71.24
	BR	24.65	17.57	37.64	16.15
	RMSE	16.89	12.37	52.90	47.01
	Coverage	95.2	82.7	96.6	43.4
	RMSE (AE)	3.32	3.37	6.51	6.74
	Coverage (AE)	97.7	-	99.0	-
$N = 100,\ T = 20$	AC	95.92	96.01	73.95	81.34
	BR	18.59	15.79	27.95	14.46
	RMSE	9.89	6.80	31.55	22.60
	Coverage	96.8	92.1	95.4	67.4
	RMSE (AE)	2.24	2.25	4.69	4.72
	Coverage (AE)	97.0	-	97.3	-

Note: Definition of the evaluating metrics is given by the footnote of Table 1.

Table 3: Distributional Heteroskedasticity

Sample Size	Metrics	$G = 2$		$G = 3$	
		QB	KM	QB	KM
$N = 100,\ T = 10$	AC	87.06	88.55	80.19	83.55
	BR	25.46	17.90	37.40	21.00
	RMSE	16.71	12.71	31.02	21.84
	Coverage	94.83	81.08	96.67	76.33
	RMSE (AE)	3.36	3.42	3.34	3.43
	Coverage (AE)	96.50	-	98.67	-
$N = 100,\ T = 20$	AC	95.90	95.98	94.28	94.50
	BR	18.85	16.08	24.83	19.18
	RMSE	10.37	7.27	16.30	11.05
	Coverage	95.08	90.50	96.44	88.67
	RMSE (AE)	2.32	2.33	2.32	2.32
	Coverage (AE)	95.67	-	97.00	-

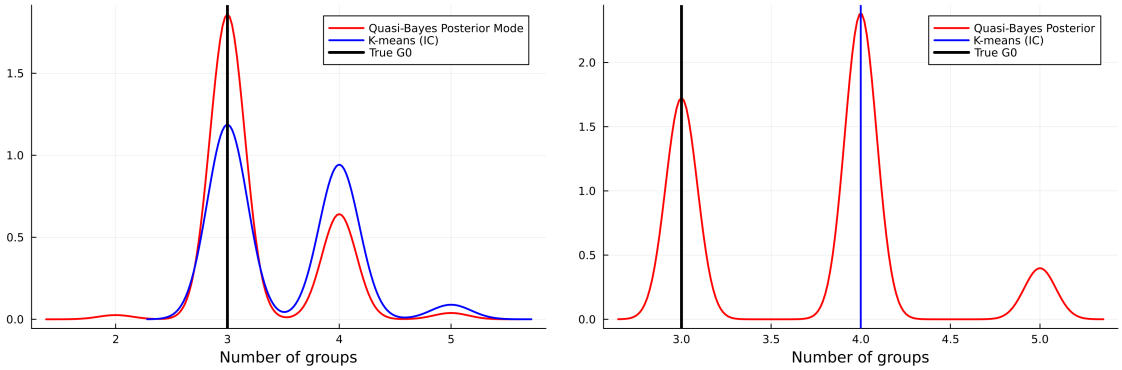
Note: Definition of the evaluating metrics is given by the footnote of Table 1.

5.3 Inference on the number of groups

The next exercise is to examine the performance of the quasi-Bayesian framework in determining the number of groups. In particular, the quasi-Bayesian approach estimates the number of groups by the posterior mode, while the k-means estimator estimates the number of groups by the information criterion (Su et al., 2016). Figure 2 shows the results with homoskedastic errors.

First, panel 2(a) shows the density of the selected group number across replications. Consistent with the theoretical results in Section 4, the posterior mode of the quasi-Bayesian approach tends to concentrate around the true number of groups. Moreover, on average it is more accurate than the frequentist information criterion. To understand the superior performance of the quais-Bayesian approach, panel 2(b) plots the quasi-posterior density of the group number in a single replication, along with the group number selected by the k-means algorithm (black line). As is clear, the frequentist information criterion fails to take into account the estimation uncertainty in the group number, whereas the quasi-Bayesian approach correctly accounts for it, by assigning non-zero probability to multiple plausible values of the group number. Overall, the simulation results highlight that encoding uncertainty in group

number inference via the posterior, rather than selecting a single G, improves discovery of the group patterns.



(A) Average Selection Probability (B) Selection on Single Run

Figure 2: Inference on Group Number

Note: The figure presents the posterior inference of the quasi-Bayesian procedure on the group number, and compare its performance with k-means clustering estimator with the number of groups selection by the information criterion (Su et al., 2016). Panel 2(a) shows the density of the selected group number across 200 replications. Panel 2(b) shows the posterior density of the quasi-Bayesian approach, along with the point estimate given by the k-means algorithm. The black vertical line indicates the true number of groups.

6 Empirical study: heterogeneous income risks

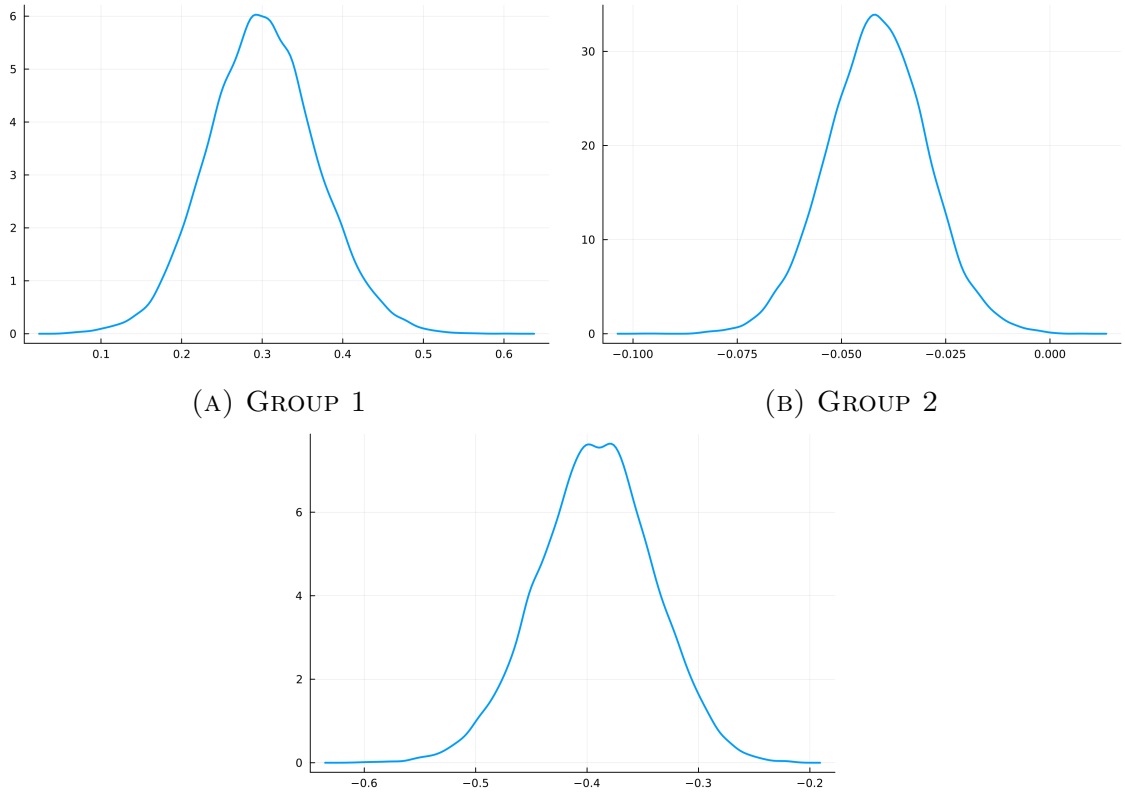
Previous studies have found substantial heterogeneity in the cyclicality of household income risk, with higher income households typically experiencing less cyclical income variation (Guvenen et al., 2014; Patterson, 2023). However, these studies have relied on ad-hoc classifications of households into groups based on observable characteristics. Although intuitive, such approach suffers from the drawback that important heterogeneity may be missed or incorrectly attributed to the assigned variable (Aguiar et al., 2020).

In this section, I apply the quasi-Bayesian approach to jointly estimate group-level income risks over the business cycle and uncover the latent group structure, using biennial panel data on households from the PSID between 1999-2009 (Arellano et al., 2017). In particular, we are interested in recovering heterogeneity in income cyclicality across groups, and examine household characteristics that help explain group membership. To do so, I estimate the following linear panel data model in the spirit of Guvenen et al. (2014)

$$\ln y_{it}^{(r)} = Y_t \beta_{\gamma_i} + \epsilon_{it} \tag{41}$$

where $\ln y_{it}^{(r)}$ is the residualized log total income, Y_t is the aggregate unemployment rate, and β_{γ_i} is the cyclicality of interest.

The quasi-Bayesian approach identifies three latent groups in the data, with the posterior distribution of the group-level parameters shown in Figure 3. As we can see, the income elasticity to unemployment rate is *positive* for Group 1, indicating that as unemployment rate increases by 1%, household (total) income increases by one-third. In contrast, Group 3 experiences income losses of similar magnitude, and Group 2 is mildly affected by the unemployment fluctuations.



(A) Group 1

(B) Group 2

(C) Group 3

Figure 3: Group Level Income Cyclicality

Note: The figure presents the group-level income cyclicality β_{γ_i} defined in model (41).

To better understand the mechanisms behind the heterogeneous income cyclicality, I show in Table 4 the average household characteristics by group. Consider first the demographic characteristics. Contrary to conventional wisdom, the three groups are remarkably similar in terms of age, education, and family size. Therefore, at least in the PSID sample, standard life-cycle considerations (Catherine, 2022) or skill distributions (Braxton et al., 2021) do not seem

to be the decisive factors behind heterogeneous income risks. Moreover, household income and wealth are correlated with the group membership, corroborating the results in Guvenen et al. (2014). As shown by the variation across rows, however, no single indicator such as total family income is able to fully capture the group differences. For example, income alone does not fully differentiate the groups. On one hand, Group 1 has the highest total family income yet the lowest labor income, highlighting the importance of variations in income categories. On the other hand, although Group 1 has the highest total family income, it also has the lowest stock values, suggesting different roles of income and wealth. Group 3 provides another example where indicators do not fully align. This group has the highest financial assets and cash holdings, and so does not fit the "wealthy hand-to-mouth" characterization (Kaplan et al., 2014). However, they are the most negatively affected by rising unemployment rates, contradicting the existing findings (Patterson, 2023). Interestingly, Group 3 also has the highest transfer income and pensions & annuities values, suggesting they are vulnerable to recessions *despite not being liquidity constrained*. To summarize, the conventional dichotomy of "rich" versus "poor", or "constrained" versus "unconstrained" households, tends to obscure the more nuanced differences in the latent group structure.[13]

Next, it is interesting to contrast the quasi-Bayesian clustering results with the k-means clustering. Importantly, although Figure 5 in the appendix shows that both methods identify a similar number of groups, the underlying group structure greatly differs. The k-means algorithm tends to partition units into balanced groups of roughly equal size. By contrast, the quasi-Bayesian approach allows for relatively small group sizes and unbalanced partitions. Such difference turns out to be important when we interpret the group assignment with group-level characteristics. As Table B.2 shows, the demographic differences across groups are less prominent in k-means clustering. Among the discernable differences, the two clustering results do exhibit some broad agreements. For example, the Group 4 households in k-means clustering have the highest stocks of pension and annuities, and suffer the most when unemployment rate increases, similar to Group 3 households discovered by the quasi-Bayesian approach. However, the confidence intervals of the k-means clustering are substantially narrower than those for the quasi-Bayesian approach. Given the extensive simulation results in the previous sections, the validity of the k-means confidence intervals is questionable. On the contrary, the quasi-Bayesian approach provides more realistic uncertainty estimates for the heterogeneous income cyclicality.

[13] Figure 4 in the appendix provides a graphical illustration. When total assets and labor income are considered in isolation, the resulting group partitions appear unrelated to these indicators. However, the full analysis shows these factors do play a role in differentiating the groups.

Table 4: Household characteristics by group (quasi-Bayes)

	Group 1	Group 2	Group 3
Age	45	45	45
Education	14	14	14
Family Size	4	3	4
Total Family Income	145120	117048	129835
Taxable Income (Head and Wife)	138145	110208	121237
Transfer Income (Head and Wife)	884	2565	3607
Labor Income (Head)	45069	73857	80745
Labor Income (Wife)	35480	29137	29559
Hours Worked (Head)	2253	2247	2170
Hours Worked (Wife)	1425	1443	1042
House Value	244989	245389	309352
Stocks Value	9564	46405	99193
Pensions & Annuities	46000	46016	87341
Cash	27291	27924	49622
Bonds	12545	15270	22379
Other Debt	7608	7324	6063
Financial Assets	228621	161973	280191
Total Assets	575585	318244	488282
Nondurable Consumption	9258	8602	8331
Services Consumption	35072	33233	35793
Total Consumption	40394	38204	40268
Count	15	713	34
β_g (%)	29.77	-2.97	-35.81
$CS(\beta_g)$ (%)	[17.19,43.72]	[-6.53,-1.76]	[-49.73,-29.27]

Note: This table reports the group-level averages of household characteristics. The last two rows report the group-level coefficients (posterior mode), and the 95% confidence sets constructed by 2.5% and 97.5% quantiles of the MCMC draws, expressed in percentage terms.

In conclusion, the quasi-Bayesian approach provides a more nuanced and flexible clustering of households compared to pre-defined grouping criteria. For example, it reveals the nuanced interactions between sub-categories of income, as well as the differences between income and wealth, in determining the group structure. Importantly, it recovers subgroup heterogeneity missed by k-means, as the latter tends to produce a balanced group structure of similar sizes, which masks across-group differences and thus impedes interpretation.

Further, quasi-Bayesian clustering also has methodological advantages, as evidenced by its realistic uncertainty estimates. While no grouping can capture all household heterogeneity, the results showcase the value of the quasi-Bayesian approach in discovering subgroup patterns. Further investigation of additional covariate relationships with the subgroup assignments could yield further economic insights into cross-sectional differences in household income dynamics.

7 Discussion

This paper develops a general quasi-Bayesian framework for grouped panels where economic agents are partitioned into latent groups such that the parameters of interest are distinct across groups but common within groups. By jointly modeling the latent group structure and the grouped parameters, the proposed framework accounts for potential propagation of estimation errors and thus significantly improves estimation accuracy and coverage relative to existing methods. Importantly, the framework is general enough to encompass a large class of popular estimators and priors, thereby facilitating straightforward adaptation for applied researchers.

Several promising extensions of this work can be explored. First, the current paper focuses on consistency and posterior contraction of the quasi-Bayesian approach. An important extension would involve deriving distributional results such as the Bernsteinvon Mises theorem as in Chernozhukov and Hong (2003). Second, the bootstrap-based learning rate selection is computationally intensive. It would be interesting to extend alternative learning rate calibration methods such as (Fasiolo et al., 2021) to the current framework. Moreover, the bootstrap learning rate selection has close connections with algorithmic stability (Zrnic and Jordan, 2023), and thus exploring the implications of using Bayesian methods in post-selection inference through the lens of stability is an interesting direction for future work.

References

Abbott, B. and Gallipoli, G. (2022). Permanent-income inequality. *Quantitative Economics*, 13(3):1023–1060.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019). Democracy Does Cause Growth. *Journal of Political Economy*, 127(1):47–100.

Aguiar, M. A., Bils, M., and Boar, C. (2020). Who Are the Hand-to-Mouth?

Aguilar, J. and Boot, T. (2022). Grouped heterogeneity in linear panel data models with heterogeneous error variances.

Almgren, M., Gallegos, J.-E., Kramer, J., and Lima, R. (2022). Monetary Policy and Liquidity Constraints: Evidence from the Euro Area. *American Economic Journal: Macroeconomics*, 14(4):309–340.

Alquier, P. (2023). User-friendly introduction to PAC-Bayes bounds.

Ando, T. and Bai, J. (2017). Clustering Huge Number of Financial Time Series: A Panel Data Approach With High-Dimensional Predictors and Factor Structures. *Journal of the American Statistical Association*, 112(519):1182–1198.

Arellano, M., Blundell, R., and Bonhomme, S. (2017). Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework. *Econometrica*, 85(3):693–734.

Arellano, M. and Carrasco, R. (2003). Binary choice panel data models with predetermined variables. *Journal of Econometrics*, 115(1):125–157.

Atchadé, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics*, 45(5).

Auclert, A. (2019). Monetary Policy and the Redistribution Channel. *American Economic Review*, 109(6):2333–2367.

Bils, M. J. (1985). Real Wages over the Business Cycle: Evidence from Panel Data. *Journal of Political Economy*, 93(4):666–689.

Bishop, C. M. and Svensen, M. (2012). Bayesian Hierarchical Mixtures of Experts.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130.

Blundell, R., Pistaferri, L., and Saporta-Eksten, I. (2016). Consumption Inequality and Family Labor Supply. *American Economic Review*, 106(2):387–435.

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, 87(3):699–739.

Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184.

Bonhomme, S. and Weidner, M. (2022). Posterior Average Effects. *Journal of Business & Economic Statistics*, 40(4):1849–1862.

Bouchard-Cote, A., Doucet, A., and Roth, A. (2017). Particle Gibbs Split-Merge Sampling for Bayesian Inference in Mixture Models. *Journal of Machine Learning Research*, 18(28):1–39.

Braxton, J. C., Herkenhoff, K. F., Rothbaum, J. L., and Schmidt, L. (2021). Changing Income Risk across the US Skill Distribution: Evidence from a Generalized Kalman Filter.

Busch, C., Domeij, D., Guvenen, F., and Madera, R. (2022). Skewed Idiosyncratic Income Risk over the Business Cycle: Sources and Insurance. *American Economic Journal: Macroeconomics*, 14(2):207–242.

Catherine, S. (2022). Countercyclical Labor Income Risk and Portfolio Choices over the Life Cycle. *The Review of Financial Studies*, 35(9):4016–4054.

Chen, W., Chen, X., Hsieh, C.-T., and Song, Z. (2019). A Forensic Examination of China's National Accounts. *Brookings Papers on Economic Activity*, 2019(1):77–141.

Cheng, X., Schorfheide, F., and Shao, P. (2019). Clustering for Multi-Dimensional Heterogeneity.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Cloyne, J., Ferreira, C., and Surico, P. (2020). Monetary Policy When Households Have Debt: New Evidence on the Transmission Mechanism. *The Review of Economic Studies*, 87(1):102–129.

Cytrynbaum, M. (2020). Blocked Clusterwise Regression.

Diebolt, J. and Robert, C. P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375.

Dolado, J. J., Motyovszki, G., and Pappa, E. (2021). Monetary Policy and Inequality under Labor Market Frictions and Capital-Skill Complementarity. *American Economic Journal: Macroeconomics*, 13(2):292–332.

Drton, M. and Plummer, M. (2017). A Bayesian Information Criterion for Singular Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):323–380.

Duan, L. L. and Dunson, D. B. (2021). Bayesian Distance Clustering. *Journal of Machine Learning Research*, 22(1):10228–10254.

Duan, L. L., Roy, A., and For the Alzheimer's Disease Neuroimaging Initiative (2023). Spectral Clustering, Bayesian Spanning Forest, and Forest Process. *Journal of the American Statistical Association*, pages 1–14.

Dustmann, C., Schönberg, U., and Stuhler, J. (2017). Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment*. *The Quarterly Journal of Economics*, 132(1):435–483.

Dzemski, A. and Okui, R. (2021). Confidence set for group membership.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast Calibrated Additive Quantile Regression. *Journal of the American Statistical Association*, 116(535):1402–1412.

Fernández-Val, I. and Lee, J. (2013). Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference. *Quantitative Economics*, 4(3):453–481.

Figueiredo, A. (2022). Wage Cyclicality and Labor Market Sorting. *American Economic Review: Insights*, 4(4):425–442.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer.

Fúquene, J., Steel, M., and Rossell, D. (2019). On Choosing Mixture Components via Non-Local Priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(5):809–837.

Gao, C., Van Der Vaart, A. W., and Zhou, H. H. (2020). A general framework for Bayes structured linear models. *The Annals of Statistics*, 48(5).

Gao, L. L., Bien, J., and Witten, D. (2022). Selective Inference for Hierarchical Clustering. *Journal of the American Statistical Association*, 0(0):1–11.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Number 44 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge ; New York.

Ghosh, P., Pati, D., and Bhattacharya, A. (2020). Posterior Contraction Rates for Stochastic Block Models. *Sankhya A*, 82(2):448–476.

Gonçalves, S. and Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495.

Gratton, G., Guiso, L., Michelacci, C., and Morelli, M. (2021). From Weber to Kafka: Political Instability and the Overproduction of Laws. *American Economic Review*, 111(9):2964–3003.

Gregory, V., Menzio, G., and Wiczer, D. G. (2021). The Alpha Beta Gamma of the Labor Market.

Grunwald, P. D. and Mehta, N. A. (2020). Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes. *Journal of Machine Learning Research*.

Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188.

Guvenen, F., Ozkan, S., and Song, J. (2014). The Nature of Countercyclical Income Risk. *Journal of Political Economy*, 122(3):621–660.

Hahn, J. and Kuersteiner, G. (2011). Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects. *Econometric Theory*, 27(6):1152–1191.

Hall, A. R., Han, S., and Boldea, O. (2012). Inference regarding multiple structural changes in linear models with endogenous regressors. *Journal of Econometrics*, 170(2):281–302.

Hamilton, J. D. (1991). A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions. *Journal of Business & Economic Statistics*, 9(1):27–39.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029.

Huang, J. (2021). Group Local Projections. *SSRN Electronic Journal*.

Huang, W., Jin, S., and Su, L. (2020). Identifying Latent Grouped Patterns in Cointegrated Panels. *Econometric Theory*, 36(3):410–456.

Huang, W., Su, L., and Zhuang, Y. (2023). Detecting Unobserved Heterogeneity in Efficient Prices via Classifier-Lasso. *Journal of Business & Economic Statistics*, 41(2):509–522.

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1).

Jiang, S. and Tokdar, S. (2021). Consistent Bayesian Community Detection.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5).

Kapetanios, G. (2008). A bootstrap procedure for panel data sets with many cross-sectional units. *The Econometrics Journal*, 11(2):377–395.

Kaplan, G., Moll, B., and Violante, G. L. (2018). Monetary Policy According to HANK. *American Economic Review*, 108(3):697–743.

Kaplan, G., Violante, G. L., and Weidner, J. (2014). The Wealthy Hand-to-Mouth. *Brookings Papers on Economic Activity*, 2014(1):77–138.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kato, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics*, 41(5):2359–2390.

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

Kim, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics*, 107(1):175–193.

Kim, Y. and Gao, C. (2020). Bayesian Model Selection with Graph Structured Sparsity. *Journal of Machine Learning Research*, 21(109):1–61.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none):354–381.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1):21–59.

Lin, C.-C. and Ng, S. (2012). Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown. *Journal of Econometric Methods*, 1(1):42–55.

Liu, R., Shang, Z., Zhang, Y., and Zhou, Q. (2020). Identification and Estimation in Panel Models with Overspecified Number of Groups. *Journal of Econometrics*, 215(2):574–590.

Low, H., Meghir, C., and Pistaferri, L. (2010). Wage Risk and Employment Risk over the Life Cycle. *American Economic Review*, 100(4):1432–1467.

Lu, Y. and Zhou, H. H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In Dey, D. and Rao, C., editors, *Bayesian Thinking*, volume 25 of *Handbook of Statistics*, pages 459–507. Elsevier.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2):1464–1482.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Mian, A., Rao, K., and Sufi, A. (2013). Household Balance Sheets, Consumption, and the Economic Slump*. *The Quarterly Journal of Economics*, 128(4):1687–1726.

Miller, J. W. (2021). Asymptotic normality, concentration, and coverage of generalized posteriors. *The Journal of Machine Learning Research*, 22(1):168:7598–168:7650.

Miller, J. W. (2023). Consistency of mixture models with a prior on the number of components. *Dependence Modeling*, 11(1):20220150.

Miller, J. W. and Harrison, M. T. (2018). Mixture Models With a Prior on the Number of Components. *Journal of the American Statistical Association*, 113(521):340–356.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, Migration, and Local Employment Elasticities. *American Economic Review*, 108(12):3855–3890.

Müller, U. K. (2013). Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix. *Econometrica*, 81(5):1805–1849.

Narisetty, N. N., Shen, J., and He, X. (2019). Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection. *Journal of the American Statistical Association*, 114(527):1205–1217.

Natarajan, A., De Iorio, M., Heinecke, A., Mayer, E., and Glenn, S. (2023). Cohesion and Repulsion in Bayesian Distance Clustering. *Journal of the American Statistical Association*, pages 1–11.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Neufeld, A. C., Gao, L. L., and Witten, D. M. (2022). Tree-Values: Selective Inference for Regression Trees. *Journal of Machine Learning Research*, 23:1–43.

Newey, W. K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, 59(4):1161.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

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

Norets, A. (2021). Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models. *Econometric Theory*, 37(1):49–81.

Papastamoulis, P. and Iliopoulos, G. (2010). An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Patterson, C. (2023). The Matching Multiplier and the Amplification of Recessions. *American Economic Review*, 113(4):982–1012.

Patton, A. J. and Weller, B. M. (2022). Testing for Unobserved Heterogeneity via k-means Clustering. *Journal of Business and Economic Statistics*, 0(0):1–15.

Petrova, K. (2019). A quasi-Bayesian local likelihood approach to time varying parameter VAR models. *Journal of Econometrics*, 212(1):286–306.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Statistics, probability and game theory*, 30:245–268.

Pollard, D. (1981). Strong Consistency of K-Means Clustering. *The Annals of Statistics*, 9(1):135–140.

Pollard, D. (1982). A Central Limit Theorem for k-Means Clustering. *The Annals of Probability*, 10(4).

Postel-Vinay, F. and Robin, J.-M. (2002). Equilibrium Wage Dispersion with Worker and Employer Heterogeneity. *Econometrica*, 70(6):2295–2350.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):557–574.

Ren, Y., Zhu, X., Lu, X., and Hu, G. (2022). Graphical Assistant Grouped Network Autoregression Model: A Bayesian Nonparametric Recourse. *Journal of Business & Economic Statistics*, pages 1–15.

Richardson, Sylvia. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792.

Rigon, T., Herring, A. H., and Dunson, D. B. (2023). A generalized Bayes framework for probabilistic clustering. *Biometrika*, page asad004.

Robert, C. P., Elvira, V., Tawn, N., and Wu, C. (2018). Accelerating MCMC algorithms. *WIREs Computational Statistics*, 10(5):e1435.

Smith, S. C. (2022). Structural Breaks in Grouped Heterogeneity. *Journal of Business & Economic Statistics*, pages 1–13.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Stephens, M. (2000). Dealing With Label Switching in Mixture Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(4):795–809.

Stock, J. H. and Watson, M. W. (2008). Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression. *Econometrica*, 76(1):155–174.

Storesletten, K., Telmer, C. I., and Yaron, A. (2004). Cyclical Dynamics in Idiosyncratic Labor Market Risk. *Journal of Political Economy*, 112(3):695–717.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying Latent Structures in Panel Data. *Econometrica*, 84(6):2215–2264.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Syring, N. and Martin, R. (2020). Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image. *The Annals of Statistics*, 48(3):1498–1513.

Syring, N. and Martin, R. (2023). Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli*, 29(2).

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111(514):600–620.

van der Vaart, A. W. and van Zanten, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Institute of Mathematical Statistics Collections*, pages 200–222. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008.

Wade, S. and Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis*, 13(2):559–626.

Wang, W., Phillips, P. C. B., and Su, L. (2018). Homogeneity Pursuit in Panel Data Models: Theory and Application. *Journal of Applied Econometrics*, 33(6):797–815.

Wang, W. and Su, L. (2021). Identifying Latent Group Structures in Nonlinear Panels. *Journal of Econometrics*, 220(2):272–295.

Yang, X., Yan, X., and Huang, J. (2019). High-Dimensional Integrative Analysis with Homogeneity and Sparsity Recovery. *Journal of Multivariate Analysis*, 174.

Yano, K. and Kato, K. (2020). On frequentist coverage errors of Bayesian credible sets in moderately high dimensions. *Bernoulli*, 26(1):616–641.

Zhang, B. (2023). Incorporating Prior Knowledge of Latent Group Structure in Panel Data Models.

Zhang, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210.

Zhang, Y., Wang, H. J., and Zhu, Z. (2019). Quantile-regression-based clustering for panel data. *Journal of Econometrics*, 213(1):54–67.

Zrnic, T. and Jordan, M. I. (2023). Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691.

A Posterior Sampling

This section discusses issues in posterior analysis.

A.1 Label switching

When the mixture-of-finite-mixture (MFM) prior is used, the quasi-posterior distribution suffers from the well known "label switching" problem, due to the fact that the posterior (12) is invariant to arbitrary permutations of the group labels.[14] In particular, this causes two issues. First, the MCMC convergence of Algorithm 2 can be slow. This is because the posterior density is by construction multi-modal, and thus the MCMC sampler may fail to explore the parameter space (Jasra et al., 2005). Second, when the MCMC sampler successfully visits all posterior modes, the posterior samples can not be used directly for inference on group-level parameters. This is because the labels in each iteration can be different (Papastamoulis and Iliopoulos, 2010). The two issues together call for a method to first enforce switching of labels so as to improve MCMC convergence, and second undo the label switching so as to facilitate posterior inference. Next I discuss the two issues in turn.

First, to improve MCMC convergence, I augment Algorithm 1 with a random label switching step as in Norets (2021). Specifically, in each iteration $\mathtt{m}$, I randomly select two groups $k, l \in \left\{1, \ldots, G^{(j)}\right\}$ and switch the labels of the two groups, i.e., $\gamma_i^{(j)} \leftarrow l$ if $\gamma_i^{(j)} = k$, and $\gamma_i^{(j)} \leftarrow k$ if $\gamma_i^{(j)} = l$.

Second, to undo the label switching, I follow Marin et al. (2005) to relabel the MCMC output. Specifically, I first select the maximum a posteriori (MAP) estimate based on the MCMC output as the pivot. Then for each MCMC draw, I permute the partitions so as to minimize the distance with respect to the pivotal parameters.[15]

As an illustration, Table A.1 presents the MCMC diagnostics results for the quasi-Bayesian approach under homoskedastic simulation design. As the table shows, the uniform label switching (Norets, 2021) significantly improves the effective sample size (ESS).

[14] For systematic review of the problem, see Frühwirth-Schnatter (2006).

[15] Alternative methods can be used, e.g., (Stephens, 2000) and Papastamoulis and Iliopoulos (2010). However, simulation results show that these alternatives are substantially slower while the improvement in terms of RMSE is limited.

With Uniform Label Switching						
parameters	mean	std	mcse	ess_bulk	ess_tail	rhat
$\beta_1(G1)$	0.8386	0.4948	0.0073	1810.9606	1398.2787	0.9999
$\beta_2(G1)$	0.9540	0.5446	0.0082	3022.6556	2988.7424	1.0004
$\beta_1(G2)$	0.8358	0.4949	0.0075	1991.6056	1437.6703	0.9999
$\beta_2(G2)$	0.9474	0.5399	0.0081	3419.3273	3136.6249	0.9999
$\beta_1(G3)$	0.8281	0.4923	0.0075	2111.8656	1689.8951	1.0005
$\beta_2(G3)$	0.9685	0.5498	0.0080	3912.0066	3189.3999	1.0001
Without Label Switching						
parameters	mean	std	mcse	ess_bulk	ess_tail	rhat
$\beta_1(G1)$	0.5302	0.0510	0.0011	2306.7401	3879.9385	1.0000
$\beta_2(G1)$	1.7061	0.0665	0.0023	858.7201	1521.8845	1.0002
$\beta_1(G2)$	1.5089	0.0624	0.0025	652.8397	1554.5412	1.0011
$\beta_2(G2)$	0.5655	0.0627	0.0026	607.7379	908.0364	1.0008
$\beta_1(G3)$	0.4301	0.1554	0.0081	373.3776	1022.4364	1.0008
$\beta_2(G3)$	0.5729	0.1471	0.0079	356.9726	339.3024	1.0032

Table A.1: MCMC Convergence Diagnostics: Label Switching

B Additional empirical results

This section collects additional results for the empirical application.

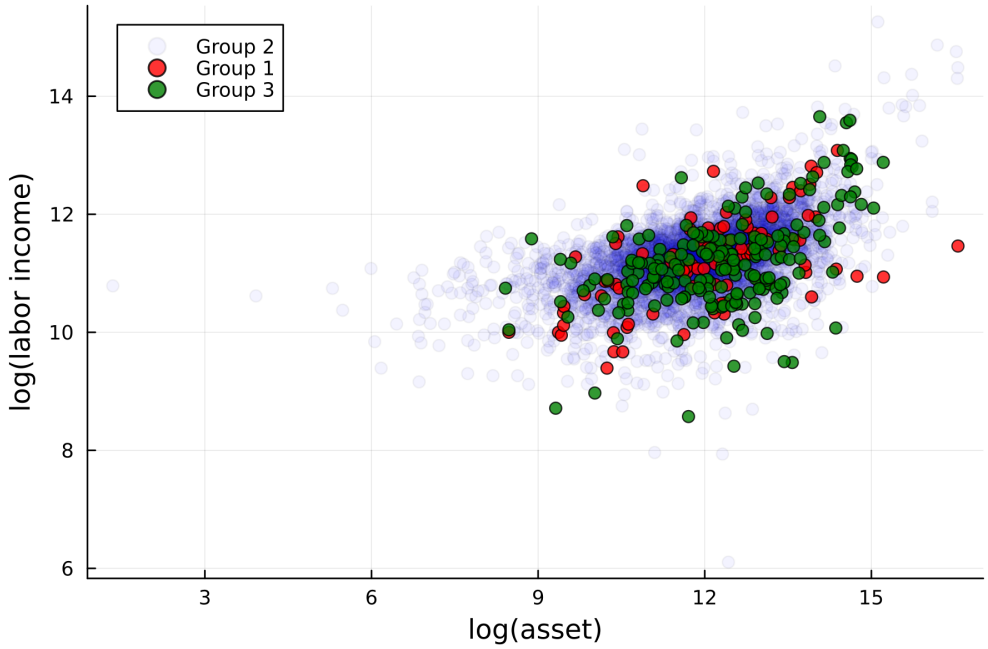


FIGURE 4: GROUP CHARACTERISTICS: ASSETS VS. LABOR INCOME

Note: The figure presents the scatter plot of household total assets and total consumption by identified group. Following definitions in Arellano et al. (2017), total assets are defined by the sum of financial assets, housing values, real estate values, and car values, minus outstanding mortgages. Both variables are transformed in log form.

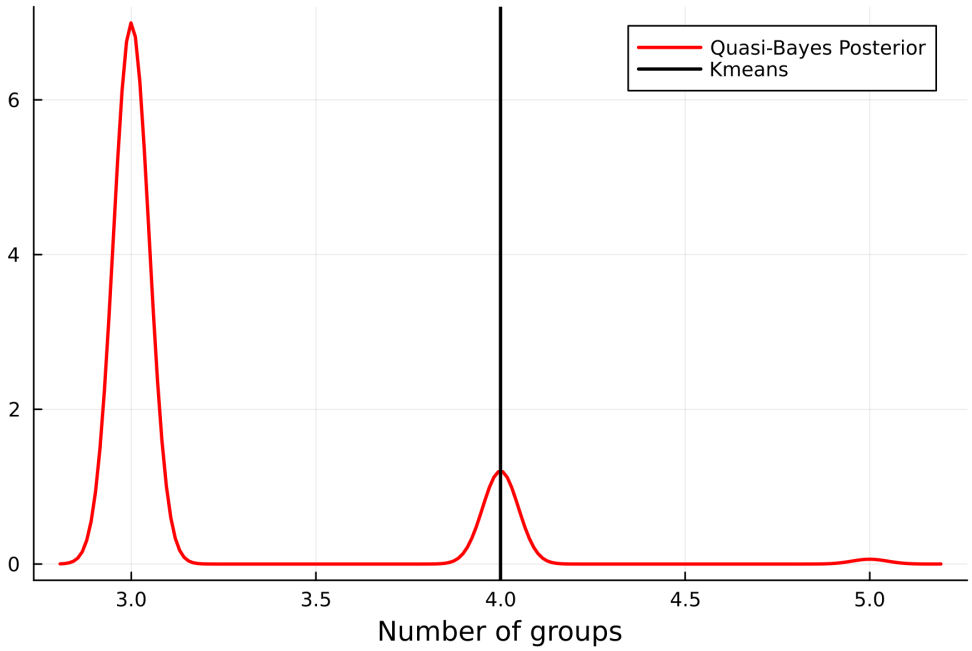


FIGURE 5: GROUP SELECTION

Table B.2: Household characteristics by group (K-means)

	Group 1	Group 2	Group 3	Group 4
Age	44	44	46	44
Education	14	14	14	14
Family Size	3	4	3	3
Total Family Income	123686	123952	110708	125468
Taxable Income (Head and Wife)	117266	118364	103557	116534
Transfer Income (Head and Wife)	2459	2034	2154	5171
Labor Income (Head)	80504	70918	67832	80701
Labor Income (Wife)	30351	27705	29670	26825
Hours Worked (Head)	2263	2191	2245	2246
Hours Worked (Wife)	1444	1431	1474	1179
House Value	254970	233456	242061	271904
Stocks Value	55928	47058	35643	75864
Pensions & Annuities	40637	44122	49785	63978
Cash	28568	32506	26542	34103
Bonds	19086	12487	13222	18781
Other Debt	7036	9284	7007	6564
Financial Assets	190936	182624	138470	208521
Total Assets	351140	360056	298291	368299
Nondurable Consumption	8510	8629	8648	8645
Services Consumption	33519	33076	32825	35533
Total Consumption	38335	38148	37878	40352
Count	232	104	338	88
β_g (%)	-14.95	25.82	1.35	-39.38
$CS(\beta_g)$ (%)	[-15.67,-14.23]	[23.86,27.78]	[0.77,1.93]	[-41.44,-37.31]

C Proofs

C.1 List of lemmas

Lemma 1 (Risk Bound)**.** Under Assumption 2, we have

$$\begin{aligned}\mathbb{E}_0\Pi_{NT}\left(\{\theta\colon L_N(\theta)-L_N(\theta_0)>5\delta_{NT}\}\right)\leq&\mathbb{P}_0\left(\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|>\delta_{NT}\right)\\&+\frac{\exp\left[-2NT\psi\delta_{NT}\right]}{\Pi\left(\{\theta\colon L_N(\theta)-L_N(\theta^0)<\delta_{NT}\}\right)}\ .\end{aligned}\tag{42}$$

Lemma 2 (Uniform Convergence - M-estimation)**.** Under Assumption 4.A, 4.B, 4.C, 4.D, and 4.E, for any $\epsilon>0$, we have as $N,T\to\infty$,

$$\mathbb{P}_0\left(\max_i\sup_{\beta\in\mathcal{B}}\left|\frac{1}{T}\sum_{t=1}^{T}h(w_{it};\beta)-\mathbb{E}h(w_{it};\beta)\right|\geq\epsilon\right)=o(N^{-1})\ .\tag{43}$$

Lemma 3 (Uniform Convergence - GMM)**.** Under Assumption 5.A, 5.B, 5.C, 5.D, 5.E, and 5.F, for any $\epsilon>0$, we have as $N,T\to\infty$,

$$\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|=o_p(N^{-1})\ .\tag{44}$$

Lemma 4 (Prior Concentration - Mixture of Finite Mixture)**.** Let $\Pi(\boldsymbol{\gamma})$ be the prior distribution of $\boldsymbol{\gamma}$ defined in (6). Then for arbitrarily small $\epsilon>0$, we have (up to permutation of the group labels)

$$\Pi\left(\boldsymbol{\gamma}=\boldsymbol{\gamma}^0\right)\geq\exp\left[-CN\ln G^0\right]\tag{45}$$

where $C>0$ is some finite constant.

Lemma 5 (Anderson's Lemma)**.** Suppose $\beta\sim N(0,\Sigma)$ where Σ is a $d\times d$ positive definite matrix, and $\beta^0\in\mathbb{R}^d$. Then for any $\epsilon>0$, we have

$$\mathbb{P}\left(\|\beta-\beta^0\|\leq\epsilon\right)\geq\exp\left[-\frac{1}{2}\beta^{0\top}\Sigma^{-1}\beta^0\right]\mathbb{P}\left(\|\beta\|\leq\epsilon\right)\ .\tag{46}$$

Lemma 6 (Prior mass condition)**.** Under the prior combination of (6), the prior mass condition 2.D holds with $c_{NT}(\epsilon)=\exp\left[-C(N\ln G^0+|\ln\epsilon|)\right]$, for the average distance $d_{MS}(\theta,\tilde{\theta})$ and the Hausdorff distance $d_H(\theta,\tilde{\theta})$. That is,

$$\begin{aligned}\Pi\left(\{\theta\colon d_{MS}(\theta,\theta^0)\leq\epsilon\}\right)&\geq\exp\left[-C(N\ln G^0+|\ln\epsilon|)\right]\\\Pi\left(\{\theta\colon d_H(\theta,\theta^0)\leq\epsilon\}\right)&\geq\exp\left[-C(N\ln G^0+|\ln\epsilon|)\right]\end{aligned}\tag{47}$$

Remark 1. Lemma 1 is the Proposition 6 in Jiang and Tanner (2008). Lemma 2 follows from Lemma S1.2 in Su et al. (2016). Lemma 4 is often referred to as the "prior concentration" result. Similar results are obtained (or directly assumed) in the community detection literature (e.g., Ghosh et al., 2020; Jiang and Tokdar, 2021). Lemma 5 is Lemma 5.2 in van der Vaart and van Zanten (2008).

C.2 Proofs of general results

Proof. (of Theorem 1) For notational simplicity, denote the event $A_\epsilon = \{\theta \colon d(\theta, \theta^0) \leq \epsilon\}$. Notice that the posterior distribution can be written as

$$\begin{aligned}\Pi_{NT}\left(A_\epsilon^c\right) =& \frac{\int \mathbf{1}\left\{A_\epsilon^c\right\} \exp\left[-NT\psi L_{NT}(\theta)\right] \Pi(d\theta)}{\int \exp\left[-NT\psi L_{NT}(\theta)\right] \Pi(d\theta)} \\ =& \frac{\int \mathbf{1}\left\{A_\epsilon^c\right\} \exp\left[-NT\psi (L_{NT}(\theta) - L_N(\theta))\right] \Pi(d\theta)}{\int \exp\left[-NT\psi (L_{NT}(\theta) - L_N(\theta))\right] \Pi(d\theta)}\end{aligned} \tag{48}$$

Therefore, we would like to derive an upper bound for the numerator and a lower bound for the denominator. Consider first the numerator, we can decompose the integrand as

$$\begin{aligned}&\exp\left[-NT\psi\left(L_{NT}(\theta) - L_N(\theta)\right)\right] \mathbf{1}\left\{A_\epsilon^c\right\} \exp\left[-NT\psi \underbrace{\left(L_N(\theta) - L_N(\theta^0)\right)}_{>0}\right] \\ \leq& \exp\left[NT\psi \sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right|\right] \exp\left[-NT\psi \inf_{A_\epsilon^c} \left(L_N(\theta) - L_N(\theta^0)\right)\right] \\ \leq& \exp\left[NT\psi \left(\sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right| - \tilde{\chi}(\epsilon)\right)\right]\end{aligned} \tag{49}$$

where the last inequality comes from the identification condition 1.A.

For the denominator, we have

$$\begin{aligned}&\int \exp\left[-NT\psi(L_{NT}(\theta) - L_N(\theta) + L_N(\theta) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta) \\ \geq& \int \exp\left[-NT\psi \sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right|\right] \exp\left[-NT\psi(L_N(\theta) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta) \\ \geq& \exp\left[-NT\psi \sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right|\right] \int \exp\left[-NT\psi(L_N(\theta) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta)\end{aligned} \tag{50}$$

where we use the fact that $\sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right|$ does not depend on θ. To further bound the right hand side, note that

$$\begin{aligned}&\int \exp\left[-NT\psi(L_N(\theta) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta) \\ \geq& \int \mathbf{1}\left\{L_N(\theta) - L_N(\theta^0) \leq \delta\right\} \exp\left[-NT\psi(L_N(\theta) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta) \\ \geq& \int \mathbf{1}\left\{L_N(\theta) - L_N(\theta^0) \leq \delta\right\} \exp\left[-NT\psi\delta\right] \Pi(\mathrm{d}\theta) \\ =& \exp\left[-NT\psi\delta\right] \Pi\left(\left\{\theta \colon L_N(\theta) - L_N(\theta^0) \leq \delta\right\}\right) \geq \exp\left[-NT\psi\delta\right] \tilde{c}_{NT}(\delta)\end{aligned} \tag{51}$$

where we use the prior mass condition 1.C. Therefore, the denominator is lower bounded by

$$\exp\left[-NT\psi\left(\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|+\delta\right)\right]\tilde{c}_{NT}(\delta)\ . \tag{52}$$

Combined with the upper bound for the numerator, we have

$$\begin{aligned}\Pi_{NT}\left(A_\epsilon^c\right)\leq&\frac{\exp\left[-NT\psi\left(\tilde{\chi}(\epsilon)-2\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|-\delta\right)\right]}{\tilde{c}_{NT}(\delta)}\\ \leq&\frac{\exp\left[-NT\psi\left(\tilde{\chi}(\epsilon)-o(1)-\delta\right)\right]}{\tilde{c}_{NT}(\delta)}\end{aligned} \tag{53}$$

where we use the uniform convergence condition 1.B.

■

Proof. (Posterior contraction rate in Theorem 2)
To start with, notice that by Assumption 2.A, and Lemma 1, we have

$$\begin{aligned}&\mathbb{E}_0\Pi_{NT}\left(\left\{\theta\colon d(\theta,\theta^0)\geq\epsilon_{NT}\right\}\right)\\ \leq&\mathbb{E}_0\Pi\left(\left\{\theta\colon L_N(\theta)-L_N(\theta^0)\geq a(\epsilon_{NT})\right\}\right)\\ \leq&\mathbb{P}_0\left(\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|\geq\frac{1}{5}a(\epsilon_{NT})\right)+\frac{\exp\left[-\frac{2}{5}NT\psi a(\epsilon_{NT})\right]}{\Pi\left(\left\{\theta\colon L_N(\theta)-L_N(\theta^0)<\frac{1}{5}a(\epsilon_{NT})\right\}\right)}\end{aligned} \tag{54}$$

By the uniform convergence condition 2.B, the first term is upper bounded by b_{NT}. For the second term, notice that by the smoothness condition 2.C, we have

$$L_N(\theta)-L_N(\theta^0)\leq\tilde{c}_M d(\theta,\theta^0) \tag{55}$$

and thus

$$\left\{d(\theta,\theta^0)\leq\tilde{c}_M^{-1}\frac{1}{5}a(\epsilon_{NT})\right\}\implies\left\{L_N(\theta)-L_N(\theta^0)\leq\frac{1}{5}a(\epsilon_{NT})\right\} \tag{56}$$

and thus by the prior mass condition 2.D, we have

$$\begin{aligned}&\Pi\left(\left\{\theta\colon L_N(\theta)-L_N(\theta^0)<\frac{1}{5}a(\epsilon_{NT})\right\}\right)\\ \geq&\Pi\left(\left\{d(\theta,\theta^0)\leq\frac{a(\epsilon_{NT})}{5\tilde{c}_M}\right\}\right)\geq c_{NT}(\frac{a(\epsilon_{NT})}{5\tilde{c}_M})\ .\end{aligned} \tag{57}$$

Therefore, we have

$$\mathbb{E}_0\Pi_{NT}\left(\left\{\theta\colon d(\theta,\theta^0)\geq\epsilon_{NT}\right\}\right)\leq b_{NT}+\frac{\exp\left[-\frac{2}{5}NT\psi a(\epsilon_{NT})\right]}{c_{NT}(a(\epsilon_{NT})/5\tilde{c}_M)}\ . \tag{58}$$

■

Proof. (Posterior ratio consistency in Theorem 3)
By definition, the quasi-marginal posterior distribution is

$$\pi_{NT}(G) = \frac{\pi(\mathcal{W}|G)\pi(G)}{\pi(\mathcal{W})} = \frac{\int \exp\left[-NT\psi L_{NT}(\theta_G)\right] \pi(\boldsymbol{\beta}_G)\pi(\boldsymbol{\gamma}_G)\mathrm{d}\boldsymbol{\beta}_G\mathrm{d}\boldsymbol{\gamma}_G\pi(G)}{\pi(\mathcal{W})}$$

where with a slight abuse of notation, $\theta_G = (\boldsymbol{\beta}_G, \boldsymbol{\gamma}_G)$, $\pi(\boldsymbol{\beta}_G)$ is the prior density of $\boldsymbol{\beta}_G = (\beta_1, \ldots, \beta_G)$ and $\pi(\boldsymbol{\gamma}_G)$ is the prior density of $\boldsymbol{\gamma}_G$ such that there are G groups in total. $\mathcal{W}$ represents the entire data set, and $\pi(\mathcal{W})$ is the marginal density of the data. Written more compactly, we have

$$\frac{\pi_{NT}(G)}{\pi_{NT}(G^0)} = \frac{\pi(G)}{\pi(G^0)} \frac{\int \exp\left[-NT\psi L_{NT}(\theta_G)\right] \Pi(\mathrm{d}\theta_G)}{\int \exp\left[-NT\psi L_{NT}(\theta_{G^0})\right] \Pi(\mathrm{d}\theta_{G^0})} \tag{59}$$

and we are now back to the setup for proving Theorem 1.

Consider the ratio of marginal density, we have

$$\begin{aligned}
&\frac{\int \exp\left[-NT\psi L_{NT}(\theta_G)\right] \Pi(\mathrm{d}\theta_G)}{\int \exp\left[-NT\psi L_{NT}(\theta_{G^0})\right] \Pi(\mathrm{d}\theta_{G^0})} \\
=&\frac{\int \exp\left[-NT\psi (L_{NT}(\theta_G) - L_N(\theta_G) + L_N(\theta_G) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta_G)}{\int \exp\left[-NT\psi (L_{NT}(\theta_{G^0}) - L_N(\theta_{G^0}) + L_N(\theta_{G^0}) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta_{G^0})} \\
\leq& \exp\left[2NT\psi \sup_{\theta\in\Theta} \left|L_{NT}(\theta) - L_N(\theta)\right|\right] \\
&\times \frac{\int \exp\left[-NT\psi (L_N(\theta_G) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta_G)}{\int \exp\left[-NT\psi (L_N(\theta_{G^0}) - L_N(\theta^0))\right] \Pi(\mathrm{d}\theta_{G^0})}
\end{aligned} \tag{60}$$

Whenever $G \neq G^0$, denote by $\sigma(\boldsymbol{\gamma}_G)$ a permutation of $\boldsymbol{\gamma}_G$, we have

$$\begin{aligned}
L_N(\theta_G) - L_N(\theta^0) =& \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[l_{iT}(\beta_{\gamma_i}) - l_{iT}(\beta^0_{\gamma_i^0})\right] \\
\geq& \frac{\min_\sigma \{\sigma(\gamma)_i \neq \gamma_i^0\}}{N} \mathbb{E}\left[l_{iT}(\beta_{\sigma(\gamma)_i}) - l_{iT}(\beta^0_{\gamma_i^0})\right] \\
\geq& c\check{\chi}(\epsilon) > 0
\end{aligned} \tag{61}$$

for some $0 < c < 1$ and $\epsilon > 0$. Here, we use (1) the identification condition 3.D, (2) the separability condition 3.C, and (3) the group size condition 3.B. Specifically, by group size condition, $\min_\sigma \{\sigma(\gamma)_i \neq \gamma_i^0\} = O(N)$ and thus $0 < c < 1$; by the separability condition, for those misclassified groups, we have $\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| > \epsilon$ for some $\epsilon > 0$; and finally by the

identification condition, the loss difference is lower bounded by $\check{\chi}(\epsilon) > 0$.

Then following the proof of Theorem 1, the posterior ratio is upper bounded by

$$\frac{\exp\left[-NT\psi\left(c\check{\chi}(\epsilon) - o(1) - \delta\right)\right]}{\tilde{c}_{NT}(\delta)} \,.$$

■

C.3 Proof for M-estimation

The proof of of Theorem 4 is separated in three parts:

Proof. (Theorem 4, mean square convergence)
We would like to verify the three conditions of Theorem 1 under Assumption 4.

Identification. The identification condition 1.A follows directly from the Assumption 4.C. Specifically, we would like to show that

$$\left\{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\geq\epsilon\right\}\implies\left\{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\geq\chi(\epsilon)\right\}. \tag{62}$$

To this extent, assume that $\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\geq\epsilon$. Let $1>\tau>0$ and denote $\tilde{\mathcal{C}}=\{i\colon\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\geq\tau\epsilon\}$ the set of units with "large" parameter estimation errors. Then by construction,

$$\begin{aligned}N\epsilon\leq\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2=&\sum_{i\in\tilde{\mathcal{C}}}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2+\sum_{i\notin\tilde{\mathcal{C}}}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\\ \leq&\left|\tilde{\mathcal{C}}\right|\mathrm{diam}(\mathcal{B})^2+(N-\left|\tilde{\mathcal{C}}\right|)\tau\epsilon\ .\end{aligned} \tag{63}$$

Rearrange terms, we have

$$\left|\tilde{\mathcal{C}}\right|\geq\frac{N(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon} \tag{64}$$

and thus

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\ \geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\ \geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\min_i\inf_{\|\beta-\beta^0_{\gamma_i^0}\|^2\geq\tau\epsilon}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\ \geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\chi(\tau\epsilon)=\frac{\left|\tilde{\mathcal{C}}\right|}{N}\chi(\tau\epsilon)\geq\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon)\ .\end{aligned} \tag{65}$$

Therefore, Condition 1.A holds with

$$\tilde{\chi}(\epsilon)=\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon) \tag{66}$$

for $\chi(\cdot)$ defined in Assumption 4.C and some $\tau \in (0,1)$.

Uniform convergence. The uniform convergence condition 1.A follows directly from Lemma 2. Specifically, since

$$\begin{aligned}&\mathbb{P}_0\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}h(w_{it};\beta_{\gamma_i})-\mathbb{E}h(w_{it};\beta_{\gamma_i})\right)>\epsilon\right\}\\ \leq&\mathbb{P}_0\left\{\max_i\sup_{\beta\in\mathcal{B}}\frac{1}{T}\sum_{t=1}^{T}h(w_{it};\beta_{\gamma_i})-\mathbb{E}h(w_{it};\beta_{\gamma_i})>\epsilon\right\}=o(\frac{1}{N}).\end{aligned} \tag{67}$$

Therefore 1.B holds with $\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|=o(\frac{1}{N})=o(1)$.

Prior mass condition. Condition 1.C holds by Lemma 6. Specifically, we use the smoothness condition 4.D to translate the prior mass in Lemma 6 into the prior mass on the distance metrics. To do so, notice that

$$\begin{aligned}|L_N(\theta)-L_N(\theta_0)|=&\left|\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbb{E}\left[h(w_{it},\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right|\\ \leq&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbb{E}|h(w_{it},\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})|\\ \leq&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbb{E}\left[M(w_{it})\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|\right]\\ =&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|\mathbb{E}M(w_{it})\\ \leq&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|\sup_i\mathbb{E}M(w_{it})\\ \leq&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|c_M^{1/q}\\ \leq&\sqrt{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|^2}c_M^{1/q}\end{aligned} \tag{68}$$

where the first inequality comes from the triangle inequality; the second from Assumption 4.D; the third follows from the assumption that $M(w_{it})$ is non-negative; the second last inequality follows from the moment bounds in Assumption 4.D, and the monotonicity of moments, i.e., $\left[\mathbb{E}\left(|X|^r\right)\right]^{1/r}\leq\left[\mathbb{E}\left(|X|^s\right)\right]^{1/s}$ for $1<r\leq s$; and finally, the last inequality follows from Cauchy-Schwarz inequality, i.e., $\left(\sum_{i=1}^{N}\|x_i\|\right)^2\leq N\left(\sum_{i=1}^{N}\|x_i\|^2\right)$ for any $x_i\in\mathbb{R}^d$.

Therefore, we have for any $\delta>0$,

$$\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|^2\leq\delta^2c_M^{-2/q}\right\}\implies\left\{|L_N(\theta)-L_N(\theta_0)|\leq\delta\right\}. \tag{69}$$

This implies that

$$\begin{aligned}&\Pi\left(\{\theta\colon L_N(\theta)-L_N(\theta_0)\leq\delta\}\right)\\ =&\Pi\left(\{\theta\colon L_N(\theta)-L_N(\theta_0)\leq\delta\}\left|\mathcal{C}=\mathcal{C}^0\right.\right)\Pi\left(\mathcal{C}=\mathcal{C}^0\right)\end{aligned}$$

$$
\begin{aligned}
&\geq \Pi\left(\left\{\theta\colon \frac{1}{N}\sum_{i=1}^{N}\left\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\right\|^2 \leq \delta^2 c_M^{-2/q}\right\}\middle|\mathcal{C}=\mathcal{C}^0\right)\Pi\left(\mathcal{C}=\mathcal{C}^0\right)\\
&\geq \exp\left[-C(N\ln G^0+\left|\ln \delta^2 c_M^{-2/q}\right|)\right] \qquad (70)
\end{aligned}
$$

where the first inequality comes from (69) and the last inequality comes from Lemma 6. Therefore, Condition 1.C holds with

$$
\tilde{c}_{NT}(\delta)=\exp\left[-C(N\ln G^0+\left|\ln \delta^2 c_M^{-2/q}\right|)\right]\ . \qquad (71)
$$

Taken together. Now combining the above results, we have for any $\epsilon>0$,

$$
\begin{aligned}
&\mathbb{E}_0\Pi_{NT}\left(\left\{\theta:\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|>\epsilon\right\}\right)\\
&\leq \exp\left[-NT\psi\left(\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon)-o(1)-\delta-\frac{C\left(N\ln G^0+\left|\ln\delta^2 c_M^{-2/q}\right|\right)}{NT\psi}\right)\right]\ . \qquad (72)
\end{aligned}
$$

It remains to show that the right hand side converges to zero as N,T go to infinity. To do so is equivalent to showing that

$$
\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon)-o(1)-\delta-\frac{C\left(N\ln G^0+\left|\ln\delta^2 c_M^{-2/q}\right|\right)}{NT\psi}>0\ . \qquad (73)
$$

Notice that for any given $\epsilon>0$, the first term above is $O(1)$. Therefore, there exists some constant $\tilde{C}\approx 1$ such that

$$
\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon)-o(1)\geq \tilde{C}\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon)\ . \qquad (74)
$$

Moreover, $\delta>0$ can be arbitrarily positive number, and thus without loss of generality we can select

$$
\delta=\frac{1}{2}\tilde{C}\frac{(1-\tau)\epsilon}{\mathrm{diam}(\mathcal{B})^2-\tau\epsilon}\chi(\tau\epsilon) \qquad (75)
$$

and thus the sum of the first three terms remains $O(1)$. Finally, notice that $\frac{CN\ln G^0}{NT\psi}=O(\frac{1}{T})=o(1)$ for any given ψ. Moreover, for any $\epsilon>0$ and δ chosen above, we have $\left|\ln\delta^2 c_M^{-2/q}\right|=O(1)$, where we use the finite moment assumption 4.D so that $c_M^{-2/q}<\infty$, rendering the last term $O(\frac{1}{NT})$.

To sum up, (73) is dominated by the first term, which is $O(1)$ for any positive $\epsilon>0$. Therefore, the right hand side converges to zero as $N,T\to\infty$, and thus the average

consistency result holds.

■

Proof. (Theorem 4, average misclassification errors $d_M(\theta, \theta^0)$)
As in the previous proof, we proceed by verifying the conditions in Theorem 1.

Identification. We have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\
\geq&\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\
\geq&\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}\min_{i:\,\gamma_i\neq\gamma_i^0}\inf_{\|\beta-\beta^0_{\gamma_i^0}\|\geq\delta_1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(w_{it};\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right]\\
\geq&\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}\chi(\delta_1)\geq\epsilon\chi(\delta_1)
\end{aligned}\tag{76}$$

when we have $\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}\geq\epsilon\right\}$.

Uniform convergence. The uniform convergence does not depend on the specific metric $d(\theta,\theta^0)$ and thus follows directly from Lemma 2.

Prior mass condition. Having established the convergence of $d_{MS}(\cdot,\cdot)$, we can simply apply the same bounds from the previous part:

$$\Pi\left(\{\theta\colon L_N(\theta)-L_N(\theta_0)\leq\delta\}\right)\geq\exp\left[-C(N\ln G^0+\left|\ln\delta c_M^{-1/q}\right|)\right]\tag{77}$$

Notice that the above bound is valid regardless of the metric used.

Taken together, we have

$$\begin{aligned}
&\mathbb{E}_0\Pi\left(\left\{\theta:\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}>\epsilon\right\}\right)\\
\leq&\exp\left[-NT\psi\left(\epsilon\chi(\delta_1)-o(1)-\delta-\frac{C\left(N\ln G^0+\left|\ln\delta c_M^{-1/q}\right|\right)}{NT\psi}\right)\right].
\end{aligned}\tag{78}$$

Then for any given $\epsilon>0$, the right hand side converges to zero as $N,T\to\infty$, and thus the average consistency result holds. ■

Proof. (Theorem 4, Hausdorff distance $d_H(\theta,\theta^0)$)

As before, we proceed by verifying the conditions in Theorem 1.

Identification. We would like to show that

$$\left\{\theta\colon d_H(\theta,\theta^0)\geq\epsilon\right\}\implies\left\{\theta\colon L_N(\theta)-L_N(\theta^0)\geq\chi(\epsilon)\right\} \tag{79}$$

for some function $\chi(\cdot)$. The complication comes from the fact that the Hausdorff distance is defined as the *maximum* of two sub-metrics:

$$d_H(\theta,\theta^0)=\max\left\{\max_{k\in\{1,\dots,G^0\}}\min_{l\in\{1,\dots,G\}}\|\beta_l-\beta_k^0\|,\max_{l\in\{1,\dots,G\}}\min_{k\in 1,\dots,G^0}\|\beta_l-\beta_k^0\|\right\}. \tag{80}$$

Therefore, although it is easy to show that $L_N(\theta)-L_N(\theta^0)$ is lower bounded by some function of the first component, the bound does not depend on the Hausdorff distance itself.

As a workaround, we introduce an intermediate step:

$$\{\theta\colon d_H(\theta,\theta^0)>\epsilon\}\implies\{\theta\colon d_{MS}(\theta,\theta^0)>f(\epsilon)\} \tag{81}$$

where $f(\epsilon)$ is some function of ϵ. Notice that once the above is proved, the result follows directly from the consistency of the average parameter estimation.

Consider the fisrt case. From (133) (in the proof of Lemma 6), we have

$$\min_{l\in\{1,\dots,G\}}\|\beta_l-\beta_k^0\|\leq\|\beta_{\gamma_i}-\beta_k^0\|=\frac{1}{N_k^0}\sum_{i=1}^N 1\{\gamma_i^0=k\}\|\beta_{\gamma_i}-\beta_{\gamma_i^0}^0\|\leq\frac{1}{N_k^0}\sum_{i=1}^N\|\beta_{\gamma_i}-\beta_{\gamma_i^0}^0\| \tag{82}$$

Taking maximum over $k\in\{1,\dots,G\}$ leads to

$$\left\{\max_{k\in\{1,\dots,G\}}\min_{l\in\{1,\dots,G\}}\|\beta_l-\beta_k^0\|>\epsilon\right\}\implies\left\{\frac{1}{N}\sum_{i=1}^N\|\beta_{\gamma_i}-\beta_{\gamma_i^0}^0\|\geq\frac{\epsilon}{C}\right\} \tag{83}$$

for some finite constant C.

Next we start from the average errors

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N\left\|\beta_{\gamma_i}-\beta_{\gamma_i^0}^0\right\|=\frac{1}{N}\sum_{g=1}^G\sum_{i\in\mathcal{C}_g}\left\|\beta_g-\beta_{\gamma_i^0}^0\right\|&\geq\frac{1}{N}\sum_{g=1}^G\sum_{i\in\mathcal{C}_g}\min_{k\in\{1,\dots,G\}}\left\|\beta_g-\beta_k^0\right\|\\&\geq\frac{1}{N}\max_{g\in\{1,\dots,G\}}\left(N_g\min_{k\in\{1,\dots,G\}}\left\|\beta_g-\beta_k^0\right\|\right)\end{aligned} \tag{84}$$

where the first inequality comes from the definition of $\min_k$ and the second from the fact that $\sum_i a_i\geq\max_i a_i$ as long as $a_i\geq 0$ for all i. This last expression illustrates why *if we allow varying number of groups*, it is difficult to establish consistency even at the group-level. For

example, in the extreme case when we allow for micro-clustering, the right hand side reduces to $\frac{1}{N}\max_g \min_k \left\|\beta_g - \beta_k^0\right\|$, which is of order $O(\frac{1}{N})$! In this case the analysis of Hausdorff distance coincide with the *supremum* error rates, which are well known to be slow.

There are two remedies available. First, we can rule out micro-clustering by assumption. For example, we can impose minimal group size constraint in the diffuse prior (6). Second, we may assume that the true number of groups is known. In this case we have

$$\max_{g\in\{1,\ldots,G\}}\left(N_g \min_{k\in\{1,\ldots,G\}}\left\|\beta_g-\beta_k^0\right\|\right) \gtrsim \min_g N_g \max_{g\in\{1,\ldots,G\}}\min_{k\in\{1,\ldots,G\}}\left\|\beta_g-\beta_k^0\right\| . \tag{85}$$

As a result, (81) holds with $f(\epsilon)$ a linear transformation of ϵ.

Uniform convergence. The uniform convergence does not depend on the specific metric $d(\theta,\theta^0)$ and thus follows directly from Lemma 2.

Prior mass condition. Following the previous proof, it suffices to show that the loss function is upper bounded by the Hausdorff distance (scaled by some finite number). Specifically, we have

$$\begin{aligned}
|L_N(\theta)-L_N(\theta_0)| = &\ \left|\tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{E}\left[h(w_{it},\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})\right]\right| \\
\leq &\ \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{E}|h(w_{it},\beta_{\gamma_i})-h(w_{it};\beta^0_{\gamma_i^0})| \\
= &\ \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{E}|h(w_{it},\beta_{\gamma_i})-\min_{\beta\in\mathcal{B}} h(w_{it};\beta)| \\
\leq &\ \tfrac{1}{N}\textstyle\sum_{i=1}^N \min_{\beta\in\mathcal{B}}\mathbb{E}|h(w_{it},\beta_{\gamma_i})-h(w_{it};\beta)| \\
\leq &\ \tfrac{1}{N}\textstyle\sum_{i=1}^N \max_{\tilde\beta\in\mathcal{B}}\min_{\beta\in\mathcal{B}}\mathbb{E}|h(w_{it},\tilde\beta(\beta))-h(w_{it};\beta)| \\
\leq &\ \tfrac{1}{N}\textstyle\sum_{i=1}^N \max_{\tilde\beta\in\mathcal{B}}\min_{\beta\in\mathcal{B}}\mathbb{E}\left[M(w_{it})\left\|\tilde\beta(\beta)-\beta\right\|\right] \\
\leq &\ \max_{\tilde\beta\in\mathcal{B}}\min_{\beta\in\mathcal{B}}\left\|\tilde\beta(\beta)-\beta\right\|\left(\tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{E}M(w_{it})\right) \\
\leq &\ d_H(\theta,\theta^0)\sup_i \mathbb{E}M(w_{it}) \\
\leq &\ d_H(\theta,\theta^0)c_M^{1/q}
\end{aligned} \tag{86}$$

Here, the first inequality comes from the triangle inequality; the second equality follows from the definition of $\beta^0_{\gamma_i^0}$, which uniquely minimizes the population loss function for each individual (Assumption 4.C); the third inequality follows from Jensen's inequality (to the minimum operator); the third last inequality follows by the definition of Hausdorff distance; and the remaining ones are from the smoothness condition 4.D. Therefore, we have for any $\delta>0$,

$$\left\{d_H(\theta,\theta^0)\leq \delta c_M^{-1/q}\right\} \implies \left\{|L_N(\theta)-L_N(\theta_0)|\leq\delta\right\} . \tag{87}$$

and thus the prior mass condition follows, with $\tilde c_{NT}(\delta)=\exp\left[-C(N\ln G^0+\left|\ln \delta c_M^{-1/q}\right|)\right]$, the remainder of the proof is similar to the one for mean square convergence and thus is

omitted. ■

Proof. (of Theorem 5)
To prove the result, we only need to find $a(\epsilon_{NT})$, b_{NT}, $c(\cdot)$ and c_M.

First, by equation (66), we have

$$\left\{\theta\colon d_{MS}(\theta,\theta^0)\geq\epsilon_{NT}\right\}\implies\left\{\theta\colon L_N(\theta)-L_N(\theta^0)\geq\frac{(1-\tau)\epsilon_{NT}}{\operatorname{diam}(\mathcal{B})^2-\tau\epsilon_{NT}}\chi(\tau\epsilon_{NT})\right\}\tag{88}$$

for arbitrarily $\tau\in(0,1)$. Therefore, $a(\epsilon_{NT})=\frac{(1-\tau)\epsilon_{NT}}{\operatorname{diam}(\mathcal{B})^2-\tau\epsilon_{NT}}\chi(\tau\epsilon_{NT})$.

Second, Lemma 6 gives $c(\cdot)$:

$$\Pi\left(\left\{d_{MS}(\theta,\theta^0)\leq\epsilon_{NT}\right\}\right)\geq\exp\left[-C\left(N\ln G^0+\left|\ln\epsilon_{NT}\right|\right)\right]\tag{89}$$

Third, by (69) we have

$$\left\{d_{MS}(\theta,\theta^0)\leq\epsilon_{NT}\right\}\implies\left\{L_N(\theta)-L_N(\theta^0)\leq c_M^{1/q}\sqrt{\epsilon_{NT}}\right\}\tag{90}$$

where c_M is given by Assumption 4.D.

Taken together, we have

$$\begin{aligned}&\mathbb{E}_0\Pi_{NT}\left(\left\{\theta\colon\frac{1}{N}\sum_{i=1}^N\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|\geq\epsilon_{NT}\right\}\right)\\\leq&\mathbb{P}_0\left(\sup_{\theta\in\Theta}\left|L_{NT}(\theta)-L_N(\theta)\right|\geq\frac{1}{5}a(\epsilon_{NT})\right)\\&+\exp\left[CN\ln G^0+C\ln\left|c_M^{1/q}a(\epsilon_{NT})\right|-\frac{2}{5}NT\psi a(\epsilon_{NT})\right]\end{aligned}\tag{91}$$

It remains to show the fastest possible convergence rate of ϵ_{NT}. First notice that $\ln\left|c_M^{1/q}a(\epsilon_{NT})\right|$ is dominated in the second term, and thus we can focus only on $CN\ln G^0$ and $-\frac{2}{5}NT\psi a(\epsilon_{NT})$. In this case, the second term converges to zero as long as $a(\epsilon_{NT})=\tilde{C}T^{-1}$ for some sufficiently large $\tilde{C}>0$. Moreover, such condition also render $b_{NT}=o(1)$, as implied by Lemma 2. Therefore, the fastest possible convergence rate of ϵ_{NT} is T^{-1}.

The result for the average misclassification errors can be simiarly found. In particular, now we have $a(\epsilon_{NT})=\epsilon_{NT}\chi(\delta_1)$ for some (fixed but) small $\delta_1>0$. Moreover, it shares the same $c(\cdot)$ and c_M as the average estimation errors, which gives

$$\mathbb{E}_0\Pi_{NT}\left(\left\{\theta\colon\frac{1}{N}\sum_{i=1}^N\mathbf{1}\left\{\gamma_i\neq\gamma_i^0\right\}\geq\epsilon_{NT}\right\}\right)$$

$$\leq \mathbb{P}_0 \left(\sup_{\theta \in \Theta} \left| L_{NT}(\theta) - L_N(\theta) \right| \geq \epsilon_{NT} \chi(\delta_1) \right)$$
$$+ \exp \left[CN \ln G^0 + C \ln \left| c_M^{1/q} \epsilon_{NT} \chi(\delta_1) \right| - \frac{2}{5} NT \psi \epsilon_{NT} \chi(\delta_1) \right] \tag{92}$$

which as before, shows that the convergence rate is constrained by the prior mass component $CN \ln G^0$. Setting $\epsilon_{NT} = O(T^{-1})$ guarantees that the right hand side converges to zero. ■

C.4 Proofs for GMM-type criterion

The proofs are similar to M-estimation. Here I only show the results for mean-square convergence for brevity.

Proof. (Theorem 6, average parameter errors $d_{MS}(\theta, \theta^0)$)
We would like to verify the conditions of Theorem 1 under Assumption 5.

Identification. First notice that the population objective function is:

$$L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right)^\top \Omega_i \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right) \tag{93}$$

where Ω_i is defined in Assumption 5.E. Since the moment condition is correctly specified under assumption 5.C, we have $\mathbb{E}m(w_{it};\beta^0_{\gamma_i^0}) = 0$, and thus $L_N(\theta) = 0$ (regardless of the choice of Ω_i). Given that, the remaining of the proof is similar to the one for M-estimation. Specifically, by the same notation in (63) and (64), we have

$$|\tilde{\mathcal{C}}| \geq \frac{N(1-\tau)\epsilon}{\operatorname{diam}(\mathcal{B}) - \tau\epsilon} \tag{94}$$

for $\tilde{\mathcal{C}} = \{i\colon \|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2 \geq \tau\epsilon\}$ the set of units with "large" parameter errors. The population loss then becomes

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right)^\top \Omega_i \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right)\\
\geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right)^\top \Omega_i \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}m(w_{it};\beta_{\gamma_i})\right)\\
\geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\rho_{\min,i}\left\|\mathbb{E}m(w_{it};\beta_{\gamma_i})\right\|^2\\
\geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\rho_{\min,i}\min_i\inf_{\|\beta-\beta^0_{\gamma_i^0}\|\geq\tau\epsilon}\left\|\mathbb{E}m(w_{it};\beta_{\gamma_i})\right\|^2\\
\geq&\frac{1}{N}\sum_{i\in\tilde{\mathcal{C}}}\rho_{\min,i}\chi(\tau\epsilon)^2 \geq \frac{|\tilde{\mathcal{C}}|}{N}\underline{\rho}\chi(\tau\epsilon) \geq \frac{(1-\tau)\epsilon}{\operatorname{diam}(\mathcal{B})-\tau\epsilon}\underline{\rho}\chi(\tau\epsilon)^2\ .
\end{aligned} \tag{95}$$

where $\rho_{\min,i}$ is the minimal eigenvalue of Ω_i, and $\underline{\rho} = \min_i \rho_{\min,i}$, which by Assumption 5.E is strictly positive and finite. Therefore, Condition 1.A holds with

$$\tilde{\chi}(\epsilon) = \frac{(1-\tau)\epsilon}{\operatorname{diam}(\mathcal{B})-\tau\epsilon}\underline{\rho}\chi(\tau\epsilon)^2 \tag{96}$$

for $\chi(\cdot)$ defined in Assumption 5.C and some $\tau \in (0,1)$.

Uniform convergence. The uniform convergence condition 1.A follows directly from Lemma 3.

Prior mass condition. Similar to the proof for M-estimation, condition 1.C holds by Lemma 6 and the smoothness condition 5.D. Specifically, we have

$$
\begin{aligned}
|L_N(\theta) - L_N(\theta_0)| = L_N(\theta) = & \tfrac{1}{N}\textstyle\sum_{i=1}^N \left(\tfrac{1}{T}\sum_{t=1}^T \mathbb{E}m(w_{it};\beta_{\gamma_i})\right)^\top \Omega_i \left(\tfrac{1}{T}\sum_{t=1}^T \mathbb{E}m(w_{it};\beta_{\gamma_i})\right) \\
\le & \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\max,i}\left\|\mathbb{E}m(w_{it},\beta_{\gamma_i})\right\|^2 \\
= & \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\max,i}\left\|\mathbb{E}m(w_{it},\beta_{\gamma_i}) - m(w_{it};\beta^0_{\gamma_i^0})\right\|^2 \\
\le & \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\max,i}\left(\mathbb{E}\left\|m(w_{it},\beta_{\gamma_i}) - m(w_{it};\beta^0_{\gamma_i^0})\right\|\right)^2 \\
\le & \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\max,i}\left(\left\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\right\|\mathbb{E}M(w_{it})\right)^2 \\
\le & \tfrac{1}{N}\textstyle\sum_{i=1}^N \left\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\right\|^2 \sup_i \rho_{\max,i} \sup_i (\mathbb{E}M(w_{it}))^2 \\
\le & \tfrac{1}{N}\textstyle\sum_{i=1}^N \left\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\right\|^2 c_M^{2/q}\overline{\rho}
\end{aligned}
\tag{97}
$$

where $\rho_{\max,i}$ is the maximal eigenvalue of Ω_i and $\overline{\rho} = \sup_i \rho_{\max,i}$. The first inequality in the above follows from matrix inequality:

$$
\forall x \in \mathbb{R}^d, \quad x^\top A x \le \rho_{\max} x^\top x = \rho_{\max}\|x\|^2 \tag{98}
$$

for any real symmetrix matrix A. The second inequality follows by applying Jensen's inequality to the expectation operator, i.e., $\|\mathbb{E}x\| \le \mathbb{E}\|x\|$. The third inequality follows from the smoothness condition 5.D, and the last inequality follows from the monotonicity of moments, as in (69).

Therefore, we have for any $\delta > 0$,

$$
\left\{\frac{1}{N}\sum_{i=1}^N \left\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\right\|^2 \le \delta c_M^{-2/q}\overline{\rho}^{-1}\right\} \implies \left\{|L_N(\theta) - L_N(\theta_0)| \le \delta\right\} . \tag{99}
$$

This implies that

$$
\begin{aligned}
& \Pi\left(\{\theta\colon L_N(\theta) - L_N(\theta_0) \le \delta\}\right) \\
=& \Pi\left(\{\theta\colon L_N(\theta) - L_N(\theta_0) \le \delta\} \,\middle|\, \mathcal{C} = \mathcal{C}^0\right)\Pi\left(\mathcal{C} = \mathcal{C}^0\right) \\
\ge& \Pi\left(\left\{\theta\colon \frac{1}{N}\sum_{i=1}^N \left\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\right\| \le \delta c_M^{-2/q}\overline{\rho}^{-1}\right\} \,\middle|\, \mathcal{C} = \mathcal{C}^0\right)\Pi\left(\mathcal{C} = \mathcal{C}^0\right) \\
\ge& \exp\left[-C(N\ln G^0 + \left|\ln \delta c_M^{-2/q}\overline{\rho}^{-1}\right|)\right]
\end{aligned}
\tag{100}
$$

where the first inequality comes from (99) and the last inequality comes from Lemma 6. Therefore, Condition 1.C holds with

$$\tilde{c}_{NT}(\delta) = \exp\left[-C(N \ln G^0 + \left|\ln \delta c_M^{-2/q} \overline{\rho}^{-1}\right|)\right] . \tag{101}$$

Taken together. Now combining the above results, we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{E}_0 \Pi_{NT} \left(\left\{ \theta : \frac{1}{N} \sum_{i=1}^{N} \|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| > \epsilon \right\} \right) \\
&\leq \exp\left[-NT\psi \left(\frac{(1-\tau)\epsilon}{\text{diam}(\mathcal{B}) - \tau\epsilon} \underline{\rho} \chi(\tau\epsilon)^2 - o(1) - \delta - \frac{C\left(N \ln G^0 + \left|\ln \delta c_M^{-2/q} \overline{\rho}^{-1}\right|\right)}{NT\psi} \right) \right] .
\end{aligned} \tag{102}$$

The remainder of the proof is almost identical to M-estimation, by noting that $0 < \underline{\rho} < \overline{\rho} < \infty$. ■

C.5 Proofs of technical lemmas

C.5.1 Proof of Lemma 3

Proof. For notational brevity, the sample loss is written as $L_{NT}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_i\hat{\Omega}_i\hat{m}_i$, and the population loss is $L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} m_i\Omega_i m_i$.

We first consider each summand of the objective function. we have for any $\beta \in \mathcal{B}$

$$
\begin{aligned}
&\left|\hat{m}_i^\top\hat{\Omega}_i\hat{m}_i - m_i\Omega_i m_i\right| \\
=&\left|\hat{m}_i^\top\hat{\Omega}_i\hat{m}_i - \hat{m}_i^\top\Omega_i\hat{m}_i + \hat{m}_i^\top\Omega_i\hat{m}_i - m_i\Omega_i m_i\right| \\
\leq&\left|\hat{m}_i^\top(\hat{\Omega}_i - \Omega_i)\hat{m}_i\right| + \left|\hat{m}_i^\top\Omega_i\hat{m}_i - m_i\Omega_i m_i\right| \\
=&\left|\hat{m}_i^\top(\hat{\Omega}_i - \Omega_i)\hat{m}_i - m_i^\top(\hat{\Omega}_i - \Omega_i)m_i + m_i^\top(\hat{\Omega}_i - \Omega_i)m_i\right| \\
&+ \left|(\hat{m}_i - m_i)^\top\Omega_i(\hat{m}_i - m_i) + 2m_i^\top\Omega_i(\hat{m}_i - m_i)\right| \\
\leq&\left|(\hat{m}_i - m_i)^\top(\hat{\Omega}_i - \Omega_i)(\hat{m}_i - m_i) + 2m_i^\top(\hat{\Omega}_i - \Omega_i)(\hat{m}_i - m_i)\right| \\
&+ \left|m_i^\top(\hat{\Omega}_i - \Omega_i)m_i\right| \\
&+ \left|(\hat{m}_i - m_i)^\top\Omega_i(\hat{m}_i - m_i)\right| + 2\left|m_i^\top\Omega_i(\hat{m}_i - m_i)\right| \\
\leq&\|\hat{m}_i - m_i\|^2\left[\|\Omega_i\| + \|\hat{\Omega}_i - \Omega_i\|\right] \\
&+ 2\|m_i\|\|\hat{m}_i - m_i\|\left[\|\Omega_i\| + \|\hat{\Omega}_i - \Omega_i\|\right] \\
&+ \|m_i\|^2\|\hat{\Omega}_i - \Omega_i\|
\end{aligned}
\tag{103}
$$

where we have repeatedly used the triangle inequality, the matrix identity

$$a^\top Aa - b^\top Ab = (a-b)^\top A(a-b) + 2b^\top A(a-b)$$

and the matrix inequalities $|a'Ab| \leq \|a\| * \|A\| * \|b\|$.

Next, we take the supremum oevr i and $\beta \in \mathcal{B}$. Since both sides are positive, it boils down to taking the supremum over for each individual terms. Let us derive the upper bounds for each elements on the right hand side.

Notice that by Assumption 5 and Lemma S1.2 in Su et al. (2016), we have

$$\sup_i \sup_{\beta\in\mathcal{B}} \left|\frac{1}{T}\sum_{t=1}^{T} m(w_{it};\beta) - \mathbb{E}m(w_{it};\beta)\right| = o_p(N^{-1}) \tag{104}$$

and thus

$$\sup_i \sup_{\beta\in\mathcal{B}} \|\hat{m}_i - m_i\|^2 \leq \sup_i \sup_{\beta\in\mathcal{B}} \|\hat{m}_i - m_i\| \leq o_p(N^{-1}) \tag{105}$$

Moreover, by Assumption 5.E, we have

$$\sup_i \|\Omega_i\| = O_p(1), \quad \sup_i \|\hat{\Omega}_i - \Omega_i\| = o_p(N^{-1}) \tag{106}$$

where the first equality comes from the fact that Ω_i is finite positive definite and thus its norm is bounded by some finite positive constant.

Finally, by assumption 5.D, we have

$$\sup_i \sup_{\beta \in \mathcal{B}} \|\mathbb{E} m(w_{it}; \beta)\| \leq \sup_i M(w_{it}) < c_M^{1/q} < \infty \tag{107}$$

and thus $\sup_i \sup_\beta \|m_i\| = O_p(1)$ and so is $\sup_i \sup_\beta \|m_i\|^2$.

Now combining the upper bounds of all the terms in (103), we have

$$\begin{aligned} & \sup_{\theta \in \Theta} \left| L_{NT}(\theta) - L_N(\theta) \right| \\ \leq \; & \sup_i \sup_\beta \|\hat{m}_i - m_i\|^2 \left[\sup_i \|\Omega_i\| + \sup_i \|\hat{\Omega}_i - \Omega_i\| \right] \\ & + 2 \sup_i \sup_\beta \|m_i\| \sup_i \sup_\beta \|\hat{m}_i - m_i\| \left[\sup_i \|\Omega_i\| + \sup_i \|\hat{\Omega}_i - \Omega_i\| \right] \\ & + \sup_i \sup_\beta \|m_i\|^2 \sup_i \|\hat{\Omega}_i - \Omega_i\| \\ \leq \; & o_p(N^{-1}) O_p(1) = o_p(N^{-1}) \end{aligned} \tag{108}$$

which gives the desired result. ■

C.5.2 Proof of Lemma 4

Proof. Notice that

$$\Pi\left(\mathcal{C}=\mathcal{C}^0\right)=\Pi\left(\mathcal{C}=\mathcal{C}^0|G=G^0\right)\Pi(G=G^0). \tag{109}$$

The prior on selecting the true number of groups is bounded below by

$$\Pi\left(G=G^0\right)=\exp[-\lambda]\frac{\lambda^{G^0}}{G^0!}\geq\exp[-\lambda+G^0\ln\lambda-G^0\ln G^0]\geq\exp[-2G^0\ln G^0]\ . \tag{110}$$

Since G^0 is fixed and finite by Assumption 4.G, it remains to show that $\Pi\left(\mathcal{C}=\mathcal{C}^0|G=G^0\right)\geq\exp\left[-CN\ln G^0\right]$ for some positive constant C.

To this end, we first use the Dirichlet-multinomial conjugacy result (e.g., Miller and Harrison, 2018): for arbitrary $\boldsymbol{\gamma}$ and G,

$$\begin{aligned}&\Pi\left(\gamma_1,\ldots,\gamma_N\middle|G\right)\\ =\ &\int_0^1\cdots\int_0^1\Pi\left(\gamma_1,\ldots,\gamma_N\middle|\eta_1,\ldots,\eta_G,G\right)\mathrm{d}\eta_1\cdots\mathrm{d}\eta_G\\ =\ &\int_0^1\cdots\int_0^1\prod_{i=1}^N\eta_{\gamma_i}\frac{\Gamma(G\alpha)}{\Gamma(\alpha)^G}\prod_{g=1}^G\eta_g^{\alpha-1}\mathrm{d}\eta_1\cdots\mathrm{d}\eta_G\\ =\ &\frac{\Gamma(G\alpha)}{\Gamma(N+G\alpha)}\prod_{g=1}^G\frac{\Gamma\left(\sum_{i=1}^N1\{\gamma_i=g\}+\alpha\right)}{\Gamma(\alpha)}\end{aligned} \tag{111}$$

and thus

$$\begin{aligned}\Pi\left(\mathcal{C}=\mathcal{C}^0\middle|G=G^0\right)=\ &\sum_{\gamma\in[G^0]^N:\,\mathcal{C}(\gamma)=\mathcal{C}^0}\Pi\left(\gamma\middle|G=G^0\right)\\ =\ &(G^0!)\frac{\Gamma\left(G^0\alpha\right)}{\Gamma(N+G^0\alpha)}\prod_{g=1}^{G^0}\frac{\Gamma\left(N_g^0+\alpha\right)}{\Gamma(\alpha)}\end{aligned} \tag{112}$$

Next we use properties of the Gamma function to derive a lower bound on the above. Note that $\alpha\geq1$ and thus $G^0\alpha\geq1$. If $G^0\alpha>2$, we have

$$\begin{aligned}\Gamma(G^0\alpha)\geq\ &\Gamma(\lfloor G^0\alpha\rfloor)\\ \Gamma(\lfloor G^0\alpha\rfloor+N+1)\geq\ &\Gamma(\lfloor G^0\alpha\rfloor+N)\end{aligned} \tag{113}$$

and thus

$$\frac{\Gamma\left(G^0\alpha\right)}{\Gamma(N+G^0\alpha)}\geq\frac{\Gamma(\lfloor G^0\alpha\rfloor)}{\Gamma(\lfloor G^0\alpha\rfloor+N+1)}=\frac{(\lfloor G^0\alpha\rfloor-1)!}{(\lfloor G^0\alpha\rfloor+N)!}\ . \tag{114}$$

One the other hand, if $1\leq G^0\alpha\leq2$ we have

$$\frac{\Gamma\left(G^0\alpha\right)}{\Gamma(N+G^0\alpha)}\geq\frac{(\min_x\Gamma(x))\cdot1}{\Gamma(\lfloor G^0\alpha\rfloor+N+1)}=\frac{(\min_x\Gamma(x))\cdot0!}{(\lfloor G^0\alpha\rfloor+N)!}=\frac{(\min_x\Gamma(x))\cdot(\lfloor G^0\alpha\rfloor-1)!}{(\lfloor G^0\alpha\rfloor+N)!}\ . \tag{115}$$

Similarly, if $\alpha > 2$, we have

$$\frac{\Gamma(\alpha + N_g^0)}{\Gamma(\alpha)} \geq \frac{\Gamma(\lfloor\alpha\rfloor + N_g^0)}{\Gamma(\lfloor\alpha\rfloor + 1)} = \frac{\big(\lfloor\alpha\rfloor + N_g^0 - 1\big)!}{(\lfloor\alpha\rfloor)!} , \tag{116}$$

while if $1 \leq \alpha \leq 2$, we have

$$\frac{\Gamma(\alpha + N_g^0)}{\Gamma(\alpha)} \geq \frac{\Gamma(\lfloor\alpha\rfloor + N_g^0)}{1} = \frac{\big(\lfloor\alpha\rfloor + N_g^0 - 1\big)!}{(\lfloor\alpha\rfloor)!} . \tag{117}$$

Combining the above results, we have for all $\alpha \geq 1$,

$$\frac{\Gamma\left(G^0\alpha\right)}{\Gamma(N + G^0\alpha)} \geq \tilde{C}_1 \frac{(\lfloor G^0\alpha\rfloor - 1)!}{(\lfloor G^0\alpha\rfloor + N)!}, \quad \frac{\Gamma(\alpha + N_g^0)}{\Gamma(\alpha)} \geq \frac{\big(\lfloor\alpha\rfloor + N_g^0 - 1\big)!}{(\lfloor\alpha\rfloor)!} . \tag{118}$$

Plug this into (112), we obtain

$$\Pi\left(\mathcal{C} = \mathcal{C}^0 \middle| G = G^0\right) \geq (G^0)!\tilde{C}_1 \frac{(\lfloor G^0\alpha\rfloor - 1)!}{(\lfloor G^0\alpha\rfloor + N)!} \prod_{g=1}^{G^0} \frac{\big(\lfloor\alpha\rfloor + N_g^0 - 1\big)!}{(\lfloor\alpha\rfloor)!} . \tag{119}$$

To bound the RHS, we use the fact that multinomial coefficients are maximized when group sizes are equal. Specifically, rewrite the cumulative product in the above as

$$\begin{aligned}
\textstyle\prod_{g=1}^{G^0} \frac{\left(\lfloor\alpha\rfloor + N_g^0 - 1\right)!}{(\lfloor\alpha\rfloor)!} &= \left(\frac{[(\lfloor\alpha\rfloor)!]^{G^0}}{\left(\lfloor\alpha\rfloor + N_1^0 - 1\right)! \cdots \left(\lfloor\alpha\rfloor + N_{G^0}^0 - 1\right)!}\right)^{-1} \\
&= \left(\frac{\left[\sum_{g=1}^{G^0}\left(\lfloor\alpha\rfloor + N_g^0 - 1\right)\right]!}{[(\lfloor\alpha\rfloor)!]^{G^0}}\right)\left(\frac{\left[\sum_{g=1}^{G^0}\left(\lfloor\alpha\rfloor + N_g^0 - 1\right)\right]!}{\left(\lfloor\alpha\rfloor + N_1^0 - 1\right)! \cdots \left(\lfloor\alpha\rfloor + N_{G^0}^0 - 1\right)!}\right)^{-1} \\
&\geq \left(\frac{\left[G^0\lfloor\alpha\rfloor + N - G^0\right]!}{[(\lfloor\alpha\rfloor)!]^{G^0}}\right)\left(\frac{\left[G^0\lfloor\alpha\rfloor + N - G^0\right]!}{[(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor - 1)!]^{G^0 - r}[(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor)!]^r}\right)^{-1} \\
&= \left\{(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor - 1) \cdots (\lfloor\alpha\rfloor + 1)\right\}^{G^0} (\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor)^r
\end{aligned} \tag{120}$$

where $r = N - \lfloor N/G^0\rfloor G^0$. Plug this into (119), we have

$$\begin{aligned}
&\Pi\left(\mathcal{C} = \mathcal{C}^0 \middle| G = G^0\right) \\
\geq\ & (G^0)!C_1 \frac{\left\{\left(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor - 1\right) \cdots (\lfloor\alpha\rfloor + 1)\right\}^{G^0} \left(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor\right)^r}{(\lfloor G^0\alpha\rfloor) \cdots (\lfloor G^0\alpha\rfloor + N)} \\
=\ & (G^0)!\tilde{C}_1 \frac{\left\{\left(\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor - 1\right) \cdots (\lfloor\alpha\rfloor + 1)\right\}^{G^0}}{\{(\lfloor G^0\alpha\rfloor + 2G^0 \cdot 1) \cdots (\lfloor G^0\alpha\rfloor + 2G^0 \cdot (\lfloor N/G^0\rfloor - 1))\}^{G^0}} \times \\
& \qquad \frac{\left\{\left(\lfloor G^0\alpha\rfloor + 2G^0 \cdot 1\right) \cdots \left(\lfloor G^0\alpha\rfloor + 2G^0 \cdot (\lfloor N/G^0\rfloor - 1)\right)\right\}^{G^0}}{(\lfloor G^0\alpha\rfloor) \cdots (\lfloor G^0\alpha\rfloor + N)} (\lfloor\alpha\rfloor + \lfloor N/G^0\rfloor)^r
\end{aligned} \tag{121}$$

For the first ratio, rewrite it as

$$\begin{aligned} & \frac{\left\{\left(\lfloor\alpha\rfloor+\left\lfloor N/G^0\right\rfloor-1\right)\cdots(\lfloor\alpha\rfloor+1)\right\}^{G^0}}{\left\{(\lfloor G^0\alpha\rfloor+2G^0\cdot 1)\cdots(\lfloor G^0\alpha\rfloor+2G^0\cdot(\lfloor N/G^0\rfloor-1))\right\}^{G^0}} \\ = & \left\{\prod_{j=1}^{\left\lfloor N/G^0\right\rfloor-1}\frac{\lfloor\alpha\rfloor+j}{2G^0(\lfloor G^0\alpha\rfloor/(2G^0)+j)}\right\}^{G^0} \geq \left(\frac{1}{2G^0}\right)^{\left\lfloor N/G^0\right\rfloor G^0-G^0} \end{aligned} \tag{122}$$

where the last inequality comes from the fact that

$$\frac{\lfloor\alpha\rfloor}{\lfloor G^0\alpha\rfloor} \geq \frac{\lfloor\alpha\rfloor}{(\lfloor\alpha\rfloor+1)G^0} \geq \frac{1}{2G^0} \implies \lfloor\alpha\rfloor \geq \frac{\lfloor G^0\alpha\rfloor}{2G^0} \ . \tag{123}$$

To analyze the second ratio, we first notice that there are

$$G^0 \times \left(\left\lfloor N/G^0\right\rfloor-1\right) = G^0\left\lfloor N/G^0\right\rfloor - G^0 \leq N$$

terms in the numerator. Collect the first $G^0\left\lfloor N/G^0\right\rfloor - G^0$ terms in the denominator and rewrite the ratio as

$$\left(\prod_{j=1}^{G^0}\frac{\lfloor G^0\alpha\rfloor+2G^0\cdot 1}{\lfloor G^0\alpha\rfloor+j}\right)\times\ldots\times\left(\prod_{j=1}^{G^0}\frac{\lfloor G^0\alpha\rfloor+2G^0\cdot(\lfloor N/G^0\rfloor-1)}{\lfloor G^0\alpha\rfloor+G^0\cdot(\lfloor N/G^0\rfloor-2)+j}\right)\geq 1 \ . \tag{124}$$

The remaining terms in the denominator are

$$\frac{1}{\lfloor G^0\alpha\rfloor+G^0\left\lfloor N/G^0\right\rfloor-G^0+1}\times\ldots\times\frac{1}{\lfloor G^0\alpha\rfloor+N} \geq \left(\frac{1}{\lfloor G^0\alpha\rfloor+N}\right)^{r+G^0} \ .$$

Combining all these results, we have

$$\begin{aligned} & \Pi\left(\mathcal{C}=\mathcal{C}^0\middle|G=G^0\right) \\ \geq & (G^0)!\tilde{C}_1\left(\tfrac{1}{2G^0}\right)^{\left\lfloor N/G^0\right\rfloor G^0-G^0}\times\left(\tfrac{1}{\lfloor G^0\alpha\rfloor+N}\right)^{r+G^0}\times(\lfloor\alpha\rfloor+\lfloor N/G^0\rfloor)^r \\ = & (G^0)!\tilde{C}_1\left(\tfrac{1}{2G^0}\right)^{\left\lfloor N/G^0\right\rfloor G^0}\times\left(\tfrac{2G^0}{\lfloor G^0\alpha\rfloor+N}\right)^{G^0}\times\left(\tfrac{\lfloor\alpha\rfloor+\left\lfloor N/G^0\right\rfloor}{\lfloor G^0\alpha\rfloor+N}\right)^r \end{aligned} \tag{125}$$

As is clear from the above expression, the lower bound is driven by the term $(2G^0)^{-\left\lfloor N/G^0\right\rfloor G^0}$. Let $C_1>0$ be some constant, we can express the lower bound as

$$\Pi\left(\mathcal{C}=\mathcal{C}^0\middle|G=G^0\right) \geq \exp\left[-C_1 N\ln G^0\right] \ . \tag{126}$$

■

C.5.3 Proof of Lemma 6

Proof. The results follow by Lemma 4 and Lemma 5. First notice that regardless of the distance metrics used, from Lemma 4 we have

$$\begin{aligned}\Pi\left(d(\theta,\theta^0)\leq\epsilon\right)=&\Pi\left(d(\theta,\theta^0)\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right)\Pi\left(\mathcal{C}=\mathcal{C}^0\right)\\ \geq&\Pi\left(d(\theta,\theta^0)\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right)\exp\left[-C_1N\ln G^0\right]\end{aligned}\tag{127}$$

where $C_1>0$ is some finite constant. Such decomposition greatly ease the proof. In particular, suppose we have recovered the true group structure. Then if $\|\beta_g-\beta_g^0\|^2\leq\epsilon$ for all the recovered groups, we have

$$\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2=\frac{1}{N}\sum_{g=1}^{G^0}\sum_{i\in\mathcal{C}_g^0}\|\beta_g-\beta_g^0\|^2\leq\frac{1}{N}\sum_{g=1}^{G^0}\sum_{i\in\mathcal{C}_g^0}\epsilon=\frac{\sum_{g=1}^{G^0}N_g^0}{N}\epsilon=\epsilon\ .\tag{128}$$

That is,

$$\left\{\forall g,\|\beta_g-\beta_g^0\|^2\leq\epsilon\right\}\implies\left\{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\leq\epsilon\right\}\tag{129}$$

and thus

$$\Pi\left(\forall g,\|\beta_g-\beta_g^0\|^2\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right)\leq\Pi\left(\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right).\tag{130}$$

Moreover, since the event $\left\{\forall g,\|\beta_g-\beta_g^0\|^2\leq\epsilon\right\}$ does not involve the group structure, we can simply write

$$\Pi\left(\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i}-\beta^0_{\gamma_i^0}\|^2\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right)\geq\Pi\left(\forall g,\|\beta_g-\beta_g^0\|^2\leq\epsilon\right)\ .\tag{131}$$

Plug this back in (127) gives

$$\Pi\left(\left\{\theta\colon d_{MS}(\theta,\theta^0)\leq\epsilon\right\}\right)\geq\Pi\left(\forall g,\|\beta_g-\beta_g^0\|^2\leq\epsilon\right)\exp\left[-C_1N\ln G^0\right]\ .\tag{132}$$

Similar results hold for the Hausdorff distance. Specifically, for any $\epsilon>0$, and an arbitrary true group k

$$\min_{l\in\{1,\ldots,G\}}\|\beta_l-\beta_k^0\|\leq\|\beta_{\gamma_i}-\beta_k^0\|$$

$$
\begin{aligned}
=&\frac{1}{N_k^0}\sum_{i=1}^{N} 1\{\gamma_i^0 = k\}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| \leq \frac{1}{N_k^0}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| \\
\leq&\frac{1}{N_k^0}\sqrt{N\left(\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2\right)} = \frac{N}{N_k^0}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2}
\end{aligned} \tag{133}
$$

where the first inequality is obtained by construction; the second inequality is trivially true; the third inequality follows from Cauchy-Schwarz inequality. Therefore, we have

$$
\begin{aligned}
\max_{k\in\{1,\ldots,G^0\}}\min_{l\in\{1,\ldots,G\}}\|\beta_l - \beta_k^0\| \leq & \max_{k\in\{1,\ldots,G^0\}}\tfrac{1}{N_k^0}\textstyle\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| \\
\leq & \left(\max_{k\in\{1,\ldots,G^0\}}\tfrac{N}{N_k^0}\right)\sqrt{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2}
\end{aligned} \tag{134}
$$

Now by Assumption 4.G, we have

$$
0 < \max_{k\in\{1,\ldots,G^0\}}\frac{N}{N_k^0} = \frac{N}{\min_{k\in\{1,\ldots,G^0\}}N_k^0} \leq \frac{N}{N^0_{\min}} < C_1 \tag{135}
$$

for some finite positive constant C_1. We can write

$$
\max_{k\in\{1,\ldots,G^0\}}\min_{l\in\{1,\ldots,G\}}\|\beta_l - \beta_k^0\| \leq C_1\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2} \tag{136}
$$

The other half of the Hausdorff distance can be bounded similarly, by noting that

$$
\min_{k\in\{1,\ldots,G^0\}}\|\beta_l - \beta_k^0\| \leq \|\beta_l - \beta^0_{\gamma_i^0}\| = \frac{1}{N_l^0}\sum_{i=1}^{N}1\{\gamma_i^0 = l\}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| \leq \frac{1}{N_l^0}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\| \tag{137}
$$

and the rest follows. The key observation is that here $\gamma_i = \gamma_i^0$ under the true group structure. Taken together, the Hausdorff distance (under true grouping) is bounded by

$$
d_H(\theta, \theta^0) \leq C\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2} \tag{138}
$$

for some finite constant C. Therefore, we also have

$$
\left\{\forall g, \|\beta_g - \beta_g^0\|^2 \leq \epsilon\right\} \implies \left\{\frac{1}{N}\sum_{i=1}^{N}\|\beta_{\gamma_i} - \beta^0_{\gamma_i^0}\|^2 \leq \epsilon\right\} \implies \left\{d_H(\theta, \theta^0) \leq C\sqrt{\epsilon}\right\} \tag{139}
$$

which implies

$$\begin{aligned}\Pi\left(d_H(\theta,\theta^0)\leq\epsilon\right)=&\Pi\left(d_H(\theta,\theta^0)\leq\epsilon\middle|\mathcal{C}=\mathcal{C}^0\right)\Pi\left(\mathcal{C}=\mathcal{C}^0\right)\\ \geq&\Pi\left(\forall g,\|\beta_g-\beta_g^0\|^2\leq\tilde{C}\epsilon^2\right)\exp\left[-C_1N\ln G^0\right]\end{aligned}\tag{140}$$

where $\tilde{C}=1/C^2$.

As a result, proving the lemma is equivalent to proving a lower bound for $\Pi\left(\forall g,\|\beta_g-\beta_g^0\|\leq\epsilon\right)$ for any positive $\epsilon>0$. Notice that the additional constant term $\tilde{C}$ in the Hausdorff distance does not require a separate proof since $\tilde{C}$ is a finite positive constant.

Now, for any arbitrarily small $\epsilon>0$, we have

$$\begin{aligned}&\Pi\left(\forall g,\|\beta_g-\beta_g^0\|\leq\epsilon\right)\\ =&\prod_{g=1}^{G^0}\Pi\left(\|\beta_g-\beta_g^0\|\leq\epsilon\right)\\ \geq&\prod_{g=1}^{G^0}\left(\exp\left[-\tfrac{1}{2}\beta_g^{0\top}\Sigma^{-1}\beta_g^0\right]\Pi\left(\|\beta_g\|\leq\epsilon\right)\right)\ .\end{aligned}\tag{141}$$

where the first inequality comes from probability algebra and the fact that each β_g is drawn independently from $N(0,\Sigma)$, and the second inequality comes from Anderson's Lemma (Lemma 5). To bound the RHS, we use the decomposition of $\Sigma=R\Lambda R^\top$ such that $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_d)$ is a diagonal matrix of eigenvalues and R is an orthogonal matrix $R^\top R=RR^\top=I$. Then we have, for the first term,

$$\begin{aligned}&\exp\left[-\tfrac{1}{2}\beta_g^{0\top}\Sigma^{-1}\beta_g^0\right]\\ =&\exp\left[-\tfrac{1}{2}\left(R^\top\beta_g^0\right)^\top\Lambda^{-1}\left(R^\top\beta_g^0\right)\right]\\ =&\exp\left[-\tfrac{1}{2}\sum_{j=1}^d\xi_{gj}^2\lambda_j^{-1}\right]\end{aligned}\tag{142}$$

where $\xi_g=R^\top\beta_g^0$. For the second term, we have

$$\begin{aligned}&\Pi\left(\|\beta_g\|\leq\epsilon\right)=\Pi\left(\beta_g^\top\beta_g\leq\epsilon^2\right)\\ =&\Pi\left(\left(\Sigma^{-1/2}\beta_g\right)^\top\Sigma\left(\Sigma^{-1/2}\beta_g\right)\leq\epsilon^2\right)\\ =&\Pi\left(\left(R^\top\Sigma^{-1/2}\beta_g\right)^\top R^\top\Sigma R\left(R^\top\Sigma^{-1/2}\beta_g\right)\leq\epsilon^2\right)\\ =&\Pi\left(\tilde{\xi}^\top\Lambda\tilde{\xi}\leq\epsilon^2\right)=\Pi\left(\sum_{j=1}^d\lambda_j\tilde{\xi}_j^2\leq\epsilon^2\right)\\ \geq&\Pi\left(\forall j,\tilde{\xi}_j^2\leq\tfrac{\epsilon^2}{d\lambda_j}\right)=\prod_{j=1}^d\Pi\left(\tilde{\xi}_j^2\leq\tfrac{\epsilon^2}{d\lambda_j}\right)\\ =&\prod_{j=1}^d\Pi\left(|\tilde{\xi}_j|\leq\tfrac{\epsilon}{\sqrt{d\lambda_j}}\right)=\prod_{j=1}^d\left[2\Phi\left(\tfrac{\epsilon}{\sqrt{d\lambda_j}}\right)-1\right]\end{aligned}\ ,\tag{143}$$

where $\tilde{\xi}=R^\top\Sigma^{-1/2}\beta_g\sim N(0,I_d)$ and hence entries $\tilde{\xi}_j$ are independent standard normal random variables, with cdf $\Phi(\cdot)$.

Combine (142) and (143), we have

$$\begin{aligned} & \Pi\left(\forall g, \|\beta_g - \beta_g^0\| \leq \epsilon\right) \\ \geq\ & \exp\left[-\tfrac{1}{2}\sum_{g=1}^{G^0}\sum_{j=1}^{d}\left(\xi_{gj}^2\lambda_j^{-1} - 2\ln\left[2\Phi\left(\tfrac{\epsilon}{\sqrt{d\lambda_j}}\right) - 1\right]\right)\right] \end{aligned} \tag{144}$$

To derive a lower bound of the above is equivalent to derive an upper bound of the summands.

To bound the logarithm term, notice that for a sufficiently small $\epsilon > 0$, we have

$$\begin{aligned} & \xi_{gj}^2\lambda_j^{-1} - 2\ln\left[2\Phi\left(\tfrac{\epsilon}{\sqrt{d\lambda_j}}\right) - 1\right] \\ \leq\ & \xi_{gj}^2\lambda_j^{-1} + 2 + 2\left|\ln\left(\tfrac{\epsilon}{\sqrt{d\lambda_j}}\right)\right| \\ \leq\ & \xi_{gj}^2\left(\lambda_{j,\min}\right)^{-1} + 2 + 2\left|\ln\left(\tfrac{\epsilon}{\sqrt{d(\lambda_{j,\max})}}\right)\right| \\ \leq\ & \xi_{gj}^2\left(\lambda_{j,\min}\right)^{-1} + 2 + 2|\ln\epsilon| + |\ln(d\lambda_{j,\max})| \end{aligned} \tag{145}$$

and thus

$$\begin{aligned} & \Pi\left(\forall g, \|\beta_g - \beta_g^0\| \leq \epsilon\right) \\ \geq\ & \exp\left[-\tfrac{1}{2}\left(\lambda_{j,\min}\right)^{-1}\sum_{g=1}^{G^0}\sum_{j=1}^{d}\xi_{gj}^2 - G^0 d - \sum_{g=1}^{G^0}\sum_{j=1}^{d}|\ln\epsilon| - \tfrac{1}{2}G^0 d|\ln(d\lambda_{j,\max})|\right] \\ \geq\ & \exp\left[-G^0 d\left(\tfrac{1}{2}\left(\lambda_{j,\min}\right)^{-1}C_\xi + 1 + \tfrac{1}{2}|\ln(d\lambda_{j,\max})| + |\ln\epsilon|\right)\right] \end{aligned} \tag{146}$$

where $C_\xi = \max_{g,j}\xi_{gj}^2$, which is a finite constant, by the compactness assumption 4.A. Moreover, since Σ is positive definite, the minimal and maximal eigenvalues are also finite, i.e., $\lambda_{j,\min},\ \lambda_{j,\max} < \infty$. As a consequence, the above lower bound is driven by ϵ. That is, there exists some positive constant C_3 such that

$$\Pi\left(\forall g, \|\beta_g - \beta_g^0\| \leq \epsilon\right) \geq \exp\left[-C_3|\ln\epsilon|\right]\ . \tag{147}$$

Combine (147) and (127), we have for any ϵ

$$\begin{aligned} \Pi\left(d(\theta,\theta^0) \leq \epsilon\right) \geq\ & \exp\left[-C_1 N\ln G^0 - C_3|\ln\epsilon|\right] \\ \geq\ & \exp\left[-C\left(N\ln G^0 + |\ln\epsilon|\right)\right] \end{aligned} \tag{148}$$

for some constant $C > 0$. Therefore, Condition 2.D is satisfied with

$$c_{NT}(\epsilon) = \exp\left[-C\left(N\ln G^0 + |\ln\epsilon|\right)\right] \tag{149}$$

and the results follow. ■