Nonlinear models with latent grouping and grouped fixed effect

Zhonghui Zhang[1]

[1]Department of Economics, University of Connecticut

January 10, 2020

Abstract

We extend the linear panel data model with grouped fixed effect and unknown group membership in Bonhomme and Manresa (2015) to nonlinear. Unlike Bonhomme et al. (2017, Working Paper), we assume the unobservable heterogeneities are from the mixture of a certain number of group-specific distributions. Our method provides information on higher moments than k-means which only offer the first moment. We also give researcher an option to relax the restrictions on the group-specific distribution and leave it to be estimated nonparametrically, and provide a guide on selecting bandwidth. We show that the coefficient of interest covariate in the grouping object or "moment" in Bonhomme et al. (2017, Working Paper) is a "nuisance" concerning grouping. This implies, in the grouping step, we can choose some particular value for the latent common parameter just for grouping purpose. Lastly, we study the distribution of unobservable heterogeneities given different value of common parameter and T through Monte Carlo simulation.

Keywords: Nonlinear panel data model, group fixed effect, mixture model

JEL Codes: C13 C23 C38 C63

1 Introduction

Panel data is widely used in empirical economics study because it has a time dimension to identify the unobservable individual-specific heterogeneities, which may be either time-invariant or time-varying, and may be correlated with the interest covariates. According to literature, two conventional approaches can be applied to deal with those individual heterogeneities: random effect and fixed effect approach. Random effect approach imposes a specific restriction on the unobservables. For example, individual heterogeneities are independent with the interest covariates, or they follow some distribution which can integrate the interest covariates out. Either one cannot be verified in practice. Fixed effect approach considers individual heterogeneities as unknown variables to be estimated, thus can release the assumptions on the unobservables. However, the price researcher has to pay is more data of the time dimension is required. Since the individual-specific parameter is estimated by the time-series data of the specific individual. Once T grows comparably slower than N, those individual-specific parameters suffer from estimation error and then contaminate the inference of common parameters, especially in the nonlinear and dynamic setup. This is the well-known incidental parameter problem firstly noted by Neyman and Scott (1948). Even if T grow at the same speed as N, the coefficient of interest covariate is still biased, but bias-correction can be used to address the issue, see, e.g., Lancaster (2000), Hahn and Newey (2004).

One method to reduce the incidental parameter problem in fixed effect approach is to assume the individuals may be grouped at different levels and assume the individuals within the same group are identical. By grouping, the number of parameters to be estimated is reduced from N to G. The observations for estimating each group-specific parameter are increased from T to $G_N * T$ (G_N is the number of individuals within the same group). Bester and Hansen (2016) show that grouped fixed effect estimates suffer from two sources of bias. One is from the incidental parameter problem which is increasing in the number of groups G. Individual fixed effect estimates is an extreme example which considers each individual is a group. Another source of bias is from ignoring the heterogeneities within a group. For example, pooling estimates is another extreme which omits all individual heterogeneities. They also provide the conditions so that the grouped effects estimator is asymptotically unbiased and normal when the group membership is assumed to be known. Bonhomme and Manresa (2015) extend the literature to the unknown group membership. They use k-means algorithm to estimate the grouping along with the time-varying fixed effects in the linear panel data model. Then k-means method is further applied to nonlinear panel data model in Bonhomme et al. (2017, Working Paper).

In this paper, we also focus on a nonlinear panel data model with unknown group membership, but we release the assumption that individuals within a group are identical. Instead, we assume that the individual heterogeneities are from a mixture of a certain number of groups with the group-specific distribution. Our method belongs to the random model approach, but we don't assume the independence between unobservable heterogeneities and interest covariates. Compare to the k-means type clustering method in Bonhomme and Manresa (2015) or Bonhomme et al. (2017, Working Paper), our method provide more information, e.g., higher moments than just the first moment. In another word, our mixture model method is equivalent to k-means if only mean is concerned. Secondly, we provide both theoretical and numerical study on the relationship between the coefficient of interest covariates (β) and grouping in the generalized linear model (GLM). The grouping object, Bonhomme et al. (2017, Working Paper) call it "moment", include a latent common parameter β in the grouping estimation step. We find that the common parameter β will only affect the magnitude of individual heterogeneities but is a "nuisance" concerning grouping when T is large. This is useful since we can choose some particular value of β to improve the estimation of clustering. For example, when there are only large T observations of response variable but short T data of interest covariates, we can choose $\beta = 0$ to release the role of independent variables. In the first step, we only use the information of the response variable to estimate the group membership. Indeed, for some distribution, e.g., logistic distribution, it is known as sufficient statistics. Once we have optimal grouping in hand, we can go back to estimate grouped fixed effects and the common parameters. Thirdly, We also introduce a nonparametric approach to relax the restriction on the distribution of heterogeneities and leave the group-specific distribution to be estimated nonparametrically. However, the price of such flexibility is the inconsistency of the cluster distributions. Lastly, similar to the literature, the result using the nonparametric approach is much more robust to kernel compare to bandwidth. So we also provide numerical evidence of selecting the optimal bandwidth using an MSE as a criterion.

The paper is organized as follows. In section 2, we introduce our model and estimation method and use GLM as an example to show the coefficient of interest covariate is a nuisance in the grouping step. In section 3, we introduce the EM algorithm and the optional step if the group-specific distribution is to be estimated nonparametrically. We also briefly discuss the relationship between k-means and EM algorithm. In section 4, we study the behavior of the distribution of the grouping object and compare the performance of different estimation methods in the finite sample case. The last section summarize the paper.

2 Model and estimation

2.1 A general model

Let's denote the observed data as $W = \{w_{it}\}$, where $i = 1, \dots, N$ refers to N individuals, and each has $t = 1, \dots, T$ periods of observations. We consider a nonlinear panel data model with group-specific heterogeneity along with the grouping are assumed to be unknown and need to be estimated. The estimators are defined as

$$(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) = \underset{\beta,\alpha,\gamma}{\text{argmax}} \sum_{g=1}^{G} \sum_{i \in A_g} \sum_{t=1}^{T} \log p(w_{it}|\beta, \Theta_{g_i}) \tag{2.1}$$

where $p(\cdot)$ is the density function, and the data is balanced; β is the common parameter for all individuals; $\gamma = \{(g_1, \ldots, g_N)' : g_i \in \{1, \ldots, G\}, i = 1, \ldots, N\}$ is the unobserved group membership of unit i and G is assumed to be fixed; Θ_{g_i} is a set of parameters captures the grouped-level heterogeneity, e.g., grouped fixed effects or group-specific distribution, for all $i \in A_g$ where A_g represents certain group g. For convenience, we introduce a $N \times G$ matrix $Z = (z_1, \ldots, z_N)'$ to represent the grouping. $z_i = (z_{i1}, \ldots, z_{iG})'$ is a vector in which one of the element z_{ig} equals to 1, and all other elements equal to 0, denoting the individual i is from group g. For example, suppose the observations are drawn from 5 clusters, and a specific individual i belongs to the cluster where $g_i = 3$, then $z_i = (0, 0, 1, 0, 0)$. Here we let $z_{ig} \in \{0, 1\}$ and $\sum_{g=1}^{G} z_{ig} = 1$, but it can be easily extended to continuous variable (Fuzzy). Now the log-likelihood function of W can be re-written as

$$\begin{aligned} l(\beta, \Theta_{g_i}) = \sum_{n=1}^{N} \log p(w_i|\beta, \Theta_{g_i}) &= \sum_{i=1}^{N} \log \sum_{z_i} p(w_i, z_i|\beta, \Theta_{g_i}) \\ &= \sum_{i=1}^{N} \log \sum_{z_i} q(z_i) \frac{p(w_i, z_i|\beta, \Theta_{g_i})}{q(z_i)} \\ &\geq \sum_{i=1}^{N} \sum_{z_i} q(z_i) \log \frac{p(w_i, z_i|\beta, \Theta_{g_i})}{q(z_i)} \qquad \text{(Jensen's inequality)} \\ &\equiv L(q(z_i), \beta, \Theta_{g_i}) \end{aligned}$$

Instead of maximizing the actual log-likelihood, it's easier to maximize $L(q(z_i), \beta, \Theta_{g_i})$ which is a lower bound of the actual log-likelihood. The tightness of the lower bound can be shown by choosing $q(z_i)$ to be the posterior probabilities, i.e., $q(z_i) = p(z_i|w_i, \beta, \Theta_{g_i})$ (Aggarwal and

Reddy (2013)). Then

$$\begin{aligned}
L(p(z_i|w_i,\beta,\Theta_{g_i}),\beta,\Theta_{g_i}) &= \sum_{n=1}^{N}\sum_{z_i} p(z_i|w_i,\beta,\Theta_{g_i}) \log \frac{p(w_i,z_i|\beta,\Theta_{g_i})}{p(z_i|w_i,\beta,\Theta_{g_i})} \\
&= \sum_{n=1}^{N}\sum_{z_i} p(z_i|w_i,\beta,\Theta_{g_i}) \log p(w_i|\beta,\Theta_{g_i}) \\
&= \sum_{n=1}^{N} \log p(w_i|\beta,\Theta_{g_i}) = l(\beta,\Theta_{g_i})
\end{aligned}$$

In another word, the lower bound of log-likelihood $L(q(z_i),\beta,\Theta_{g_i})$ is maximized by $q(z_i) = p(z_i|w_i,\beta,\Theta_{g_i})$. From now, the maximization problem in equation (2.1) has been transformed into

$$\left(\hat{p}(z_i|w_i,\hat{\beta},\hat{\Theta}_{g_i}),\hat{\beta},\hat{\Theta}_{g_i}\right) = \underset{q,\beta,\Theta_{g_i}}{\operatorname{argmax}} L\left(q(z_i),\beta,\Theta_{g_i}\right) \tag{2.2}$$

Denote $\{\pi_g = p(g_i = g) : g = 1,\ldots,G\}$ (also denoted as $\pi_g = p(z_{ig} = 1)$) is the prior probability of unit i is from the cluster g, so $0 \leq \pi_g \leq 1, g = 1,\ldots,G$, and $\sum_{g=1}^{G} \pi_g = 1$. The marginal probability of z_i and the conditional distribution of w_i, given (β,Θ_{g_i}), are

$$\begin{aligned}
p(z_i|\beta,\Theta_{g_i}) &= \prod_{g=1}^{G} \pi_g^{z_{ig}} \\
p(w_i|z_i,\beta,\Theta_{g_i}) &= \prod_{g=1}^{G} p(w_i|\beta,\Theta_{g_i})^{z_{ig}}
\end{aligned}$$

respectively. Then the joint distribution is

$$\begin{aligned}
p(w_i,z_i|\beta,\Theta_{g_i}) &= p(z_i|\beta,\Theta_{g_i})p(w_i|z_i,\beta,\Theta_{g_i}) \\
&= \prod_{g=1}^{G} \left(\pi_g \cdot p(w_i|\beta,\Theta_{g_i})\right)^{z_{ig}}
\end{aligned}$$

and the marginal distribution of w_i, given (β, Θ_{g_i}), is

$$
\begin{aligned}
p(w_i|\beta, \Theta_{g_i}) = \sum_{z_i} p(w_i, z_i|\beta, \Theta_{g_i}) &= \sum_{z_i} \prod_{g=1}^{G} (\pi_g \cdot p(w_i|\beta, \Theta_{g_i}))^{z_{ig}} \\
&= \sum_{g=1}^{G} \pi_g \cdot p(w_i|\beta, \Theta_{g_i})
\end{aligned}
$$

According to Bayes' theorem, the posterior probability of $z_{ig} = 1$, given $(w_i, \beta, \Theta_{g_i})$ is

$$
\begin{aligned}
p(z_{ig} = 1|w_i, \beta, \Theta_{g_i}) &= \frac{p(z_{ig} = 1) \cdot p(w_i|z_{ig} = 1, \beta, \Theta_{g_i})}{p(w_i|\beta, \Theta_{g_i})} \\
&= \frac{\pi_g \cdot p(w_i|\beta, \Theta_{g_i})}{\sum_{g=1}^{G} \pi_g \cdot p(w_i|\beta, \Theta_{g_i})}
\end{aligned}
\tag{2.3}
$$

Recall we have shown that the lower bound of the actual likelihood $L(q(z_i), \beta, \Theta_{g_i})$ is maximized by (2.3). Then the optimization problem $L(Z, \Theta)$ $(\Theta = (\pi_g, \beta, \Theta_{g_i}))$ in (2.2) can be solved by iteratively solving the following two sub-problems:

1. Fix $\Theta = \hat{\Theta}$, solving the sub-problem $L(Z, \hat{\Theta})$ by letting

$$
z_{ig} = \begin{cases} 1, \text{ if } \hat{p}(z_{ig} = 1|w_i, \hat{\beta}, \hat{\Theta}_{g_i}) = \max\limits_{1 \le j \le G} \hat{p}(z_{ij} = 1|w_i, \hat{\beta}, \hat{\Theta}_{g_i}) \\ 0, \text{ otherwise} \end{cases}
$$

   This step assigns the individual i to the cluster g with the largest posterior probability. When fuzzy clustering is preferred, we can simply let $z_{ig} = p(z_{ig} = 1|w_i, \hat{\beta}, \hat{\Theta}_{g_i})$ so that z_{ig} is a continuous variable.

2. Fix $Z = \hat{Z}$, the solutions to the sub-problem $L(\hat{Z}, \Theta)$ are defined as

$$
\begin{aligned}
(\hat{\beta}, \hat{\Theta}) &= \underset{\beta, \Theta}{\text{argmax}} \sum_{i=1}^{N} \sum_{g=1}^{G} \hat{z}_{ig} \log p(w_i|\beta, \Theta_{g_i}) \\
\hat{\pi}_g &= \frac{\sum_{i=1}^{N} \hat{z}_{ig}}{N}
\end{aligned}
$$

This step can be considered as finding the MLE when the group membership is known. The prior probability of cluster g, π_g, is obtained by the number of individuals in cluster g divided by the total number of individuals.

In practice, the two-steps procedures above can be implemented using EM algorithm which will be introduced in detail in the next section.

Mixture of unspecific densities: So far we have been assuming the densities of each cluster, $p(w_i|\beta, \Theta_{g_i})$, follows a probability distribution based on a fixed set of parameters. But we can also relax this assumption and estimate the densities using a non-parametric method in Benaglia et al. (2009b). The model is defined as

$$\begin{aligned} p(w_i|\beta, \Theta_{g_i}) &= p_{g_i}(w_i(\beta)) \\ \text{where} \quad p_{g_i}(u) &= \frac{1}{N h_g \pi_g} \sum_{i=1}^{N} z_{ig} K\left(\frac{u - w_i(\beta)}{h_g}\right) \end{aligned} \tag{2.4}$$

where $K(\cdot)$ is a kernel density function, h_g is the bandwidth for the g-th component density estimate. Since β is the common parameter so that we can "extract" it from the estimation of group-specified density. In another word, all the observations will be first rescaled by β and then used to estimate component densities. In many cases, we estimate group membership from the scaled data $W(\beta)$ rather than the actual data W. For example, in the case of linear model with grouped fixed effects, the grouping is recovered from the object $Y - X\beta$. Even if we know the population distribution of W, it's still not the clear the distribution is the same if the grouping object is $W(\beta)$. With this non-parametric grouping method, we don't need to worry about the distribution of $W(\beta)$ and can leave it to be estimated. However, the higher flexibility comes with a price. The identification of model (2.4) is still under discussion. For example, Bordes et al. (2006) and Hunter et al. (2007) found that if the univariate data are drawn from a mixture of up to three symmetric components, the problem is identifiable except in certain cases that are easily enumerable. We use the term

”non-parametric” because symmetric distribution is the only requirement but the others are completely unspecified.

2.2 Example: A glm with group fixed effects

In this section, we use glm with grouped fixed effects as an example to illustrate our method. Let the observed data be $W = \{Y, X\}$. Y is the $N \times T$ binary response variable, and X is a set of exogenous variables, which has dimension $K \times N \times T$. The model is

$$P(y_{it} = 1 | x_{it}) = G(x'_{it}\beta + \alpha_{g_i})$$

From example, the model is probit, if $G(\cdot) = \Phi(\cdot)$ ($\Phi(\cdot)$ is the C.D.F. of standard normal distribution), and the model is logit if $G(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$. The model is known up to a finite dimensional common parameters, β, and a set of group specific parameters, $\{\alpha_{g_i}\}$. The log-likelihood is

$$l(Z, \beta, \alpha) = \frac{1}{N} \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \phi_i(\beta, \alpha_{g_i})$$

$$\phi_i(\beta, \alpha_{g_i}) = \frac{1}{T} \sum_{t=1}^{T} (y_{it} \log(G(x'_{it}\beta + \alpha_{g_i})) + (1 - y_{it}) \log(1 - G(x'_{it}\beta + \alpha_{g_i})))$$

Given any β, define

$$a_i(\beta) = \underset{\alpha}{\operatorname{argmax}} \, \phi_i(\beta, \alpha)$$

$a_i(\beta)$ can be considered as the MLE of individual fixed effect which only depends on β and the T periods observations of individual i. Denote the random variables, $\mathcal{X}_N(\beta) = \{a_i(\beta)\}_{i=1}^N$, and assume there exists a cdf $P_{\mathcal{X}}(\beta)$, such that $\mathcal{X}_N(\beta)$ converge in distribution to random variable $\mathcal{X}(\beta)$ with cdf $P_{\mathcal{X}}(\beta)$, as $N, T \to \infty$.

We also assume the distribution $P_{\mathcal{X}}(\beta)$ is a mixture of finite G component distributions from location-scaled family, and each component is parameterized by a global parameter β and component-specific parameters $\{\Theta_g = (\mu_g, \sigma_g^2) : g = 1, \ldots, G\}$, that is

$$P_{\mathcal{X}}(\beta) = \sum_{g=1}^{G} \pi_g p_g \left(\Theta_g(\beta)\right)$$

A special case is grouped fixed effect if $\mathcal{X}$ is drawn from a set of G discrete numbers μ_g and $\sigma_g^2 = 0$ for $g = 1, \ldots, G$. Since $\mathcal{X}$ captures the time-invariant fixed effects so it is one-dimensional. Then the mean and variance of $\mathcal{X}(\beta)$ are

$$\begin{aligned}\mu(\beta) &= E[\mathcal{X}(\beta)] = \sum_{g=1}^{G} \pi_g \mu_g(\beta)\\ \sigma^2(\beta) &= Var(\mathcal{X}(\beta)) = \sum_{g=1}^{G} \pi_g \left(\mu_g(\beta)^2 + \sigma_g(\beta)^2\right) - \left(\sum_{g=1}^{G} \pi_g \mu_g(\beta)\right)^2\end{aligned}$$

We can express the component-specific parameters $(\mu_g(\beta), \sigma_g(\beta))$ which depend the global parameter β as

$$\begin{aligned}\mu_g(\beta) &= \mu(\beta) + \sigma(\beta) v_g\\ \sigma_g(\beta) &= \sigma(\beta)\zeta_g\end{aligned}$$

where $\mu(\beta) \in \mathbf{R}$, and $\sigma(\beta) > 0$. By this reparameterization, we divide the actual component-specific parameters into two parts: the global parameters which depend the β, and the new component-specific parameters, (v_g, ζ_g), which are indenpendent from β. It's easy to show that (v_g, ζ_g) satisfies

$$\sum_{g=1}^{G} \pi_g v_g = 0 \text{ and } \sum_{g=1}^{G} \pi_g (v_g^2 + \zeta_g^2) = 1$$

Then, we can recover the grouping from a new random variable $\tilde{\mathcal{X}}(\beta) = \frac{\mathcal{X}(\beta) - \mu(\beta)}{\sigma(\beta)}$ since the posterior distribution can be expressed as following

$$p(z_{ig}=1|a_i(\beta),\Theta_{g_i}(\beta)) = \frac{\pi_g\cdot p(a_i(\beta)|\Theta_{g_i}(\beta))}{\sum_{g=1}^{G}\pi_g\cdot p(a_i(\beta)|\Theta_{g_i}(\beta))}$$
$$= \frac{\pi_g\cdot p\left(\frac{a_i(\beta)-\mu'(\beta)}{\sigma'(\beta)}\middle|\Theta_{g_i}(1)\right)}{\sum_{g=1}^{G}\pi_g\cdot p\left(\frac{a_i(\beta)-\mu'(\beta)}{\sigma'(\beta)}\middle|\Theta_{g_i}(1)\right)}$$

where $\Theta_{g_i}(1)=(v_{g_i},\zeta_{g_i})$ does not depend on β any more, and so the component distributions. This is saying that after we re-centered and re-scaled the original random variable $\mathcal{X}$ by some global components (a function of global parameter β) among all groups, the new random variable $\tilde{\mathcal{X}}$ follows the mixture of distributions which are independent from β. In another word, the group membership of $\mathcal{X}$ is robust to the common parameter β. β only matters for the magnitude location and scale parameters. Note that we are study the clustering of fixed effects which only contains one dimensional feature. For the multi-dimensional case, this may not be true in general. For example, the group pattern may vary if the we give different dimension a different scale.

Then, it's straightforward to see two equivalent ways to recover the grouping. The first one is, for any β, grouping $\{a_i(\beta)\}_{i=1}^{N}$ by iteratively doing the following two steps:

1. Fix $(\pi_g,\Theta_{g_i}(\beta))=\left(\hat{\pi}_g,\hat{\Theta}_{g_i}(\beta)\right)$, computing the posterior probability

$$p(z_{ig}=1|a_i(\beta),\Theta_{g_i}(\beta)) = \frac{\pi_g\cdot p(a_i(\beta)|\Theta_{g_i}(\beta))}{\sum_{g=1}^{G}\pi_g\cdot p(a_i(\beta)|\Theta_{g_i}(\beta))}$$

   and then let

$$z_{ig}=\begin{cases}1, \text{ if } \hat{p}(z_{ig}=1|a_i(\beta),\hat{\Theta}_{g_i}(\beta)) = \max\limits_{1\le j\le G}\hat{p}(z_{ij}=1|a_i(\beta),\hat{\Theta}_{g_i}(\beta))\\ 0, \text{ otherwise}\end{cases}$$

2. Fix $Z = \hat{Z}$, the solutions to the sub-problem $L(\hat{Z}, \Theta)$ are defined as

$$\{\hat{\Theta}_{g_i}(\beta)\}_{g_i=1}^{G} = \underset{\Theta}{\text{argmax}} \sum_{i=1}^{N} \sum_{g=1}^{G} \hat{z}_{ig} \log p(a_i(\beta)|\Theta_{g_i}(\beta))$$

$$\hat{\pi}_g = \frac{\sum_{i=1}^{N} \hat{z}_{ig}}{N}$$

The second one is almost the same except we group the random variable $\tilde{\mathcal{X}}$. For any β, we first estimate the global component $(\hat{\mu}(\beta), \hat{\sigma}(\beta))$ and compute $\tilde{\mathcal{X}}(\beta) = \frac{\mathcal{X}(\beta)-\hat{\mu}(\beta)}{\hat{\sigma}(\beta)}$. The rest is exactly the same with the two steps we introduced above. All notations are the same with section (2.1) so we don't restate them here.

With the optimal grouping $Z^* = (z^*_{ig})$ in hands, the common parameter and the group effects are

$$(\hat{\beta}, \{\hat{\alpha}_{g_i}\}) = \underset{\beta, \{\alpha_{g_i}\}_{g_i=1}^{G}}{\text{argmax}} \left[\sum_{i=1}^{N} \sum_{g=1}^{G} z^*_{ig} \phi_i(\beta, \alpha_{g_i}) \right]$$

The large sample properties of the panel data model with grouped fixed effects when the group membership is known has been well studied in Bester and Hansen (2016).

3 Algorithm

3.1 EM algorithm

In this section, we introduce the EM (Expectation-Maximization) algorithm which is a popular iteration approach to find the latent group membership in the probabilistic models. Denote the current estimates of the parameters are $\Theta^{(t)}$ (In our case, $\Theta = (\pi_g, \alpha_{g_i})$, $g, g_i = 1, \ldots, G$). Then the EM algorithm maximizes the likelihood by doing the following two steps iteratively untill the convergence criterion is satisfied (Aggarwal and Reddy (2013)).

E-step: Maximize $L(q(z_i), \Theta)$ with respect to $q(z_i)$ while holing Θ fixed.

$$q(z_i) = \underset{q}{\operatorname{argmax}} L(q(z_i), \Theta^{(t)}) = p(z_i|w_i(\beta), \Theta^{(t)})$$

We have shown that $L(q(z_i), \Theta) = l(\Theta)$ (i.e., $L(q(z_i), \Theta)$ reachs the upper bound) when $q(z_i)$ is the posterior distribution $p(z_i|w_i(\beta), \Theta^{(t)})$, and $l(\Theta)$ doesn't depends on $q(z_i)$. So the maximization in E-step is just computing the posterior distribution $p(z_i|w_i(\beta), \Theta^{(t)})$ given $\Theta^{(t)}$.

M-step: The distribution $q^{(t)}(z_i)$ is held fixed and the estimates of parameters are updated by

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} L(q^{(t)}(z_i), \Theta)$$

When $q^{(t)}(z_i) = p(z_i|w_i(\beta), \Theta^{(t)})$, the lower-bound likelihood can be written as

$$\begin{aligned} L(p(z_i|w_i(\beta), \Theta^{(t)}), \Theta) &= \sum_{n=1}^{N} \sum_{z_i} p(z_i|w_i(\beta), \Theta^{(t)}) \log \frac{p(w_i(\beta), z_i|\Theta)}{p(z_i|w_i(\beta), \Theta^{(t)})} \\ &= \sum_{n=1}^{N} \sum_{z_i} p(z_i|w_i(\beta), \Theta^{(t)}) \log p(w_i(\beta), z_i|\Theta) \\ &\qquad - \sum_{n=1}^{N} \sum_{z_i} p(z_i|w_i(\beta), \Theta^{(t)}) \log p(z_i|w_i(\beta), \Theta^{(t)}) \end{aligned}$$

As the second term is independent of Θ, the maximization problem of $L(p(z_i|w_i(\beta), \Theta^{(t)}), \Theta)$ is equivalent to

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{z_i} p(z_i|w_i(\beta), \Theta^{(t)}) \log p(w_i(\beta), z_i|\Theta)$$

which is in fact maximizing the expectation of complete data log-likelihood.

Nonparametric density estimation step: In addition to the E-M steps, we need a third step to estimate the density functions of each component nonparametrically if we assume no prior knowledge of them. Follow Benaglia et al. (2009b), for any $u \in \mathcal{R}$, the density function is estimated by

$$\begin{aligned}
\phi_g^{t+1}(u) &= \frac{\frac{1}{h_g}\sum_{i=1}^{N} q^{(t)}(z_i)K((u - w_i(\beta))/h_g)}{\sum_{i=1}^{N} z_{ig}^{t}} \\
&= \frac{1}{h_g N \pi_g^{t+1}} \sum_{i=1}^{N} q^{(t)}(z_i) K\left(u - w_i(\beta)/h_g\right)
\end{aligned}$$

The result of this step depends the choice of kernel and bandwidth, but choosing bandwidth is more important issue in nonparametric estimation since the estimate is often robust to kernel function. So we use the Gaussian density as the kernel density. About how to decide the bandwidth is another complicated question in the nonparametric literature. Here, we don't expand this question just provide the simulation results using different bandwidth in the simulation section.

Algorithm 1 EM Algorithm

```
Start with an initial valuse of $\Theta^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_G^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_G^{(0)}\}$
Compute the posterior distribution $q(z_{ig} = 1|\Theta^{(0)}) = \frac{\pi_g \phi(w_i|\alpha_g)}{\sum_{j=1}^{G} \pi_j \phi(w_i|\alpha_j)}$ and then the initial log-likelihood $L\left(q^{(0)}, \Theta^{(0)}\right)$
while $\left(\left|L\left(q^{(t+1)}, \Theta^{(t+1)}\right) - L\left(q^{(t)}, \Theta^{(t)}\right)\right| < \epsilon\right)$ do
    E-step: compute the $q^{(t)} \equiv q(z_{ig} = 1|\Theta^{(t)})$
    M-step: update $\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} L(q^{(t)}, \Theta) = \underset{\Theta}{\operatorname{argmax}} L\left(\Theta^{(t)}, \Theta\right)$
    (Optional)
    if (distribution $\phi_g(W)$ is not specified) then
        Kernel density estimation step: $\phi_g^{t+1}(u) = \frac{1}{hN\pi_g^{t+1}} \sum_{i=1}^{N} q^{(t)} K\left(\frac{u - w_i}{h}\right)$
```

The general convergence properties of EM algorithm has been discussed in detail in Aggarwal and Reddy (2013). In practice, there often exists a closed-form solution of M-step, but there are many cases for which it is intractable to find the maximizer Θ. To address those situations, a generalized EM algorithm (GEM) is defined, which chooses a better $\Theta^{(t+1)}$ to increase the objective function rather than to find the local maximum of it. GEM is widely used in practice since it reduces the computation burden and the convergence can still be guaranteed.

Relation of EM algorithm with K-means In the last part of this section, we show that K-means can also be achieved through EM-steps which has been well known in the literature, e.g., Aggarwal and Reddy (2013).

Given some G initial values of α_{g_i}, K-means is equivalent to doing the following EM-steps iteratively until the convergence condition is met.

E-step: Let

$$z_{ig}^{(s+1)} = \begin{cases} 1, \text{ if } d_{euc}(w_i, \alpha_{g_i}^{(s)}) = \underset{g_i}{\text{argmin}} \left\{ d_{euc}(w_i, \alpha_{g_i}^{(s)}) \right\}_{g_i=1}^{G} \\ 0, \text{ otherwise} \end{cases}$$

M-step:

$$\alpha_{g_i}^{(s+1)} = \frac{\sum_{i=1}^{N} z_{ig}^{(s)} w_i}{\sum_{i=1}^{N} z_{ig}^{(s)}}$$

In E-step, the individuals are assigned to the nearest cluster. In the following M-step, α_{g_i} are updated based on the new group membership z_{ig}.

4 Asymptotic properties

Let the observations $x_1, \ldots, x_n$ be i.i.d., and drawn form a finite mixture model. The log-likelihood function is

$$l_n(\mathcal{G}) = \sum_{i=1}^{N} f(x_i|\mathcal{G})$$

$$\text{where } f(x|\mathcal{G}) = \sum_{g=1}^{G} \pi_g p_g(\theta_g)$$

Define the MLE $\hat{\mathcal{G}}$ of $\mathcal{G}$ as

$$l_n(\hat{\mathcal{G}}) = \sup_{\mathcal{G} \in \mathbb{G}} l_n(\mathcal{G}) \tag{4.1}$$

Note that when the number of groups $G = 1$, the problem degenerate to the ordinary MLE of $\mathcal{G}$, and its consistency has been proved by Wald (1949). Here we consider the nontrivial case when $G > 1$, then (4.1) becomes the MLE of $\mathcal{G}$ under finite mixture model.

4.1 Consistency of Nonparametric MLE under mixture model

Kiefer and Wolfowitz (1956) show the consistency of nonparametric MLE. In their case, $\mathcal{G}$ are allowed to contain infinite many parameters, but some of their assumptions are difficult to verify. We first briefly introduce their approach below. Let $\mathcal{G} = (\theta, F)$ be the generic point in $\mathbb{G} = \Omega \times \Gamma$, where Ω is the parameter space and Γ is the distribution space. Denote $\bar{\mathbb{G}} = \bar{\Omega} \times \bar{\Gamma}$ as the completed space of $\mathbb{G}$, so $\mathbb{G}$ is compact. We also need to define a metric $D_{KW}(\cdot, \cdot)$ on $\mathbb{G}$ which is the same with the metric defined in Kiefer and Wolfowitz (1956) as

$$\begin{aligned} D_{KW}(\mathcal{G}_1, \mathcal{G}_2) = &|\arctan \theta_1 - \arctan \theta_2| \\ &+ \int_{\tau} |F_1(\tau), F_2(\tau)| \exp(-|\tau|) d\tau \end{aligned}$$

where $|\theta| = |\theta_1| + \cdots + |\theta_k|$ and $d\tau = d\tau_1 \dots d\tau_k$. $D_{KW}(\cdot, \cdot)$ can be understood as the distance of a bijective transformation of θ plus a Laplace transformation of F. Let $\mathcal{G}^0 \in \mathbb{G}$ be the true mixture distribution and $\hat{\mathcal{G}} \in \mathbb{G}$ be the MLE of $\mathcal{G}^0$. Then $\hat{\mathcal{G}}$ is consistent if $D_{KW}(\hat{\mathcal{G}}, \mathcal{G}^0) \to 0$ almost for sure.

Assumptions in Kiefer and Wolfowitz (1956):

KW1(Identifiability): Let $F(x|\mathcal{G})$ be the c.d.f. of $f(x|\mathcal{G})$. If $F(x|\mathcal{G}) = F(x|\mathcal{G}^0)$ for any x, then $D_{KW}(\mathcal{G}, \mathcal{G}^0) = 0$.

KW2(Continuity): The set of the component parameter is closed. For all x and any $\mathcal{G}^0$, there is $\lim_{\mathcal{G} \to \mathcal{G}^0} f(x|\mathcal{G}) = f(x|\mathcal{G}^0)$.

KW3(Integrability): For any $\mathcal{G} \neq \mathcal{G}^0$, there exist an $\epsilon > 0$ and an open ball $B_\epsilon(\mathcal{G}) = \{\mathcal{G} : D_{KW}(\mathcal{G}, \mathcal{G}^0) < \epsilon\}$ such that

$$E^0[\log\{f(x|B_\epsilon(\mathcal{G}))/f(x|\mathcal{G}^0)\}]^+ < \infty$$

where $E^0[\cdot]^+$ represents the expectation of the positive part of the quantinty in $[\cdot]$ with respect to true mixture distribution.

KW4(Compactness): The definition of $f(x|\mathcal{G}^0)$ can be extended to a compact space $\bar{\mathbb{G}}$ while KW3 still hold.

Theorem 4.1. *With assumptions KW1-KW4 hold, the nonparametric MLE of $\mathcal{G}^0$, $\hat{\mathcal{G}}$, is strongly consistent, i.e., $D_{KW}(\hat{\mathcal{G}}, \mathcal{G}^0) \to 0$ almost for sure as $n \to \infty$.*

The proof of Theorem 4.1 can be found in Kiefer and Wolfowitz (1956) and a simplified version can be found in Chen (2017). Note this result is valid for both finite and infinite mixtures. However, the generality comes from the high-level assumptions of KW3 and KW4 which are difficult to verify. For example, even when is component distribution is Poisson or Gaussian.

4.2 Consistency of MLE under finite normal mixture (univariate)

In this section, we sacrifice some flexibility in order to get more practical results. We assume the i.i.d. observations $x_1, \ldots, x_n$ are from a mixture of G univariate normals. Formally, the likelihood function is

$$l_n(\mathcal{G}) = \sum_{i=1}^{N} f(x_i|\mathcal{G})$$
$$\text{where } f(x|\mathcal{G}) = \sum_{g=1}^{G} \pi_g N(\mu_g, \sigma_g^2)$$

where G is finite and assumed to be known. The support space of $\mathcal{G}$ is then

$$\mathbb{G} = \left\{ \Omega = (\pi_1, \ldots, \pi_G, \mu_1, \ldots, \mu_G, \sigma_1, \ldots, \sigma_g) : \sum_{g=1}^{G} \pi_g = 1, \pi_g \geq 0, \sigma_g \geq 0, g = 1, \ldots, G \right\}$$

As we have confined the model to normal mixture, KW1 can then be proved to be satisfied by the following lemma (Proposition 1 in Teicher (1963)).

Lemma 4.1. *The class of all finite mixtures of normal distributions is identifiable.*

Though the identifiability is no long an issue, we still encounter a problem that the MLE may not be well-defined. An example in Day (1969) is by letting $\mu_1 = x_1$ and $\sigma_1 \to 0$ with the other parameters fixed, the log-likelihood function

$$l_n(\mathcal{G}) = \sum_{i=1}^{N} \log f(x_i|\mathcal{G}) = \sum_{i=1}^{N} \log \left(\sum_{g=1}^{G} \frac{\pi_g}{\sqrt{2\pi}\sigma_g} \exp \left(-\frac{(x_i - \mu_g)^2}{2\sigma_g^2} \right) \right) \to \infty$$

Same problem also makes the finite normal mixture model violate KW3 so the consistency is no longer guaranteed. To address this issue, Hathaway (1985) impose the constraints on σ which can be written as

$$\min_{g,g' \in \{1,\ldots,G\}} \left(\frac{\sigma_g}{\sigma_{g'}} \right) \geq c > 0 \tag{4.2}$$

where c is some constant. Then there is following nice result

Theorem 4.2 (Theorem 3.3 in Hathaway (1985))**.** *Let the parameter space Ω satisfy (4.2), the true value of $\Theta^0 = (\pi^0, \mu^0, \sigma^0) \in \Omega$, and $c \in (0, 1]$. Let $\hat{\Theta}_n = (\hat{\pi}, \hat{\mu}, \hat{\sigma})$ be the maximizer of $l_n(\mathcal{G})$. Then $\hat{\Theta}_n$ is strongly consistent.*

The inconsistency of the MLE under finite normal mixture is mainly due to the nonregularity of parameter space, or more specific, σ. Beside the constraint approach in Hathaway (1985), in literature, there is another approach known as penalized MLE (PMLE). The basic idea of PMLE is maximizing the penalized likelihood function as follows

$$\bar{l}_n(\mathcal{G}) = l_n(\mathcal{G}) + p(\sigma)$$

where $p(\sigma)$ is the penalty put on σ to avoid the unboundness issue of $l_n(\mathcal{G})$. Some advantages of PMLE are, for example, the original parameter space is unchanged by the constraints, and the convergence speed is at least as fast as the MLE (Green (1990)). However, how to choose a suitable penality function $p(\sigma)$ to obtain the consistency is an important issue. We list the assumptions on $p(\sigma)$ in Ciuperca et al. (2003) and the corresponding result below. A more general proof can also be found in Chen et al. (2008)

Theorem 4.3 (Theorem 2 in Ciuperca et al. (2003))**.** *Let the penalty function $p(\sigma)$ satisfy*

C1: $lim_{\sigma \to 0} \frac{p(\sigma)}{\sigma^n} = 0$ for all n.

C2: Given G, $p(\sigma)$ is many-to-one from $(0, \infty)$ onto $(0, G]$, G.

C3: $p(\sigma)$ is measurable and increasing in an arbitrary small open interval $(0, \delta)$.

C4: $p(\sigma)$ is continuously differentialble on $(0, \infty)$.

Let the true value of $\Theta^0 = (\pi^0, \mu^0, \sigma^0) \in \bar{\Omega}$, and $\hat{\Theta}_n = (\hat{\pi}, \hat{\mu}, \hat{\sigma})$ be the maximizer of $\bar{l}_n(\mathcal{G})$. Then $\hat{\Theta}_n$ is strongly consistent.

Assumption C1 is to make sure the penalty term dominant the likelihood. In another word, the penalized estimator

$$\hat{\mathcal{G}} = \operatorname*{argmax}_{\mathcal{G} \in \mathbb{G}} \bar{l}_m(\mathcal{G})$$

exists for all n. C4 is for the consistency of a penalized estimator over a compact set.

4.3 Convergence rate

The optimal convergence rate also have been well studied. It is shown that $\sqrt{n}$ convergence rate is achievable when the number of group G is known. For example, Ciuperca et al. (2003) show the case of univariate normal mixture. When the number of group is only known up to an upper bound, Chen (1995) find the convergence rate of MLE under finite mixture of the location and scale families, such as normal and Cauchy distribution, is at most $n^{\frac{1}{4}}$. The $n^{\frac{1}{4}}$ convergence rate is achievable by choosing, for example, the Kolmogorov-Smirnov distance.

4.4 Consistency of EM algorithm

In this section, we discuss the consistency of EM algorithm which is firstly formally proposed by Dempster et al. (1977). Continue using the notation in section 3, we can see that, after E-step, the object function

$$l(\Theta^{(t)}) = L(q^{(t)}, \Theta^{(t)}) \geq L(q^{(t-1)}, \Theta^{(t)})$$

We have shown E-step is computing the posterior probability given $\Theta^{(t)}$. The M-step is the ordinary MLE problem, and thus

$$L(q^{(t)}, \Theta^{(t+1)}) \geq L(q^{(t)}, \Theta^{(t)})$$

Putting two inequalities together, we can see, after each iteration,

$$L(q^{(t+1)}, \Theta^{(t+1)}) \geq L(q^{(t)}, \Theta^{(t)})$$

this guarantees the EM algorithm convergence to a local optimum. A more detailed discussion and proof can be found in Dempster et al. (1977). EM algorithm has the same issue with most of the optimal algorithm, which is the program can be trapped in a local optimum rather than the global one. One way to address this problem is starting search from many different initial values.

4.5 Asymptotic distribution

In the previous sections, we have shown the grouping can be estimated from

$$a_i(\beta) = \underset{\alpha}{\operatorname{argmax}} \lim_{T \to \infty} l_i(\beta, \alpha)$$

and the common parameter β is a nuisance in the grouping step. We model $a_i(\beta)$ is drawn from a mixture of finite number of component distributions and show the consistency of nonparametric MLE under such a setting. However, the nice result is mainly form the high level assumptions which are not valid even for the Poisson mixture or Gaussian mixture. A good thing is $a_i(\beta)$ represents the fixed effect over time and hence is univariate. Also, assuming $a_i(\beta)$ is a mixture of densities from location-scale family is often good enough in practice. These restrictions make the problem much simpler and thus give us very nice result. We have shown the MLE of $\Theta = (\pi, \mu, \sigma)$, $\hat{\Theta} = (\hat{\pi}, \hat{\mu}, \hat{\sigma})$ is strongly consistent if $a_i(\beta)$, is from a univariate Gaussian mixture (Cauchy mixture is the same). This implies the grouping can be consistently estimated as

$$\operatorname*{plim}_{N\to\infty} \hat{z}_{ig} = z_{ig}$$
$$\text{where } \hat{z}_{ig} = \mathbb{1}\left[\hat{p}(z_{ig} = 1|a_i(\beta), \hat{\Theta}_{g_i}) = \max_{1\leq j\leq G} \hat{p}(z_{ij} = 1|a_i(\beta), \hat{\Theta}_{g_i})\right]$$

After that, our problem can be considered as a nonlinear panel data model with grouped fixed effect and known group membership.

5 Simulation

In this section, we report the finite sample behavior of unconditional logit estimator (UCL), the bias-corrected unconditional logit estimator (BCL) and the conditional logit estimator (CL) in the following cases: grouped fixed effects model with known group membership, individual fixed effects model, and the grouped fixed effects model with unknown group membership. In the case that the grouping needs to be estimated, we only report the behavior of the unconditional logit estimator but using three grouping methods: K-means, Gaussian mixture model (Mixture), and nonparametric mixture model (MixNP). Besides, all individuals are assigned to one group (Pooling) is also considered. We use mean square error (MSE) as our criterion to evaluate the performance the each estimator. The conditional estimator

5.1 Data generating process

Our data generating process (DGP) is

$$
\begin{aligned}
x_{it} &= \gamma\alpha_{g_i} + e_{it} \\
v_{it} &= \log\left(\frac{u_{it}}{1-u_{it}}\right) \\
w_{it} &= \beta x_{it} + \alpha_{g_i} \\
y_{it} &= \mathbb{1}[w_{it} + v_{it} > 0]
\end{aligned}
$$

where $\beta = 1, \gamma = 1, e_{it}, u_{it} \sim N(0, 1)$. We set the number of groups G=3, the group fixed effect α_{g_i} is generated as follows:

DGP 1: $\{\alpha_{g_i}\}_{g_i=1}^{G} \sim \sum_{g=1}^{G} \pi_g \cdot \mathcal{N}(\mu_g, \sigma_g)$, where $\{\mu_g\}_{g=1}^{G} = \{-1, 0, 1\}$ and $\{\sigma_g\}_{g=1}^{G} = \{0.1, 0.01, 0.5\}$.

DGP 2: $\{\alpha_{g_i}\}_{g_i=1}^{G} \sim \sum_{g=1}^{G} \pi_g \cdot \mathcal{U}\{-1, 0, 1\}$.

DGP 3: $\{\alpha_{g_i}\}_{g_i=1}^{G} \sim \sum_{g=1}^{G} \pi_g \cdot \phi_g$, where $\phi_1 = N(-1, 0.1)$, $\phi_2 = U\{0\}$, and $\phi_3 = \chi^2(1)$.

In DGP 1, α_{g_i} is drawn from a mixture of G Gaussian distributions with group-specific parameters. In DGP 2, α_{g_i} is identical within group and different between groups. DGP 3 assumes α_{g_i} is from different type of distribution. The prior probability of group is $\pi_g = 1/G$ for $g = 1, \ldots, G$. We allow the independent variable x_{it} to be arbitrarily correlated with the group heterogeneities α_{g_i}. The number of individuals $N = 90$, and each experiment is iterated for 500 times, at which for each iteration x_{it}, α_{g_i}, and y_{it} vary.

5.2 Common parameter is a nuisance for finding the grouping in GLM

We continue our glm example in the second section and provide some numerical evidence that common parameter β is a nuisance in the step of estimation of grouping. Denote β^0,

$\alpha^0_{g_i}$, and $Z^0 = \{z^0_{ig}\}$ as the true values of β, α_{g_i}, and the true group membership, respectively. Then by law of large numbers, under some regularity conditions, we have

$$\sup_{\beta,\alpha_{g_i}\in\Theta} |\phi_i(\beta,\alpha_{g_i}) - E[\phi_i(\beta,\alpha_{g_i})]| \to 0 \text{ a.s. } T \to \infty$$

Lemma. *(Lemma 2.2 in Newey and McFadden (1994)): If θ^0 is identified ($\theta \neq \theta^0$ and $\theta \in \Theta$ implies $f(z|\theta) \neq f(z|\theta^0)$), and $E[|\log f(z|\theta)|] < \infty$ for all θ, then $Q^0(\theta) = E[\log f(z|\theta)]$ has a unique maximum at the θ^0.*

By Lemma and definition, respectively,

$$(\beta^0, \{\alpha^0_{g_i}\}^G_{g_i=1}) = \underset{\beta,\alpha_{g_i}}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} z^0_{ig} E[\phi_i(\beta,\alpha_{g_i})]$$

$$(\hat{\beta}, \{a_i(\hat{\beta})\}^N_{i=1}) = \underset{\beta,\alpha}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N} \phi_i(\beta,\alpha)$$

Because of the consistency of MLE, we know plim $\left(\hat{\beta}, a_i(\hat{\beta})\right) = \left(\beta^0, \alpha^0_{g_i}\right)$ as $N, T \to \infty$.

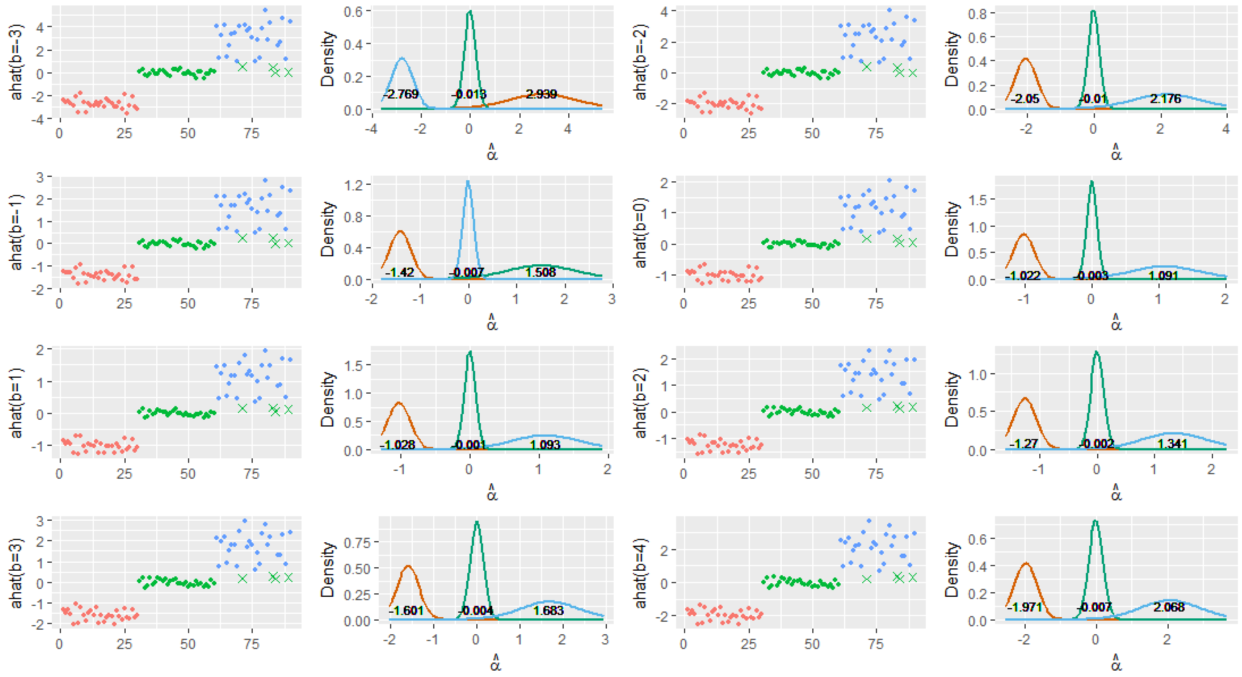


Figure 5.1: $\{a_i(\beta)\}^N_{i=1}$ and estimated distribution given different β (N=90, T=600, DGP1).

When T is large: Figure 5.1 shows $\{a_i(\beta)\}_{i=1}^N$ and the estimated grouping and component densities given $\beta \in \{-3,-2,-1,0,1,2,3,4\}$. Recall the true value $\beta^0 = 1$ in the DGP. We set (N,T)=(90,600), T is much larger than N so the incidental parameter problem is negligible. For now, we only focus on the point that the true value of β is not necessary to find the group membership. β only affects the magnitude of the group-specific parameters. And we can always get the better estimation of β along with the group heterogeneities in the second step when we have the optimal grouping. Among all values of β, $\beta = 0$ is quite special, since it implies the regressor x_{it} can be released from grouping. This is useful when we only have large T for y_it but very short T for x_it. In another word, we first use only large T information of y_{it} to estimate group membership. In the second step, with the optimal grouping in hand, we bring x_{it} back and use fixed effects model to reduce the incidental parameter problem on the common parameter β.

When T is fixed: To study the behavior of the distribution of $\{a_i(\beta)\}_{i=1}^N$ when T is comparably small. We define the following Kullback-Leibler divergences (KLD hereafter)

$$
\begin{aligned}
KDL_\infty(\beta) &= \int p(a_i(\beta)) \log \frac{p(a_i(\beta))}{p(a_i(\hat{\beta}))} da = \int p(a_i(\beta)) \log \frac{p(a_i(\beta))}{p(\alpha^0_{g_i})} da, \quad N, T \to \infty \\
KDL_T(\beta) &= \int p(a_i(\beta)) \log \frac{p(a_i(\beta))}{p(a_i(\hat{\beta}))} da, \quad N \to \infty, T \text{ is fixed} \\
KDL(T) &= \int p(a_i(\hat{\beta})) \log \frac{p(a_i(\hat{\beta}))}{p(\alpha^0_{g_i})} da, \quad N \to \infty
\end{aligned}
$$

As we know KLD can be considered as the "distance" between two distributions, and solving likelihood maximization problem is equivalent to minimizing the KLD. Thus, $KDL_\infty(\beta)$ is a function of β represents the dissimilarity between the distribution of $p(a_i(\beta))$ and the population distribution $p(\alpha^0_{g_i})$ given different value of β. From the first panel in Figure 5.2, the entire fitted curve is close to zero which implies the estimated distribution is very similar to the population distribution. The more important is the curve doesn't vary too much if we choose different β for grouping. Especially, the $KDL_\infty(\beta)$ are close when $\beta = 0$ and

$\beta = 1$(true value), this is an numerical evidence that we can artificially make $\beta = 0$ to release the information of regressor in the grouping step if necessary. $KDL_T(\beta)$ is still a function of β captures the dissimiliarity between $p(a_i(\beta))$ and $p(a_i(\hat{\beta}))$ under different T. We want to study the asymptotic performance of $p(a_i(\beta))$ under different T. In the second panel in Figure 5.2, we see the convergence of distribution as T increase. The confident interval of the fitted curve isalso shrinking because larger T helps identify the individual heterogeneities and so the distribution $p(a_i(\beta))$. We can also see the variation of $KDL_T(\beta = 0)$ and $KDL_T(\beta = 1)$ is mainly from the estimation error of $\{a_i(\beta)\}_{i=1}^N$, and it is vanishing in T. The third KLD we defined, $KDL(T)$, which is a function of T, can provide us some intuition of the convergence rate of $p(a_i(\beta))$. The result is shown in the last panel in Figure 5.2,

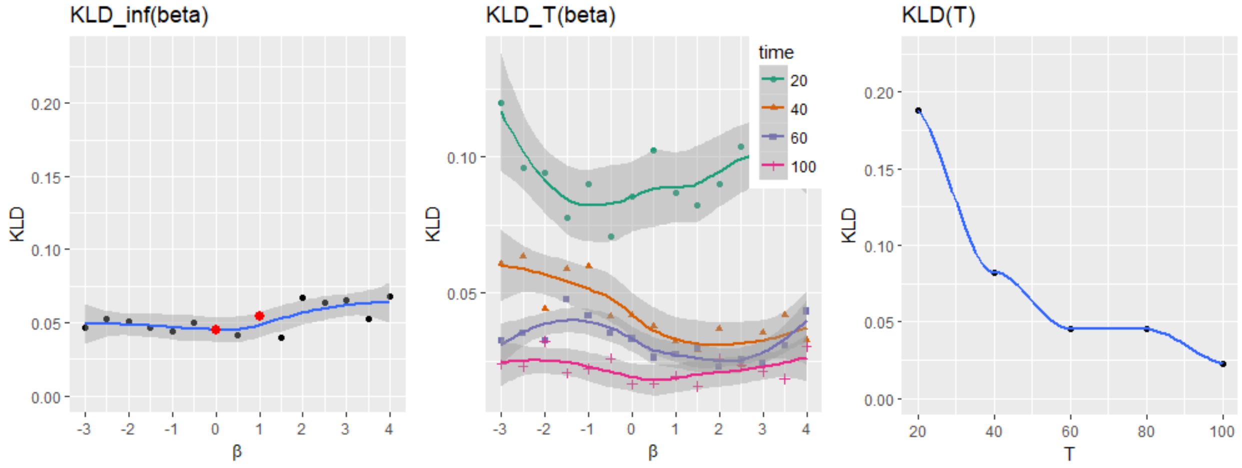


Figure 5.2: The behavior of KLD given different β or T (N=90, DGP1).

5.3 Estimation methods

The unconditional estimator (UCL) is obtained by creating a bunch of dummy variables for each group or individual which depends on the model setup. Within each cluster, common parameter and fixed effects are estimated simutaneously, so we can relax the restriction on the fixed effect. However, the price we have to pay is the inconsistency from the incidental paramiter problem and the computation cost from high dimensional Hessian matrix. The bias-corrected unconditional estimator (BCL) is UCL plus a bias-correction procedure. As it has been shown in Hahn and Newey (2004), the fixed effects estimators are still asymptotically biased even if T grows at the same rate as N, say $N/T \to \tau$. They also provide the analytical bias correction term $\hat{B}$ so that

$$\sqrt{NT}(\hat{\beta} - \beta^0) + \sqrt{\tau}(\hat{B} - B) \xrightarrow{d} N(0, \Omega)$$

For simplicity, we skip the complicated form of $\hat{B}$ but it's always avaliable in Hahn and Newey (2004). The conditional estimator (CL) is defined as the maximizer of the conditional likelihood function $l(y_{it}|\beta, \hat{\alpha}_i)$, where $\hat{\alpha}_i = 1/T \sum_{t=1}^{T} y_{it}$ is a sufficient statistic for α_i. Chamberlain (1980) shows the conditional ML estimator of β is consistent if the regularity conditions are satisfied. But it does not deliver estimates of the fixed effects and becomes computationally costly if T is large (Stammann et al. (2016))

Table 4.1 reports the MSE of common parameter β using different setting and clustering methods. The data inherent group level heterogeneities by construction. When individual fixed effect (IFE) approach is used, we see the unconditional MLE of β suffers from large estimation error when T is small due to the incidental parameter problem. After grouping, the estimation error reduces significantly. However, as T increases, IFE start to outperform grouped fixed effect (GFE) approach. This is because the bias from incidental parameter problem vanishes in T but the GFE approach suffers from the bias due to ignoring the individual heterogeneities within group. Second, the unconditional MLE using IFE approach improves remarkable after bias-correction procedure even when T is large. But bias-correction doesn't affect GFE too much because GFE is unbiased. These results are consistent with Hahn and Newey (2004) and Bester and Hansen (2016). Surprisingly, in the case of DGP 1 and DGP 2, the GFE with unknown group membership is even better than GFE with known group membership. This is because we allow individuals are different even if they are in the same group. For example, we know student A and B are in the same class, student C is in another class. However, student A is more similar to student C based

Table 5.1: MSE($\hat{\beta}$) using different methods

(a) Number of groups = 3 (DGP 1)

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0286	0.0268	0.0272	0.1757	0.0294	0.0327	0.1784	0.0274	0.0269	0.0269
90	10	0.0151	0.0145	0.0147	0.0376	0.0134	0.0145	0.1562	0.0098	0.0099	0.0100
90	20	0.0104	0.0101	0.0102	0.0115	0.0061	0.0067	0.1494	0.0056	0.0079	0.0074
90	50	0.0067	0.0065	0.0066	0.0029	0.0022	0.0023	0.1431	0.0033	0.0080	0.0061
90	100	0.0055	0.0054	0.0055	0.0013	0.0011	0.0011	0.1432	0.0027	0.0091	0.0054

(b) Number of groups = 3 (DGP 2)

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0212	0.0206	0.0208	0.1808	0.0331	0.0341	0.1619	0.0281	0.0273	0.0279
90	10	0.0102	0.0100	0.0101	0.0390	0.0140	0.0150	0.1485	0.0104	0.0113	0.0104
90	20	0.0051	0.0050	0.0050	0.0109	0.0055	0.0065	0.1386	0.0049	0.0071	0.0061
90	50	0.0022	0.0022	0.0022	0.0028	0.0022	0.0025	0.1330	0.0021	0.0042	0.0025
90	100	0.0010	0.0010	0.0010	0.0011	0.0010	0.0010	0.1320	0.0010	0.0024	0.0019

(c) Number of groups = 3 (DGP 3)

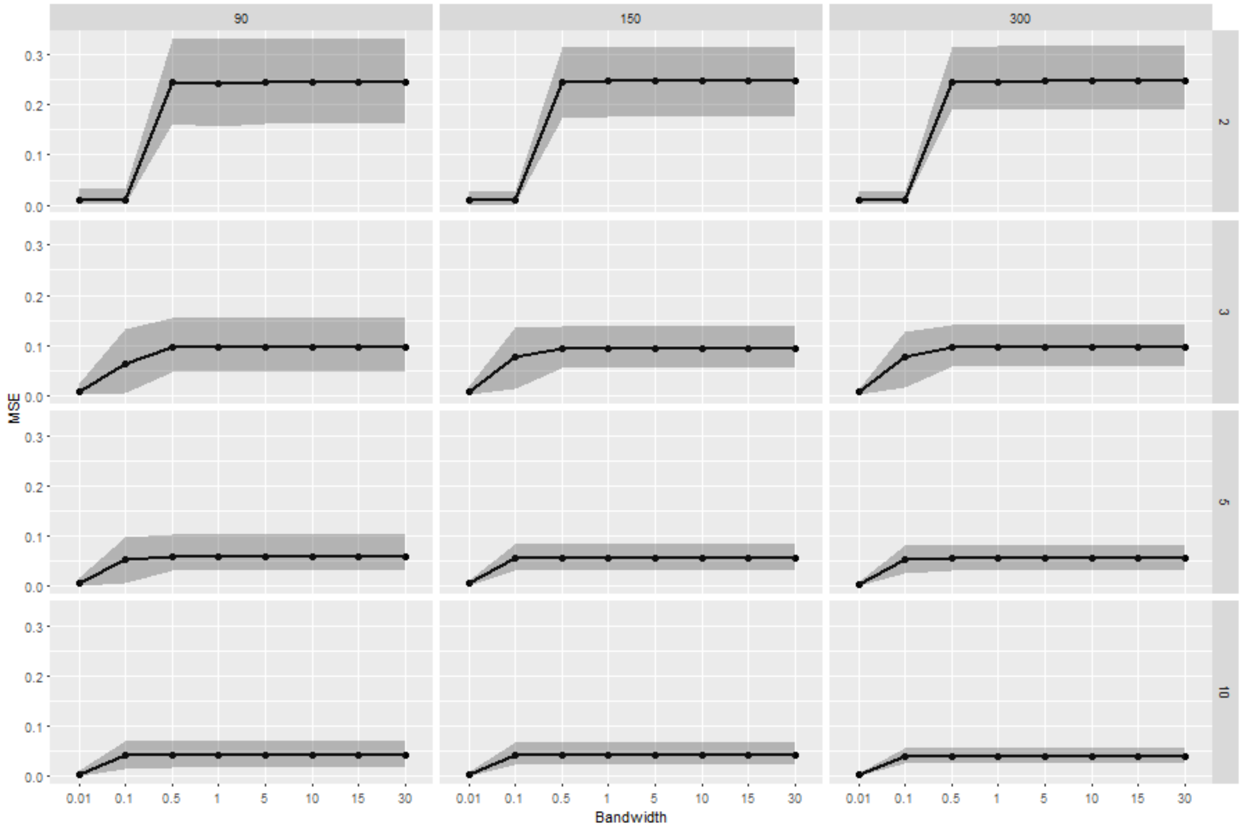
		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0562	0.0520	0.0544	0.1602	0.0296	0.0221	0.1052	0.0243	0.0244	0.0241
90	10	0.0447	0.0427	0.0437	0.0346	0.0123	0.0095	0.0930	0.0096	0.0096	0.0094
90	20	0.0404	0.0395	0.0400	0.0103	0.0054	0.0046	0.0900	0.0048	0.0080	0.0070
90	50	0.0366	0.0362	0.0365	0.0031	0.0023	0.0022	0.0859	0.0039	0.0106	0.0064
90	100	0.0342	0.0340	0.0342	0.0013	0.0011	0.0012	0.0826	0.0042	0.0147	0.0062

on the personality. In that case, the prior information on grouping may be a misleading. Fourth, all three clustering methods work well this might because the group signal is informative. Among threes methods, k-means is robust and dominant the other two methods in all cases of DGP. This is reasonable because K-means is actually lease square which is the most efficient approach. However, Mixture and MixNP can still provide acceptable results. More importantly, Mixture and MixNP can provide more information on the distribution of group-level heterogeneities. Especially, in the cases of DGP 2 and 3, the MixNP works better than Mixture since the unobervables are from distributions other than Gaussian. Lastly, pooling estimator is seriously biased due to the omitted variable bias.

In the last part of this section, we use cross-validation based on MSE to choose the optimal bandwidth in the nonparametric mixture model over different combinations of $N =$

$(90, 150, 300)$ and $G = (2, 5, 10, 30)$. T is fixed at 50 since we are grouping α_i which is time invariant. The bandwidth is selected for the set $\{0.01, 0.1, 0.5, 1, 5, 10, 15, 30\}$. From figure 1, it's clear that the smaller bandwidth is preferred among all cases. This may not be true in general because smaller bandwidth will lead to larger variance. But in terms of grouping, we assume the observations are drawn from a mixture of a certain number of component distributions. Intuitively, we prefer the components are very different from each other, but a large bandwidth will "melt" the components and hence makes it hard to identify the group membership. As a result, all the individuals will be assigned in one group (Pooling) and then lead to large MSE of the common parameter.

Figure 5.3: MSE($\hat{\beta}$) over different bandwidth(h) in mixNP method



6 Conclusion

In this paper, we extend the linear group fixed effect model with unknown group membership to nonlinear. We use the EM algorithm which is a more general method to estimate grouping. As a result, we can provide not only the group-level means but also the group-level distributions. We allow the researcher to estimate group-specific distribution nonparametrically and give a guide on choosing bandwidth. More importantly, we show the latent common parameter in the grouping object doesn't change the grouping but the magnitude of group-specific parameters in GLM. Thus, we can use some particular value to improve estimation in the grouping step. For example, when there is only plenty of time dimension data of the response variable.

The paper can be extended to several directions. For example, many empirical studies can be revisited using the unsupervised clustering method introduced in this paper, especially for the cases the prior information on grouping is not convincible. Another question that can be asked is when individuals are correlated, then finding the joint distribution of all individuals will be challenging and exciting.

References

C. C. Aggarwal and C. K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC, 1st edition, 2013.

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

T. Benaglia, D. Chauveau, D. Hunter, and D. Young. mixtools: An r package for analyzing mixture models. *Journal of Statistical Software, Articles*, 32(6):1–29, 2009a.

T. Benaglia, D. Chauveau, and D. R. Hunter. An em-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009b.

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

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

S. Bonhomme, L. Thibaut, and E. Manresa. Discretizing unobserved heterogeneity. 2017, Working Paper.

L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232, 2006.

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

J. Chen. Optimal rate of convergence for finite mixture models. *Ann. Statist.*, 23(1):221–233, 02 1995.

J. Chen. Consistency of the mle under mixture models. *Statist. Sci.*, 32(1):47–63, 02 2017.

J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18(2):443–465, 2008.

G. Ciuperca, A. Ridplfi, and J. Idier. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30(1):45–59, 2003.

N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56 (3):463–474, 1969.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

P. J. Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452, 1990.

J. Hahn and H. R. Moon. Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881, 2010.

J. Hahn and W. Newey. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319, 2004.

R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, 13(2):795–800, 06 1985.

D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 35(1):224–251, 2007.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27(4):887–906, 12 1956.

T. Lancaster. The incidental parameter problem since 1948. *Journal of Econometrics*, 95 (2):391 – 413, 2000.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier, 1994.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948.

R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.*, 9(1):225–228, 01 1981.

A. Stammann, F. Heiß, and D. McFadden. Estimating fixed effects logit models with large panel data. 2016.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, 20(4):595–601, 12 1949.

Appendix

Table 6.1: MSE($\hat{\beta}$) using different methods (DGP1)

(a) Number of groups = 2

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0354	0.0339	0.0354	0.2088	0.0369	0.0621	0.3360	0.0332	0.0336	0.0328
90	10	0.0213	0.0207	0.0213	0.0513	0.0195	0.0233	0.3159	0.0155	0.0158	0.0155
90	20	0.0123	0.0121	0.0123	0.0136	0.0069	0.0083	0.3026	0.0090	0.0104	0.0092
90	50	0.0080	0.0079	0.0080	0.0038	0.0026	0.0026	0.2972	0.0070	0.0075	0.0070
90	100	0.0060	0.0060	0.0060	0.0014	0.0012	0.0012	0.2851	0.0054	0.0065	0.0055

(b) Number of groups = 3

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0286	0.0268	0.0272	0.1757	0.0294	0.0327	0.1784	0.0274	0.0269	0.0269
90	10	0.0151	0.0145	0.0147	0.0376	0.0134	0.0145	0.1562	0.0098	0.0099	0.0100
90	20	0.0104	0.0101	0.0102	0.0115	0.0061	0.0067	0.1494	0.0056	0.0079	0.0074
90	50	0.0067	0.0065	0.0066	0.0029	0.0022	0.0023	0.1431	0.0033	0.0080	0.0061
90	100	0.0055	0.0054	0.0055	0.0013	0.0011	0.0011	0.1432	0.0027	0.0091	0.0054

(c) Number of groups = 5

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0263	0.0234	0.0239	0.1634	0.0267	0.0260	0.1228	NA	0.0218	NA
90	10	0.0165	0.0154	0.0155	0.0377	0.0134	0.0139	0.1109	0.0121	0.0117	0.0119
90	20	0.0100	0.0095	0.0096	0.0112	0.0059	0.0061	0.1025	0.0051	0.0058	0.0066
90	50	0.0061	0.0059	0.0061	0.0027	0.0021	0.0022	0.0974	0.0020	0.0044	0.0075
90	100	0.0048	0.0047	0.0048	0.0011	0.0009	0.0009	0.0961	0.0010	0.0029	0.0075

(d) Number of groups = 10

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0297	0.0239	0.0221	0.1715	0.0297	0.0271	0.0990	NA	0.0237	NA
90	10	0.0165	0.0144	0.0136	0.0354	0.0129	0.0137	0.0870	NA	0.0125	NA
90	20	0.0090	0.0081	0.0078	0.0099	0.0053	0.0060	0.0780	0.0055	0.0056	0.0065
90	50	0.0050	0.0047	0.0047	0.0025	0.0019	0.0021	0.0749	0.0020	0.0021	0.0092
90	100	0.0040	0.0039	0.0039	0.0010	0.0009	0.0009	0.0730	0.0009	0.0012	0.0086

Table 6.2: MSE($\hat{\beta}$) using different methods (DGP2)

(a) Number of groups = 2

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0278	0.0271	0.0278	0.2095	0.0365	0.0674	0.3407	0.0317	0.0332	0.0317
90	10	0.0141	0.0139	0.0141	0.0485	0.0184	0.0277	0.3165	0.0149	0.0153	0.0149
90	20	0.0055	0.0054	0.0055	0.0119	0.0059	0.0093	0.3007	0.0054	0.0067	0.0054
90	50	0.0024	0.0024	0.0024	0.0036	0.0024	0.0028	0.3016	0.0024	0.0030	0.0024
90	100	0.0013	0.0013	0.0013	0.0015	0.0014	0.0015	0.2961	0.0013	0.0019	0.0013

(b) Number of groups = 3

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0212	0.0206	0.0208	0.1808	0.0331	0.0341	0.1619	0.0281	0.0273	0.0279
90	10	0.0102	0.0100	0.0101	0.0390	0.0140	0.0150	0.1485	0.0104	0.0113	0.0104
90	20	0.0051	0.0050	0.0050	0.0109	0.0055	0.0065	0.1386	0.0049	0.0071	0.0061
90	50	0.0022	0.0022	0.0022	0.0028	0.0022	0.0025	0.1330	0.0021	0.0042	0.0025
90	100	0.0010	0.0010	0.0010	0.0011	0.0010	0.0010	0.1320	0.0010	0.0024	0.0019

(c) Number of groups = 5

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0224	0.0213	0.0212	0.1676	0.0299	0.0319	0.1077	NA	0.0271	NA
90	10	0.0103	0.0099	0.0101	0.0382	0.0120	0.0120	0.1008	0.0100	0.0096	0.0098
90	20	0.0051	0.0050	0.0050	0.0107	0.0056	0.0062	0.0899	0.0051	0.0056	0.0065
90	50	0.0017	0.0017	0.0017	0.0024	0.0018	0.0020	0.0844	0.0017	0.0026	0.0063
90	100	0.0010	0.0010	0.0010	0.0011	0.0010	0.0011	0.0826	0.0010	0.0017	0.0028

(d) Number of groups = 10

		Known grouping			Individual FE				Unknown grouping		
N	T	UCL	BCL	CL	UCL	BCL	CL	Pooling	Kmeans	Mixture	MixNP
90	5	0.0238	0.0210	0.0195	0.1615	0.0281	0.0255	0.0852	NA	0.0231	NA
90	10	0.0102	0.0097	0.0097	0.0327	0.0121	0.0138	0.0692	NA	0.0118	NA
90	20	0.0045	0.0044	0.0044	0.0091	0.0046	0.0057	0.0640	0.0050	0.0048	0.0065
90	50	0.0019	0.0018	0.0018	0.0026	0.0019	0.0020	0.0619	0.0019	0.0023	0.0094
90	100	0.0010	0.0010	0.0010	0.0011	0.0010	0.0011	0.0603	0.0010	0.0011	0.0092