Modelling latent grouped patterns with heterogeneous distributions

Erasmus School of Economics

Bachelor Thesis (Econometrics and Operational Research)

Author: Sabine Vlasblom

Student number: 497859

Supervisor: dr. W. Wang

Second assessor: dr. C. Cavicchia

Date final version: July 3, 2022

Abstract

Modelling unobserved heterogeneity is quite relevant in panel data, as it can give insightful information about grouped patterns. Current literature mainly focus on cluster methods using mean heterogeneity, but in this paper we want to model the distributional heterogeneity of the covariates effects. We try and replicate a simplified version of the multi-dimensional quantile regression from Leng et al. (2021) and use one-dimensional clustering. Moreover, we extend their research by trying to validate the claim of common group memberships. We test the quantile-invariance with the use of the clustering consensus statistic from Zhang et al. (2019). When group memberships vary across quantiles the use of a composite-quantile function in the estimation process is no longer correct. We cannot pool information across quantiles and thus need to change our estimation process. We generate data using the location-scale shift model and define a one-dimensional quantile regression model. We estimate our regression coefficients using a composite-quantile approach, for which we need to verify the quantile-invariance of the group memberships. We find that the group memberships are not always common across quantiles and this can lead to inaccurate results. Therefore, it is relevant to validate claims on group memberships to ensure accuracy of the estimation process.[1]

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of contents

1 Introduction

Panel data is widely used in economics and finance. However, we do come across the issue of unobserved heterogeneity in cross-sectional and time series dimensions. Leng et al. (2021) argue that empirical evidence in panel data applications has shown the effects of covariates often exhibits a grouped pattern of heterogeneity and a homogeneous effect within a group. Existing studies mostly focus on the mean effects and cluster units based on the mean heterogeneity. Leng et al. (2021) shift the focus to the distributional effects of covariates and aim to model multi-dimensional distributional heterogeneity using a panel quantile regression model with additive cross-section and time fixed effects. They distinguish the following two group memberships: cross-section fixed effects and slope coefficients. Moreover, Leng et al. (2021) aim to identify these two latent group structures leading to a multi-dimensional clustering problem.

In this paper we replicate the model from Leng et al. (2021) without any fixed effects and only focusing on the slope coefficients. We also focus on modelling the distributional heterogeneity of covariates and do not use the cluster based on the mean heterogeneity. Moreover, we cluster units based on multiple quantiles, rather than using single-quantile regression as in Gu and Volgushev (2019) and Zhang et al. (2019). When the group structure is invariant to quantiles, we can use the composite-quantile estimation method from Leng et al. (2021) to get more accurately estimated group memberships. The group memberships should be common over all quantiles and then the multiple-quantile approach is more accurate than the single-quantile approach.

According to Brand and Xie (2010), there are so-called inertial factors that hardly vary across the distribution of the dependent variable. Group memberships are driven by these factors and thus are likely to be quantile-invariant. A quantile-invariant group structure gives the opportunity to pool information across quantiles and this improves membership estimates, as argued by Leng et al. (2021). Moreover, when the group structure is common across conditional distribution of the dependent variable, using multiple quantiles that contain clustering information, is expected to be more accurate in classification than existing approaches. In Leng et al. (2021) the quantile-invariance of the group memberships is never validated nor substantiated. Therefore, we extend the approach of Leng et al. (2021) by verifying this assumption of common group memberships on a one-dimensional clustering problem. We introduce the following research question:

"How can we model single-dimensional latent grouped patterns with heterogeneous distributions and validate the existing estimation method?"

In order to answer the main research question, this paper will distinguish the following subquestions: (i) *how can we model and estimate grouped patterns of distributional heterogeneity of covariate effects?*; (ii) *how do we verify the quantile-invariance of group memberships?*; (iii) *how can we evaluate the finite-sample performance of the proposed method?*. For the first subquestion we will use a one-dimensional quantile regression model which is a more simpler form of the multi-dimensional group structure quantile

regression model as introduced in Leng et al. (2021). This model excludes the fixed effects from the model and focuses on using the slope coefficients to model the distributional heterogeneity. Moreover, we will use an algorithm to minimize the so-called composite-quantile function and estimate the group patterns for different dimensions. The relevance of this part of the research lies solely in the importance of more knowledge on patterns in unobserved heterogeneity. We expect our model to be useful in modelling grouped patterns of distributional heterogeneity. We contribute to the research of Leng et al. (2021) by trying to verify the quantile-invariance of the group memberships. To verify the quantile-invariance we use an interesting statistic from Zhang et al. (2019) and apply it to our model. With the results from the statistic we compare single-quantile regressions with our multiple-quantile model. Using this comparison we can answer subquestion (ii) and verify that group memberships are common among quantiles. Finally, we will evaluate the finite-sample performance of our model and the used estimation method with the use of the Monte Carlo simulation. We define two data generating processes based on the location-scale shift model. They differ in errors, the second data generating process has group specific distributions of errors. Furthermore, we evaluate the performance by comparing the misclustering frequencies and evaluate the accuracy of the slope coefficient estimates based on their bias and root mean squared error.

When we test for quantile-invariance of the group memberships we find contradicting results for the different data generating processes. The group memberships do not always seem to be common across quantiles and sometimes using a single-quantile regression gives more stable group memberships. We incorrectly use the composite-quantile function as times and get inaccurate results. However, our model performs well in estimating slope coefficients accurately, especially for the fifth quantile.

This research contributes to existing literature as we combine the works of Leng et al. (2021) and Zhang et al. (2019). We replicate a simplified version of the model in Leng et al. (2021) and use their estimation method that uses the composite-quantile approach. Furthermore, we add to the work of Leng et al. (2021) by trying to validate the claim of common group memberships with the use of Zhang et al. (2019) their clustering consensus statistic. This extension is relevant, because without common group memberships the composite-quantile approach is no longer valid. Then the estimation method is incorrectly used and we would have to change the method to estimate quantile-specific group memberships. We find that sometimes this is the case and thus verifying the assumption of common group memberships can help validate the method needed to estimate correctly.

The structure of the remaining sections is as follows, in Section 2 we discuss in detail the current literature on modelling unobserved heterogeneity in panel data. Section 3 describes how we model distributional effects of heterogeneity and how we estimate our model. Moreover, we explain how we simulate our data and evaluate the model performance. In Section 4 we present our main findings for clustering accuracy, slope coefficient accuracy and the quantile-invariance of the group memberships. Finally, Section 5 summarises our results, presents the limitations of our research and suggests further research.

2 Literature review

Over the past couple of years plenty research has been conducted based on panel data and grouped heterogeneity. In econometrics is it relevant to find underlying patterns and connections that might not be observable. There have been different approaches on how to model the unobserved heterogeneity in panel data. In addition, for these different models there have been developed different estimation methods. In this section we discuss the current literature on the modelling and estimation of unobserved heterogeneity in panel data.

As previously mentioned, our paper tries to replicate a more basic version of the Multi-dimensional Group Structure Quantile Regression (MuGS-QR) model as introduced in Leng et al. (2021). This MuGS-QR model considers panel quantile regression models with additive cross-section and time-fixed effects and allows group-specific quantile slope coefficients. Leng et al. (2021) make use of a composite-quantile check function to jointly estimate the multi-dimensional group memberships. The estimation method is similar to K-means or Lasso-based clustering algorithms, but not quite the same since the estimation method does not cluster units based on the mean or single quantiles. As a result, they find that using multiple quantiles improves the clustering accuracy of the group memberships over the existing methods in which clustering is only based on the mean or some single quantile.

An example of such a method based on mean clustering is presented in the paper of Bonhomme and Manresa (2015). They propose a flexible and parsimonious approach that will allow for the unobserved heterogeneity in panel data. The model introduced allows for clustered time patterns that are common among individuals of the groups. Bonhomme and Manresa (2015) introduce the Grouped Fixed-effects (GFE) estimator, which is based on the optimal grouping of N cross-sectional units. A least squares criterion is used to obtain the estimator and the estimation process is similar to K-means. More specifically, the paper studies the effect of income on democracy for a panel of countries, with the focus on the last part of the twentieth century. The paper concludes that there is robust evidence of heterogeneous and also group-specific paths of democratization in their data. Furthermore, the GFE approach is proven to be useful in applications where time-varying grouped effects may occur in the panel data. However, as proven by Leng et al. (2021) is it not the most optimal approach to model heterogeneity in panel data.

Next, Cheng et al. (2021) propose a multi-dimensional clustering approach to account for unobserved heterogeneity in panel data using a nonlinear framework. While one-dimensional clustering is based on separate groups, their multi-dimension approach associates each unit with multiple clusters. Cheng et al. (2021) estimate the group memberships using a cluster-specific and common parameters in the nonlinear Generalized Method of Moments (GMM) framework. They find that their multi-dimensional clustering is robust to sparse interactions among different features. Moreover, Cheng et al. (2021) mentions that multi-dimensional clustering is preferred over one-dimensional clustering, because the one-dimensional approach cuts the data finer, leaving only a few units per group of much smaller size. Furthermore, the K-means algorithm is used as an estimation method for the group memberships and it is based on a GMM criterion, whereas for example Bonhomme and Manresa (2015) uses the least squares criterion.

We also consider the work of Gu and Volgushev (2019). When time-fixed effects are absent and

the group-specific slope coefficients are cross-sectionally homogeneous, the model of Leng et al. (2021) reduces to the quantile grouped-specific fixed effects model from Gu and Volgushev (2019). The model is a linear quantile regression model that accommodates grouped fixed effects and to estimate it they use a convex clustering penalty. Gu and Volgushev (2019) propose an attractive fixed effect approach as minimal assumptions are imposed on the structure of latent effects and on the correlation between the latent effects and observed covariates. It does assume that the fixed effects have a grouped structure and use individual fixed effects while assuming common slope structures. However, the major challenge of Gu and Volgushev (2019) their approach is that is introduces a large number of parameters that grow linearly with the number of individuals.

Then we consider the work of Zhang et al. (2019), which is closely related to Gu and Volgushev (2019). They also consider panel quantile regressions with grouped fixed effects and homogeneous slope coefficients, whereas Leng et al. (2021) considered both cross-section and time fixed effects and allowing for multi-dimensional group heterogeneity. Zhang et al. (2019) developed a algorithm similar to K-means to identify subgroups with heterogeneous slopes at a single-quantile level and across multiple quantiles. Their paper proposes a clustering method for panel data using quantile regressions to identify subgroups of units with the same covariate effect. Moreover, Zhang et al. (2019) consider the assumption of common group memberships across quantiles. They introduce a stability measurement to choose the empirically optimal quantile that gives the most stable results. This measurement leads to a data-adaptive criterion and in addition improves the accuracy of identifying subgroups. In this paper, we will use their measurement and apply it to our own model in order to extend the research of Leng et al. (2021).

Su and Ju (2018) uses a model similar to Gu and Volgushev (2019) with fixed-effects and slope coefficients. In contrast, they consider a different estimation method using Penalized Principal Component (PPC) estimation. This estimation method extends the penalized-profile-likelihood-based C-Lasso method and estimates panel data with cross-section dependence. Su and Ju (2018) establish a oracle property based on the uniform classification consistency. Moreover, the paper considers Interactive Fixed-Effects (IFEs), where both the cross-sectional dimension and time series dimension pass to infinity. Following the paper of Su et al. (2016), the penalized estimations are extended with IFEs. Su and Ju (2018) find that the clustering error, root mean squared errors and biases shrink to zero as the time series dimension increases. Moreover, the post-Lasso outperforms the C-Lasso, with the expectation when the time series dimension is too small and the classification error to big.

Finally, we discuss the new model and estimation procedure for panel data from Okui and Wang (2021). Their model allows us to identify heterogeneous structural breaks in panel data. Previous detection techniques would only focus on common structural breaks, whereas the paper of Okui and Wang (2021) provides a model and estimation method to detect heterogeneous structural breaks. The model is a linear panel data model with heterogeneous and time-varying coefficients. Moreover, individual heterogeneity is modeled via a grouped pattern. A hybrid estimation procedure is introduced that combines the GFE estimator from Bonhomme and Manresa (2015) and the Adaptive Group Fused Lasso (AGFL) approach by Qian and Su (2016). The so-called Grouped AGFL (GAGFL) uses the GFE to estimate the group memberships and the AGFL estimates the break dates by minimizing a penalized least

squares objective function. The performance of model and estimation method is examined using Monte Carlo simulations and also compared to other break detection techniques.Okui and Wang (2021) find that their model performs well and if heterogeneity in breaks is ignored this would lead to inconsistent estimates of the number of breaks and also less accurate breakpoint estimates.

3 Methodology

In this section we elaborate on the setup of our model and how we plan to estimate it. In addition, we describe our extension of how we test the quantile-invariance of the group memberships. As seen in the previous section there are plenty estimation methods developed, but we will replicate the one introduced and used in Leng et al. (2021). Moreover, in this section we discuss how we simulate our data with the use of the Monte Carlo simulation. The performance evaluation is discussed at the end of this section.

3.1 Model setup

In Leng et al. (2021), the Multi-dimensional Group Structure Quantile Regression (MuGS-QR) model is introduced. We replicate a more simpler version of this model and only include the group-specific slope coefficient. Our version of the MuGS-QR model, the One-dimensional Quantile Regression (1D-QR) model is written as follows:

$$Q_\tau(y_{it}|x_{it}) = x'_{it}\beta_{g_i}(\tau), \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{1}$$

where $Q_\tau(y_{it}|x_{it})$ is the conditional τ-quantile of y_{it} given x_{it}. We define y_{it} as the scalar dependent variable of individual i observed at time t and x_{it} is a vector of p exogenous regressors. In this model we do not allow any fixed effects, but only consider the quantile regression coefficient β_{g_i}, which can be different for each group membership $g_i \in (1, ..., G)$. The number of groups, G, is assumed to be fixed and finite. We adopt the assumptions from Leng et al. (2021): (i) group structures of slope coefficients are latent, which enables the capturing heterogeneity in a wide range of applications; (ii) group membership is time-invariant and independent of range of quantiles. Thus we can have quantile-specific estimates of the parameters, but the group memberships are assumed to be quantile-invariant. In Leng et al. (2021) these group memberships are assumed to be common among quantiles and in this paper we validate this assumption on our 1D-QR model.

3.2 Clustering Consensus Statistic

We have assumed the group memberships to be quantile-invariant. However, Leng et al. (2021) never validates this claim of quantile-invariance. Therefore, we add to their research and try to substantiate these claims. Zhang et al. (2019) proposes an interesting test statistic for the quantile-invariance of group memberships, the Clustering Consensus (CC) statistic. With the use of the CC statistic we can assess the stability of clusterings and choose the most stable one. In this paper, we will compare the clustering consensus of each single quantile with the multiple-quantile approach. On our 1D-QR model we perform single-quantile regressions for each τ_k and also perform a multiple-quantile regression using all quantiles.

Following slightly different approach than in Zhang et al. (2019), we use our simulated data and consider 500 (possibly) different clusterings. How we obtain our simulated data is discussed in more detail in Section 3.4.1. Furthermore, for each combination of cross-sectional dimensions N and time series dimensions T, we calculate the CC statistic. Moreover, we repeat this for each data generating process.

For the calculation of the statistic we need to construct a $N \times N$ consensus matrix $\mathcal{M} = (\mathcal{M}(i,j))$. This matrix stores the proportion of times the pair of units i and j are clustered together among our 500 clustering results. The diagonal of this matrix will contain ones and off-diagonal elements will be ranging between 0 and 1. The CC statistic is defined as the average consensus among G groups and computed as follows:

$$CC = \frac{1}{G}\sum_{g=1}^{G} \frac{2}{N_g(N_g-1)} \sum_{i,j\in\mathcal{G}_g, i<j} \mathcal{M}(i,j), \tag{2}$$

where $\mathcal{G}_g$ denotes the set of units assigned to group g and N_g is the carnality of this set $\mathcal{G}_g$. For each quantile level τ_k we will calculate the consensus of the single quantile regressions. In addition, we compute the consensus of the multiple quantile regression.

The highest CC statistic implies the highest consensus and thus gives us the most stable clustering. We compare the CC statistics from the single-quantile regressions with the multiple-quantile regressions. When the multiple-quantile approach shows a higher consensus, related to a higher CC statistic, than the individual single quantiles, we can conclude that the group memberships are quantile-invariant. This is because using multiple quantiles leads to a higher consensus and thus the group memberships are not dependent on the individual quantiles, but are estimated more accurately when the quantiles are all used in the multiple-quantile approach.

As for the relevance of quantile-invariant group memberships we know from Leng et al. (2021) that if group memberships turn out to be quantile-specific then their framework should be adjusted to allow group structures to differ over quantiles. This can be seen as a special case of applying the estimation procedure for each quantile separately. Moreover, with common group memberships we can make use of the so-called composite-quantile check function, which uses information of all quantiles to help improve the group memberships. Leng et al. (2021) argues that using this composite approach the accuracy of the group memberships improves compared to using existing methods in which the clustering is based on the mean or single quantiles.

3.3 Estimation method

In this section we focus on the method used to estimate the parameters in our 1D-QR model. The model 1 has two types of parameters: (i) group membership variables g_i, defined for each $i = 1, ..., N$; (ii) regression quantile parameters $\beta_g(\tau)$ for each fixed quantile τ. We follow the definition of parameters from Leng et al. (2021) and define $\boldsymbol{\beta}(\tau) = (\beta_1'(\tau), ..., \beta_G'(\tau))'$ for each $\tau \in (\tau_1, ..., \tau_K)$. We collect these parameters as $\theta(\boldsymbol{\tau}) = \{\boldsymbol{\beta}(\boldsymbol{\tau})\} \in \Theta$. For the group membership variables we define the parameter $\gamma_g = (g_1, ..., g_N)$ as the partition of N individuals into G groups. We hope to show that γ_g is quantile-invariant for all quantiles τ_k, with the use of the CC statistic. In addition, we will assume G to be known and finite by using the results from Leng et al. (2021) regarding the minimization of an information criterion. We use

their resulting true number of groups throughout this paper and do not confirm this number ourselves, but solely rely on the substantiation of Leng et al. (2021).

In order to obtain the estimator of the two types of parameters, we minimize the following composite quantile function:

$$\min_{(\gamma_g,\theta(\boldsymbol{\tau}))} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \rho_{\tau_k} \left[y_{it} - x'_{it} \beta_{g_i}(\tau_k) \right], \tag{3}$$

where $\rho_\tau(u) = [\tau - I(u < 0)]u$ is the check function and we consider K equally spaced quantiles. To solve the optimization problem in Equation 3, we use the iterative algorithm from Leng et al. (2021), which is defined as follows:

Algorithm 1. Define a randomized $\gamma_g^{(0)}$ as the initial estimate of γ_g. Set s=0.

Step 1 For given ($\gamma_g^{(s)}$, estimate quantile regression parameters for each quantile τ_k.

Step 2 Given the resulting $\theta^{(s)}(\boldsymbol{\tau})$ from step 1, assign unit i to a g-group as follows:

$$g_i^{(s+1)} = \arg\min_{g \in (1,...,G)} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} \rho_{\tau_k} \left(y_{it} - x'_{it} \beta_g^{(s)}(\tau_k) \right),$$

for $i = 1, ..., N$.

Step 3 Set $s = s + 1$. Go to Step 1 until numerical convergence of $\theta(\boldsymbol{\tau})$ is reached.

This algorithm is set to iterate between estimation and clustering steps. In Step 1 we estimate the quantile regression slope coefficients, given the current group memberships estimates. We perform Step 1 in RStudio (R Core Team, 2021) with the use of the $rq()$ function. In Step 2 we estimate the next group memberships using the composite-quantile check function, which is summing the check function $\rho_{\tau_k}(\cdot)$ over all quantiles k. When the group memberships are common across the conditional distribution of the dependent variable, we can more accurately cluster observations using multiple quantiles. Moreover, our convergence condition in Step 3 is reached when the current estimate $\theta^{(s)}(\boldsymbol{\tau})$ is equal to the previous estimate $\theta^{(s-1)}(\boldsymbol{\tau})$.

3.4 Monte Carlo Simulation

The finite-sample performance of our 1D-QR model will be evaluated with data simulated with the use of data generating processes. We examine the clustering accuracy and evaluate the accuracy of the slope coefficient estimates at each quantile.

3.4.1 Data generating process

In order the simulate our data we use different data generating processes (DGPs), which have different distributions of errors. We consider two of the four DGPs as defined in Leng et al. (2021), which are both based on the location-scale shift model that is written as follows:

$$y_{it} = \beta_{g_i} x_{it} + (1 + \psi x_{it}) \epsilon_{it}, \quad g_i = 1, ..., G_0. \tag{4}$$

DGP.1: The first DGP is the model in Equation 4, where $\psi = 0.5$ and x_{it} is independently and identically generated by $\chi^2(5)$. The error term ϵ_{it} is independently and identically distributed (i.i.d.) and follows a standard normal distribution. The true number of groups G_0 are based on the results in

Leng et al. (2021). Thus we use two groups for the slope coefficients. Moreover, we fix the ratio among the g-groups such that half is in one group and the other half is in the other group. We set the slope coefficient as $(\beta_1, \beta_2)' = (-0.75, 0.75)'$.

DGP.2: The second DGP generates two different groups of heterogeneously distributed errors. The first group follows the same distribution as in DGP.1, whereas the second group follows a Weibull distribution where the theoretical mean is subtracted to make it heterogeneous. Therefore, we generate the following error term: $\epsilon_{it} \sim \big\{\text{if } g_i = 1 \text{ i.i.d. } N(0,1); \text{ if } g_i = 2 \text{ i.i.d. } Weibull(3,1) - E\left(Weibull(sh, sc)\right)\big\}$. The remaining parameters are defined and determined that same way as for DGP.1.

For both of the DGPs we define $N = (80, 160)$, which are two cross-sectional sample sizes and $T = (20, 40)$, which are two lengths of time series. Combining all possible N and T gives us four combinations of dimensions. For each combination we will generate the DGPs and replicate it 500 times.

3.4.2 Performance evaluation

We apply Algorithm 1 on our 1D-QR model using the composite quantiles $\tau \in (0.1, 0.2, ..., 0.9)$. We refer to the resulting estimates of Model 1 as single-dimensional composite-quantile clustering (1D-CQ). The performance of the 1D-CQ is evaluated on clustering accuracy and accuracy of the coefficients estimated across quantiles. We measure the clustering accuracy given the correct number of groups by taking the average of the Misclustering Frequency (MF) across all simulations. We define a MF for the slope coefficients for the 1D-CQ estimates. Let $I(\cdot)$ be the indicator function, then slope MF is computed as follows:

$$MF_G = 1 - \frac{1}{N}\sum_{i=1}^{N} I(\hat{g}_i = g_i^0), \tag{5}$$

where $\hat{g}_i$ is the estimated group of observation i and g_i^0 is the true group membership of observation i.

Moreover, the evaluation of the accuracy of the slope coefficient estimates is done at the quantiles $\tau_k \in (0.3, 0.5, 0.7)$ and is based on the bias and Root Mean Squared Error (RMSE). The bias and RMSE are computed as follows:

$$\text{Bias}(\hat{\beta}(\tau)) = \frac{1}{G}\sum_{g=1}^{G}\left[\hat{\beta}_g(\tau) - \beta_g^0(\tau)\right], \tag{6}$$

$$\text{RMSE}(\hat{\beta}(\tau)) = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left[\hat{\beta}_g(\tau) - \beta_g^0(\tau)\right]^2}, \tag{7}$$

where $\hat{\beta}_g$ is the estimated coefficient and β_g^0 is the true coefficient based on the data generation, which is defined as $\beta + \psi\epsilon$.

4 Results

In this section we provide the results for the clustering performance and the accuracy of the quantile regression estimates of our model. We present our main findings for the CC statistic first, then discuss the bias and RMSE and end with the misclustering frequencies.

4.1 Clustering consensus

The Clustering Consensus statistics for both DGP.1 and DGP.2 and for each combination of dimensions N and T are shown in Table 1. The CC statistics are computes for each single quantile and for the multiple-quantile (MQ) approach. The first block of four rows consider all dimensions for DGP.1 and the second block of four rows consider all dimensions for DGP.2.

Table 1: Clustering consensus for the single-quantile regressions and the multiple-quantile (MQ) regressions.

			Single-quantile regressions									MQ
			$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$	
DGP.1	*N=80*	*T=20*	0.8058	0.8146	0.8151	0.8155	0.8156	0.8161	0.8160	0.8126	0.8053	0.8162
		T=40	0.8151	0.8157	0.8159	0.8156	0.8167	0.8156	0.8164	0.8157	0.8152	0.8158
	N=160	*T=20*	0.8094	0.8206	0.8204	0.8244	0.8222	0.8224	0.8237	0.8207	0.8091	0.8225
		T=40	0.8215	0.8215	0.8215	0.8214	0.8214	0.8220	0.8215	0.8251	0.8236	0.8211
DGP.2	*N=80*	*T=20*	0.8148	0.8159	0.8161	0.8158	0.8153	0.8140	0.8135	0.8006	0.7484	0.8151
		T=40	0.8158	0.8151	0.8156	0.8156	0.8157	0.8160	0.8171	0.8151	0.7933	0.8147
	N=160	*T=20*	0.8244	0.8224	0.8232	0.8246	0.8233	0.8235	0.8200	0.8040	0.7536	0.8233
		T=40	0.8225	0.8234	0.8219	0.8218	0.8238	0.8238	0.8224	0.8211	0.7968	0.8224

For DGP.1 we find the highest clustering consensus among the simulation result for $N = 160$ and $T = 20$ at the single quantile level of 0.4, where the consensus is about 82.4%. Moreover, mostly we find that the quantile levels 0.1 and 0.9 perform the worst considering their consensus. However, this does not hold true for $N = 160$ and $T = 40$, where we find a consensus of 82.36% for the 0.9th quantile. Furthermore, we can conclude for $N = 80$ that the multiple-quantile regressions obtain the highest consensus. Therefore, for this cross-sectional dimension we find more stability when using multiple quantiles and thus the assumption of common group memberships holds true. Contrary to the cross-sectional dimension of $N = 160$, we we do not find the highest consensus for the multiple-quantile approach. However, the difference with the highest consensus is rather small. The difference could be due to some calculation or computational errors. Another option would be to use more simulation results to remove some of the uncertainty from the results. We could argue that for $N = 160$, either quantile-specific group memberships could be used or common group memberships.

Then for DGP.2 we find that the 0.9th quantile performs significantly worse than the other single quantiles and the multiple-quantile approach. The consensus for this quantile ranges between 75 and 79 percent. The multiple-quantile approach for DGP.2 does not seem to outperform all single-quantile regressions. The best performing quantiles do differ across the different dimensions. Therefore, we can conclude that for DGP.2 the group memberships are likely to be quantile-specific since they are more stable for specific quantiles. Therefore, we suggest using group memberships that vary for each quantile to get more accurate results and change the multiple-quantile regression estimation accordingly.

4.2 Accuracy of quantile regression estimates

Next, we examine the bias and RMSE of for both DGPs to explore the accuracy of the slope coefficients. The bias and RMSE are computed with the use of Equations 6 and 7, respectively. Table 2 presents the bias and RMSE for all dimensions and DGPs for our one-dimensional model. We do not evaluate the bias and RMSE at all quantile levels, but only consider the levels 0.3, 0.5 and 0.7 and the other quantile levels are excluded from our evaluation. When we consider the results of DGP.1 we find that the bias ranges between approximately -0.1 and 0.1. This is relatively close to zero, but the bias is closest to zero for the fifth quantile level. This to be expected since the true coefficients are equal to β at the 5th quantile value, because this is the median of a standard normal distribution, which is zero. Therefore, the smallest room for a computational error is at the fifth quantile. In addition, we find the same results for the RMSEs and we can conclude that the slope coefficient estimates at the 5th quantile are the most accurate for DGP.1. The results for DGP.2 imply that this data generating process also estimates accurate slope coefficients. The bias ranges between -0.07 and 0 and the RMSE is between 0.01 and 0.07. Once again the fifth quantile has the bias closest to zero and the smallest RMSE.

Table 2: Accuracy of slope coefficient estimates for all combinations of dimensions and the different data generating processes.

		N=80				N=160			
		T=20		T=40		T=20		T=40	
	τ	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
DGP.1	0.3	−0.0916	0.0990	−0.0939	0.0970	−0.0943	0.0978	−0.0951	0.0966
	0.5	0.0015	0.0383	−0.0003	0.0259	0.0007	0.0270	0.0005	0.0197
	0.7	0.0969	0.1022	0.0946	0.0975	0.0951	0.0981	0.0960	0.0987
DGP.2	0.3	−0.0635	0.0731	−0.0648	0.0728	−0.0621	0.0706	−0.0642	0.0717
	0.5	0.0003	0.0275	−0.0008	0.0185	−0.0001	0.0196	−0.0010	0.0134
	0.7	0.0626	0.0730	0.0618	0.0700	0.0622	0.0713	0.0623	0.0706

4.3 Clustering accuracy

We examine the clustering performance based on the results shown in Table 3. This table presents the MF for the slope coefficients from Equation 1, for each combination of N and T and each DGP. From Table 3 we see that the MFs are all around 0.2, which is quite higher than the results from Leng et al. (2021), which had the MFs ranging between 0 and 0.09. Their research found that using multiple quantiles is preffered over using single-quantile regressions. Moreover, Leng et al. (2021) also demonstrate the benefits of using composite-quantiles when clustering by comparing different models and clustering methods. Although our results give twice as high misclustering frequencies, the same trend is shown between different dimensions. When T increase, we see that the MF decreases. In addition, the highest clustering errors are found for large N and small T.

Table 3: Misclustering frequencies for each combination of dimensions and each data generating process.

	N=80		N=160	
	T=20	T=40	T=20	T=40
DGP.1	0.2225	0.2151	0.2304	0.2219
DGP.2	0.2155	0.2028	0.2375	0.2224

5 Conclusion

The modelling of distributional effects of covariates can give helpful insights on grouped patterns of unobserved heterogeneity. In this paper we answer the following research question: *"How can we model single-dimensional latent grouped patterns with heterogeneous distributions and validate the existing estimation method?"*. We use three subquestions to help us answer the research question, these subquestions are: (i) how can we model and estimate grouped patterns of distributional heterogeneity of covariate effects?; (ii) how do we verify the quantile-invariance of group memberships?; (iii) how can we evaluate the finite-sample performance of the proposed method?.

To start with our most prominent findings, we first discuss the test results for our group memberships. We find that the group memberships are not always quantile-invariant. For different data generating processes we find contradicting results. As a result, using quantile-specific group memberships in the estimation process could be preferred over using quantile-invariant group memberships. Moreover, using the incorrectly defined group memberships can affect the accuracy of our quantile regression coefficients. In this case the multiple-quantile approach would not be accurate nor is it more efficient than a single-quantile approach. In addition, we do find that using the quantiles at the end of the tails, leads to less stable clusters and thus is not preferred. Then for the quantile regression coefficient estimates, we find the most accurate coefficients at the fifth quantile. This is not surprising considering this is the median quantile and thus the true coefficient is closely related to the estimated coefficient. In addition, the third and seventh quantile seem to give quite accurate results as well, because their bias close to zero and the root mean squared error is not large. Finally, we found that our estimation process results in rather significant misclustering frequencies. Around 20 percent is misclustered, which is quite a large proportion when compared to the results from Leng et al. (2021), which had misclustering frequencies between 0 and 9 percent. However, we do find similar trends as previous works related to ours. When the time series dimension increases, the misclustering frequency descrease as well.

Considering our results, we might need to reconsider our estimation method and use quantile-specific groups to get more accurate coefficient estimates. Perhaps this could also descrease our misclustering frequency to a lower expected level. However, we do show that it is useful and insightful to test the group memberships for quantile-invariance. Since knowledge is power and is it better to know for sure that the group memberships are quantile-invariant than to assume it based on previous works. Therefore, we contribute to current literature by showing that sometimes group memberships vary per quantile and using incorrect methods, like the composite-quantile approach, can lead to inaccurate results. Thus validating this assumption beforehand gives insightful information about the correctness of the estimation

method, which then might need to be changed accordingly.

5.1 Limitations and further research

When we consider possible limitations of our research we do need to take into account the set up of our model. We try and replicate a more simpler version of the multi-dimensional model from Leng et al. (2021). However, we make use of a one-dimensional quantile regression, changing the problem to a one-dimensional clustering problem. As shown by Cheng et al. (2021), one-dimensional clustering cuts the data finer. As a result, some groups are left with only a few units and are much smaller in size. This could be the reason for some of our results. Another limitation is that we evaluate our group memberships on quantile-invariance with only one statistic. It might be beneficial to also consider other ways to evaluate the group memberships. For example, we could consider the approach taken by Dzemski et al. (2021). They consider comparing confidence sets of group membership estimates across quantiles and seeing if they are compatible.

As for further research, we suggest trying to test the group memberships for the multi-dimensional quantile regression from defined as in Leng et al. (2021). Then, we can compare the effects of the one-dimensional clustering on the group memberships with the effects of multi-dimensional clustering. In addition, we can confirm the assumption of common group memberships on their model. Moreover, one could also try to use a different estimation approach, like in Su and Ju (2018), where they use penalized principal components as estimation method.For this different estimation method we can if the quantile-invariance of the group memberships influences the estimation results.

6 Bibliography

References

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

J. E. Brand and Y. Xie. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75(2):273–302, 2010.

X. Cheng, F. Schorfheide, and P. Shao. Clustering for multi-dimensional heterogeneity. *Working paper*, 2021.

A. Dzemski, R. Okui, et al. Confidence set for group membership. *Working paper*, 2021.

J. Gu and S. Volgushev. Panel data quantile regression with grouped fixed effects. *Journal of Econometrics*, 213(1):68–91, 2019.

X. Leng, H. Chen, and W. Wang. Multi-dimensional latent group structures with heterogeneous distributions. *Journal of Econometrics*, 2021.

R. Okui and W. Wang. Heterogeneous structural breaks in panel data models. *Journal of Econometrics*, 220(2):447–473, 2021.

J. Qian and L. Su. Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics*, 191(1):86–109, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org`.

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

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

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