Journal of Econometrics 213 (2019) 68–91

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Panel data quantile regression with grouped fixed effects ☆

Jiaying Gu [a,*], Stanislav Volgushev [b]

[a] Department of Economics, University of Toronto, 150 St. George St. Toronto, M5S 3G3 Ontario, Canada
[b] Department of Statistical Sciences, University of Toronto, 100 St. George St. Toronto, M5S 3G3 Ontario, Canada

ARTICLE INFO

Article history:
Available online 8 June 2019

ABSTRACT

This paper introduces estimation methods for grouped latent heterogeneity in panel data quantile regression. We assume that the observed individuals come from a heterogeneous population with a finite number of types. The number of types and group membership is not assumed to be known in advance and is estimated by means of a convex optimization problem. We provide conditions under which group membership is estimated consistently and establish asymptotic normality of the resulting estimators. Simulations show that the method works well in finite samples when T is reasonably large.

1. Introduction

It is widely accepted in applied Econometrics that individual latent effects constitute an important feature of many economic applications. When panel data are available, a common approach is to incorporate latent structures in completely nonrestrictive way, i.e. the fixed effect approach. The fixed effect approach is attractive as it imposes minimal assumptions on the structure of the latent effects and on the correlation between the latent effects and the observed covariates and hence has become a very common empirical tool (see Hsiao (2003) for a textbook treatment).

A major challenge of the fixed effects approach lies in the fact that it introduces a large number of parameters which grows linearly with the number of individuals. For a few specific models this can be avoided by differencing out individual effects and learning about the common parameter of interest. However, for most models, including quantile regression, this simple differencing method no longer exists. The literature contains various approaches that put additional structure on latent effects in order to reduce the number of parameters and obtain more interpretable models. One popular approach is to introduce some parametric distributional structure on the latent effects, see for example Mundlak (1978), Chamberlain (1982) and the correlated random effects literature. An alternative is to assume that the fixed effects have a group structure and hence only take a few distinct values which is the approach we take in this paper.

There is ample evidence from empirical studies that it is often reasonable to consider a number of homogeneous groups (clusters) within a heterogeneous population. This discrete approach was taken by Heckman and Singer (1982) for duration analysis of unemployment spells of a heterogeneous population of workers. Bester and Hansen (2016) argue that in many applications individuals or firms are grouped naturally by some observable covariates such as classes, schools or industry codes. It is also widely accepted in the discrete choice model literature that individual agents are classified as a number of latent types (for instance Keane and Wolpin (1997) among many others).

☆ The authors would like to thank Jacob Bien for bringing the convex clustering literature to their attention. We are also grateful to two anonymous Referees and the Associate Editor whose comments helped to considerably improve the presentation of this manuscript.

* Corresponding author.
E-mail addresses: jiaying.gu@utoronto.ca (J. Gu), stanislav.volgushev@utoronto.ca (S. Volgushev).

https://doi.org/10.1016/j.jeconom.2019.04.006
0304-4076/

Estimating cluster structure has a long history in Statistics and Economics, and has generated a rich and mature literature. A general overview is given in Kaufman and Rousseeuw (2009). Among the many available clustering algorithms, the k-means algorithm (MacQueen (1967)) is one of the most popular methods. It has been successfully utilized in many economic applications, for instance Lin and Ng (2012), Bonhomme and Manresa (2015) and Ando and Bai (2016). Finite mixture models provide an alternative, likelihood based approach. In the latter, grouping is usually achieved by maximizing the likelihood of the observed data. Sun (2005) builds a multinomial logistic regression model to infer the group pattern while nonparametric finite mixture models are considered in Allman et al. (2009) and Kasahara and Shimotsu (2009) among many others.

The focus of the present paper is on quantile regression for panel data with grouped individual heterogeneity. Panel data quantile regression has recently attracted a lot of attention, and there is a rich and growing literature that proposes various approaches to dealing with individual heterogeneity in this setting. In a pioneering contribution, Koenker (2004) takes the fixed effect approach and introduces individual latent effects as location shifts. These individual effects are regularized through an ℓ_1 penalty which shrinks them towards a common value. Lamarche (2010) proposes an optimal way to choose the corresponding penalty parameter in order to optimize the asymptotic efficiency of the common parameters of the conditional quantile function, see Harding and Lamarche (2017) for an extension of this approach. Another line of work that focuses on estimating common parameters while putting no structure on individual effects includes Kato et al. (2012), Galvao and Wang (2015) and Galvao and Kato (2016). Alternative approaches have also emerged. Abrevaya and Dahl (2008) take a random effect view of these individual latent effects. They consider a correlated random-effect model in the spirit of Chamberlain (1982) where the individual effects are modeled through a linear regression of some covariates. This is further developed in Arellano and Bonhomme (2016) and Chetverikov et al. (2016) where the conditional quantile function of the unobserved heterogeneity is modeled as a function of observable covariates.[1]

Our contribution, which builds upon Koenker (2004), is a linear quantile regression method that accommodates grouped fixed effects. The advantages of our proposal over existing proposals are twofold. First, grouped fixed effects maintain the merit of unrestricted correlation between the latent effects and the observables and strike a good balance between the classical fixed effects approach and the other extreme which completely ignores latent heterogeneity. Second, in contrast to Koenker (2004), where the fixed effects are treated as nuisance parameters and are regularized to achieve a more efficient estimator for the global parameter, our method allows the researcher to learn the particular group structure of the latent effects together with common parameters of interest in the model. To the best of our knowledge, panel data quantile regression with grouped fixed effects has not been considered in the literature before. The only paper that goes in this direction is Su et al. (2016). While the general framework developed in this paper does include a version of quantile regression with smoothed quantile objective function, the theoretical analysis requires the smoothing parameter to be fixed. This results in a non-vanishing bias and hence does not correspond to quantile regression in a strict sense.

We do not assume any prior knowledge of the group structure and combine the quantile regression loss function with the recently proposed convex clustering penalty of Hocking et al. (2011). The convex clustering method introduces a ℓ_1-constraint on the pair-wise difference of the individual fixed effects, which tends to push the fixed effects into clusters. The number of clusters is controlled by a penalty parameter. The resulting optimization problem remains convex and can be solved in a fast and reliable fashion. Further modifications and a theoretical analysis of convex clustering were considered in Zhu et al. (2014), Tan and Witten (2015) and Radchenko and Mukherjee (2017). All of those authors combine ℓ_1 penalties with the classical ℓ_2 loss, and only consider clustering for cross-sectional data. Their theoretical results are not directly applicable to panel data or the non-smooth quantile loss function which is the main objective in this paper (all of the available theoretical results explicitly make use of the differentiability of the ℓ_2 loss function in their proofs).

Our main theoretical contribution is to show consistency of the estimated grouping for a suitable range of penalty parameters when n and T tend to infinity jointly. We also propose a completely data-driven information criterion that facilitates the practical implementation of the method and prove its consistency for group selection as well as asymptotic normality of the resulting parameter estimators.

The remaining part of this paper is organized as follows. Section 2 contains a detailed description of the proposed methodology and provides details on its practical implementation. Assumptions and theoretical results are included in Section 3. Section 4 presents the convex optimization problem and its computational details. Monte Carlo simulation results are included in Section 5 where we investigate the final sample behavior of the proposed methodology. All proofs are collected in Section 7. In the online Appendix, we apply the method to an empirical application in studying the effect of the adoption of Right-to-Carry concealed weapon law on violent crime rate using a panel data of 51 U.S. states from 1977–2010. Some additional simulation results are also reported.

2. Methodology

Assume that for individuals $i = 1, \dots, n$ we observe repeated measures $(X_{it}, Y_{it})_{t=1,\dots,T}$ where X_{it} denote covariates and Y_{it} are responses.[2] We shall maintain the assumption that data are i.i.d. within individuals and independent across

[1] For related literature on non-separable panel data models see Evdokimov (2010), Chernozhukov et al. (2015) and the references therein.

[2] Here, T is assumed to be the same across individuals for notational simplicity. All results that follow can be extended to individual-specific values of T_i as long as the ratio $(\max_{i=1,\dots,n} T_i)/(\min_{i=1,\dots,n} T_i)$ is uniformly bounded. In this case the theory goes through without changes if all instances of T are replaced by $n^{-1} \sum_{i=1,\dots,n} T_i$

individuals. The main object of interest in this paper is the conditional τ-quantile function of Y_{i1} given X_{i1}, which we will denote by $q_{i,\tau}$. We assume that $q_{i,\tau}$ is of the form

$$q_{i,\tau}(x) = \beta_0(\tau)^\top x + \alpha_{0i}(\tau), \quad i = 1, \ldots, n$$

with individual fixed effects $\alpha_{0i}(\tau)$ taking only a finite number, say K, of different values, say $\alpha_{(01)}(\tau), \ldots, \alpha_{(0K)}(\tau)$.[3] We explicitly allow the group membership, and even the number of groups to be unknown and to depend on τ but will not stress this dependence in the notation for the sake of simplicity. Our main objective is to jointly estimate the number of groups, unknown group structure, and parameters $\alpha_{(01)}, \ldots, \alpha_{(0K)}, \beta_0$ from the observations. To achieve this, we consider penalized estimators of the form

$$(\hat\alpha_1, \ldots, \hat\alpha_n, \hat\beta) := \arg\min_{\alpha_1,\ldots,\alpha_n,\beta} \Theta(\alpha_1, \ldots, \alpha_n, \beta) \tag{1}$$

where[4]

$$\Theta(\alpha_1, \ldots, \alpha_n, \beta) := \sum_{i,t} \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_i) + \sum_{i\neq j} \lambda_{ij}|\alpha_i - \alpha_j|.$$

Here ρ_τ denotes the usual 'check function' and the weights $\lambda_{i,j}$ are allowed to depend on n, T and the data; one particular choice is discussed in below. The form of the penalty is motivated by the work of Hocking et al. (2011). Intuitively, large values of λ_{ij} will push different coefficients closer together and result in clustered structure of the estimators $\hat\alpha_i$. High-level conditions on the weights $\lambda_{i,j}$ which guarantee consistency of the resulting grouping procedure are provided in Theorem 3.1.

There are various possible choices for the penalty parameters $\lambda_{i,j}$. We propose to use weights of the form

$$\check\lambda_{ij} := \lambda|\check\alpha_i - \check\alpha_j|^{-2} \tag{2}$$

where $(\check\alpha_1, \ldots, \check\alpha_n)$ are the fixed effects quantile regression estimators

$$(\check\alpha_1, \ldots, \check\alpha_n, \check\beta) := \arg\min_{\alpha_1,\ldots,\alpha_n,\beta} \sum_{i,t} \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_i) \tag{3}$$

of Kato et al. (2012)[5] and λ is a tuning parameter. This form of weighting by preliminary estimators is motivated by the work of Zou (2006) on adaptive lasso. Intuitively, weighting by preliminary estimated distances tends to give smaller penalties to coefficients from different groups thus reducing some of the bias that is typically present in the classical lasso.

Given the developments above, it remains to find a value for the tuning parameter λ. The high-level results in Theorem 3.1 together with findings in Kato et al. (2012) provide a theoretical range for those values (see the discussion following Theorem 3.1 for additional details), but this range is not directly useful in practice since only rates and not constants are provided. Moreover, despite the fact that the weights $\check\lambda_{ij} := \lambda|\check\alpha_i - \check\alpha_j|^{-2}$ lead to asymptotically unbiased estimates, bias can still be a problem in finite samples. A typical approach in the literature to reduce bias which results from lasso-type penalties is to view the lasso problem solution as a candidate model (in our case, a candidate grouping of α_i) and re-fit based on this candidate model (see Belloni and Chernozhukov (2009) or Su et al. (2016) among many others).

To deal with bias issues and the choice of λ in practice, we propose to combine the re-fitting idea with a simple information criterion which will simultaneously reduce the bias problem and provide a simple way to select a final model. A formal description of our approach is given in Algorithm 1.

Theorem 3.2 provides a formal justification of Algorithm 1 under high-level conditions on the tuning parameters $\hat C$, $p_{n,T}$. In particular we prove that the group structure is estimated consistently with probability tending to one and derive the asymptotic distribution of the resulting estimators $\hat\alpha_i^{IC}$, $\hat\beta^{IC}$. In order to make the proposed estimation procedure fully data-driven, we need to specify a choice for the tuning parameters $\hat C$ and $p_{n,T}$. In our simulations, we found that the following choices lead to good results[6]:

$$p_{n,T} = nT^{1/4}/10, \quad \hat C := \tau(1-\tau)\hat s(\tau) \tag{4}$$

[3] We follow Koenker (2004) in treating the α_i as fixed parameters. An alternative approach which leads to equivalent results is to treat the α_i as random (with no restrictions placed on the dependence with X_{it}). In this case the model can be written as $Q_{Y_{it}|X_{it},\alpha_i(\tau)}(\tau) = \beta_0(\tau)^\top x + \alpha_{0i}(\tau)$; here $Q_{Y_{it}|X_{it},\alpha_i(\tau)}$ denotes the conditional quantile function of Y_{it} given $(X_{it}, \alpha_i(\tau))$ (see for instance Kato et al. (2012), Galvao and Wang (2015) and Galvao and Kato (2016) for this interpretation). Both interpretations lead to the same asymptotic results.

[4] As pointed out by a Referee, one could also consider combining the objective functions corresponding to several quantiles as was done in Koenker (2004) and force all coefficients α_i to be independent of τ. This would result in efficiency gains if all α_i are purely location-shift effects but can introduce bias otherwise. We leave this extension for future research.

[5] As pointed out by a Referee, an alternative approach to obtain preliminary estimators for α_{i0} would be to run separate quantile regressions for each individual. This did not improve the performance of our procedure in the simulations that we tried.

[6] The exact constant 1/10 in the factor $p_{n,T}$ does not matter asymptotically. The value 1/10 was found to work well for a wide range of values of n, T and for various models, details are provided in the Monte Carlo Section 5.2 There we also show that the impact of the precise form of the factor in $\hat p_{n,T}$ becomes less pronounced as T increases.

input : Data (X_{it}, Y_{it}), grid of values $\lambda_1, ..., \lambda_L$, quantile level of interest τ
output: Estimated number of groups $\hat{K}^{IC}$, estimated group membership $\hat{I}_1^{IC}, ..., \hat{I}_{\hat{K}}^{IC}$, estimated coefficients $\hat{\alpha}_k^{IC}$, $\hat{\beta}^{IC}$
for $i \leftarrow 1$ **to** n **do**
  compute $\check{\alpha}_i$ given in (3)
end
for $l \leftarrow 1$ **to** L **do**
  Compute

$$(\hat{\alpha}_{1,\ell}, ..., \hat{\alpha}_{n,\ell}, \hat{\beta}_\ell) := \underset{(\alpha_1,...,\alpha_n,\beta)}{\operatorname{argmin}} \left\{ \sum_{i,t} \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_i) + \lambda_\ell \sum_{i \neq j} \frac{|\alpha_i - \alpha_j|}{|\check{\alpha}_i - \check{\alpha}_j|^2} \right\}$$

  Let $\hat{\alpha}_{(1,\ell)} < ... < \hat{\alpha}_{(\hat{K}_\ell,\ell)}$ denote the unique values of $\hat{\alpha}_{1,\ell}, ..., \hat{\alpha}_{n,\ell}$, and define $\hat{I}_{j,\ell} := \{i : \hat{\alpha}_i = \hat{\alpha}_{(j,\ell)}\}$ as the estimated groups. Compute re-fitted estimators

$$(\widetilde{\alpha}_{1,\ell}, ..., \widetilde{\alpha}_{\hat{K}_\ell,\ell}, \widetilde{\beta}_\ell) := \underset{(\alpha_1,...,\alpha_{\hat{K}_\ell},\beta)}{\operatorname{argmin}} \sum_{k=1}^{\hat{K}_\ell} \sum_{i \in \hat{I}_{k,\ell}} \sum_t \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_k).$$

  Compute the IC criterion

$$IC(\ell) := \sum_{k=1}^{\hat{K}_\ell} \sum_{i \in \hat{I}_{k,\ell}} \sum_t \rho_\tau(Y_{it} - X_{it}^\top \widetilde{\beta}_\ell - \widetilde{\alpha}_{k,\ell}) + \hat{C} \hat{K}_\ell p_{n,T},$$

  where the choice of $\hat{C}$ and $p_{n,T}$ is given in (4).
end

Set $\hat{\ell}^{IC} := \operatorname{argmin}_{\ell=1,...,L} IC(\ell)$ and denote by $\hat{K}^{IC} := \hat{K}_{\hat{\ell}^{IC}}$ the corresponding number of groups. Set $\hat{I}_k^{IC} := \hat{I}_{k,\hat{\ell}^{IC}}$, $\hat{\alpha}_k^{IC} := \widetilde{\alpha}_{1,\hat{\ell}^{IC}}$, $\hat{\beta}^{IC} := \widetilde{\beta}_{\hat{\ell}^{IC}}$.

Algorithm 1: Grouping via IC criterion

with

$$\hat{s}(\tau) := (\hat{F}^{-1}(\tau + h_{n,T}) - \hat{F}^{-1}(\tau - h_{n,T}))/(2h_{n,T})$$

where $\hat{F}(y) := \frac{1}{nT} \sum_{i,t} I\{Y_{it} - X_{it}^\top \check{\beta} - \check{\alpha}_i \leq y\}$ denotes the empirical cdf of the regression residuals from the fixed effects quantile regression estimator given in (3), $\hat{F}^{-1}$ denotes the corresponding empirical quantile function, and $h_{n,T} \to 0$ is a bandwidth parameter (we use the Hall–Sheather rule in our simulations, see Koenker (2005)).

To motivate this particular choice of constant $\hat{C}$, observe the following expansion, which is derived in detail in the proof of Theorem 3.2

$$\sum_{i,t} \rho_\tau(Y_{it} - X_{it}^\top \check{\beta} - \check{\alpha}_i) - \rho_\tau(Y_{it} - X_{it}^\top \beta_0 - \alpha_{0i}) = -\sum_i \frac{\tau(1-\tau)}{2\mathbb{E}[f_{Y_{i1}|X_{i1}}(q_{i,\tau}(X_{i1})|X_{i1})]} + o_P(n).$$

This shows that plugging in the estimated (by fixed effects quantile regression) instead of true errors underestimates the objective function evaluated at the residuals by roughly the first term on the right-hand side in the above expression. This term needs to be dominated by the penalty if we want to avoid selecting models that are too large, and so it is natural to scale the penalty by a constant which is proportional to $\sum_i 1/\mathbb{E}[f_{Y_{i1}|X_{i1}}(q_{i,\tau}(X_{i1})|X_{i1})]$ in order to ensure reasonable performance across different data generating processes. Under the simplifying assumption that $f_{Y_{i1}|X_{i1}}(q_{i,\tau}(X_{i1})|X_{i1}) =: f_\varepsilon(0)$ does not depend on i, X_{i1}, this term equals $n/f_\varepsilon(0)$, and under the same assumptions $\hat{s}$ provides a consistent estimator for the latter, see Koenker (2005). Note that the sparsity term introduced here plays a similar role as the noise variance in classical information criteria such as AIC and BIC in least squares regression.

3. Theoretical analysis

In this section we provide a theoretical analysis of the methodology proposed in Section 2. We begin by stating an assumption on the true (but unknown) underlying group structure.

(C) For each quantile τ of interest, there exists a fixed number K_τ, values $\alpha_{(01)}(\tau) < \cdots < \alpha_{(0K_\tau)}(\tau)$ and disjoint sets $I_1(\tau), \ldots, I_{K_\tau}(\tau)$ with $\cup_k I_k(\tau) = \{1, \ldots, n\}$, $|I_k(\tau)|/n \to \mu_k(\tau) \in (0, 1)$, $\alpha_{0i}(\tau) = \alpha_{0j}(\tau) = \alpha_{(0k)}(\tau)$ for $i, j \in I_k(\tau)$. There

exists $\varepsilon_0 > 0$ independent of τ with

$$\min_{k=1,\dots,K_\tau-1} |\alpha_{(0k)}(\tau) - \alpha_{(0k+1)}(\tau)| \geq \varepsilon_0.$$

Assumption (C) implies that the individual fixed effects are grouped into K distinct groups and that the group specific coefficients $\alpha_{(0j)}(\tau)$ are separated. Note that the number of groups as well as group membership is allowed to differ across quantiles. For the sake of a concise notation, the dependence of the number of groups and group centers on τ will from now on be dropped unless there is risk of confusion. Note also that we require the number of groups to be fixed (i.e. independent of n, T and non-random) and exogenous, i.e. independent of the covariates X_{it}.

Next we collect some technical assumptions on the data generating process. Define $Z_{it}^\top = (1, X_{it}^\top)$ and let $\mathcal{Z}$ denote the support of Z_{it}.

(A1) Assume that $\sup_i \|Z_{it}\| \leq M < \infty$ a.s. and that

$$c_\lambda \leq \inf_i \lambda_{\min}(\mathbb{E}[Z_{it}Z_{it}^\top]) \leq \sup_i \lambda_{\max}(\mathbb{E}[Z_{it}Z_{it}^\top]) \leq C_\lambda$$

for some fixed constants $c_\lambda > 0$ and $C_\lambda < \infty$.

(A2) The conditional distribution functions $F_{Y_{i1}|Z_{i1}}(y|z)$ are twice differentiable w.r.t. y, with the corresponding derivatives $f_{Y_{i1}|Z_{i1}}(y|z)$ and $f'_{Y_{i1}|Z_{i1}}(y|z)$. Assume that

$$f_{max} := \sup_i \sup_{y\in\mathbb{R}, z\in\mathcal{Z}} |f_{Y_{i1}|Z_{i1}}(y|z)| < \infty, \qquad \overline{f'} := \sup_{y\in\mathbb{R}, z\in\mathcal{Z}} |f'_{Y_{i1}|Z_{i1}}(y|z)| < \infty.$$

(A3) Denote by $\mathcal{T}$ an open neighborhood of τ. Assume that there exists a constant $f_{\min} \leq f_{max}$ such that

$$0 < f_{\min} \leq \inf_i \inf_{\eta\in\mathcal{T}} \inf_{z\in\mathcal{Z}} f_{Y_{i1}|Z_{i1}}(q_{i,\eta}(z)|z).$$

Assumptions (A1)–(A3) are fairly standard and routinely imposed in the quantile regression literature. Similar assumptions have been made, for instance in Kato et al. (2012) [see assumptions (B1)–(B3) in that paper].

3.1. Analysis of the group structure estimators in (1)

To state our first main result on the asymptotically correct grouping define[7]

$$\Lambda_D := \sup_{i\in I_k, j\in I_{k'}, k\neq k'} \lambda_{i,j}, \quad \Lambda_S := \inf_k \inf_{i,j\in I_k} \lambda_{i,j}.$$

In words, Λ_D corresponds to the largest penalty corresponding to the difference between two individual effects from different groups while Λ_S describes the smallest penalty between two effects from the same group. Our first result provides high-level conditions on Λ_S, Λ_D that guarantee asymptotically correct grouping.

Theorem 3.1. *Let assumptions (A1)–(A3), (C) hold and assume that* $\min(n, T) \to \infty$, $\log n = o(T)$ *and*

$$\frac{\Lambda_D}{\Lambda_S} = o_P(1), \quad n\Lambda_D = o_P(T^{1/2}), \quad \frac{T^{3/4}(\log n)^{3/4}}{n\Lambda_S} = o_P(1). \tag{5}$$

Denote the ordered unique values of $\hat\alpha_1, \dots, \hat\alpha_n$ *by* $\hat\alpha_{(1)} < \cdots < \hat\alpha_{(\hat K)}$ *(i.e.* $\hat K$ *denotes the number of distinct values taken by* $\hat\alpha_1, \dots, \hat\alpha_n$ *which we interpret as the estimated number of groups) and define the sets* $\hat I_k := \{i : \hat\alpha_i = \hat\alpha_{(k)}\}$, $k = 1, \dots, \hat K$. *Then*

$$P\left(\hat K = K, \hat I_k = I_k, k = 1, \dots, K\right) \to 1.$$

Next we discuss the implications of this general result for the specific choice $\check\lambda_{i,j}$ given in (2). Define

$$\check\Lambda_D := \sup_{i\in I_k, j\in I_{k'}, k\neq k'} \check\lambda_{i,j}, \quad \check\Lambda_S := \inf_k \inf_{i,j\in I_k} \check\lambda_{i,j}.$$

From Kato et al. (2012)[8] we obtain the bound

$$\sup_{i=1,\dots,n} |\check\alpha_i - \alpha_{0i}| = O_P((\log n)^{1/2}/T^{1/2}).$$

[7] Here, we explicitly allow the values of $\lambda_{i,j}$ to be random variables that depend on the original sample. This is needed since the choice of values for $\lambda_{i,j}$ discussed in (2) and used throughout this paper is based on the first step estimators given in (3).

[8] More precisely, from the paragraph following equation (A.14) in Kato et al. (2012); note that this result is derived under the assumption that n grows at most polynomially with T.

Now if $i, j \in I_k$ then $\alpha_{0i} = \alpha_{0j}$ and thus

$$1/\check{\Lambda}_S = \left\{ \inf_k \inf_{i,j \in I_k} |\check{\alpha}_i - \check{\alpha}_j|^{-2} \lambda \right\}^{-1} = \lambda^{-1} \sup_k \sup_{i,j \in I_k} |\check{\alpha}_i - \check{\alpha}_j|^2 = O_P\left(\frac{\log n}{\lambda T} \right).$$

Moreover, under (C) we have

$$\inf_{k \neq k'} \inf_{i \in I_k, j \in I_{k'}} |\alpha_{0i} - \alpha_{0j}| \geq \varepsilon_0 > 0$$

and thus

$$\check{\Lambda}_D \leq \lambda \left\{ \inf_{k \neq k'} \inf_{i \in I_k, j \in I_{k'}} |\check{\alpha}_i - \check{\alpha}_j| \right\}^{-2} \leq \lambda/(\varepsilon_0 - o_P(1))^2 = O_P(\lambda).$$

Given this choice of weights, the condition $\frac{\Lambda_D}{\Lambda_S} = o_P(1)$ is satisfied provided that $T/\log n \to \infty$. The other conditions in (5) take the form

$$T^{1/2} \gg n\lambda \gg T^{-1/4}(\log n)^{7/4}.$$

Assuming that $(\log n)^{7/3} = o(T)$, this provides a range of possible values for λ which will ensure that (5) holds.

3.2. Analysis of the information criterion in Algorithm 1

In this section we provide theoretical guarantees for the performance of the information criterion based estimators $\hat{\beta}^{IC}, \hat{\alpha}_k^{IC}, \hat{I}_k^{IC}, \hat{K}^{IC}$ Our main result shows that, under fairly general conditions on the penalty parameter $p_{n,T}$, the procedure described in Algorithm 1 selects the correct number of groups with probability tending to one. Moreover, the estimators $(\hat{\alpha}_1^{IC}, \ldots, \hat{\alpha}_{\hat{K}^{IC}}^{IC}, \hat{\beta}^{IC})$ are shown to enjoy the 'oracle property', i.e. they have the same asymptotic distribution as estimators which are based on the true (but unknown) grouping of individuals. Before making this statement more formal, we need some additional notation. Let

$$(\hat{\alpha}_{(1)}^{(OR)}, \ldots, \hat{\alpha}_{(K)}^{(OR)}, \hat{\beta}^{(OR)}) := \underset{(\alpha_1, \ldots, \alpha_K, \beta)}{\arg\min} \sum_k \sum_{i \in I_k} \sum_t \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_k) \tag{6}$$

denote the infeasible 'oracle' which uses the true group membership. The asymptotic variance of the oracle estimator is conveniently expressed in terms of the following two limits which we assume to exist[9]

$$\Sigma_{1,\tau} := \tau(1-\tau) \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \mathbb{E}[\tilde{Z}_{ik} \tilde{Z}_{ik}^\top],$$

$$\Sigma_{2,\tau} := \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \mathbb{E}[\tilde{Z}_{ik} \tilde{Z}_{ik}^\top f_{Y_{i1}|X_{i1}}(q_{i,\tau}(X_{i1})|X_{i1})],$$

where $\tilde{Z}_{ik} := (e_k^\top, X_{i1}^\top)^\top$ and e_k denotes the k'th unit vector in $\mathbb{R}^K$. Additionally, we need the following condition on the grid $\lambda_1, \ldots, \lambda_L$

(G) For each (n, T), denote the grid values by $\lambda_{1,n,T}, \ldots, \lambda_{L,n,T}$ where L can depend on n, T. There exists a sequence j_n such that $T^{-1/2}(\log n) \ll n\lambda_{j_n} \ll T^{1/2}$.

Assumption (G) is fairly mild. It only requires that among the candidate values for λ there exists one value so that $\check{\lambda}_{i,j}$ satisfies the assumptions of Theorem 3.1. In practice, we recommend choosing a grid of values that results in sufficiently many different numbers of groups.

Theorem 3.2. *Let assumptions (A1)–(A3), (C), (G) hold and assume that* $\min(n, T) \to \infty$ *and n grows at most polynomially in T (i.e.* $n = O(T^b)$ *for some* $b < \infty$*) and* $\frac{(\log T)^3 (\log n)^2}{T} \to 0$*. Assume that there exists* $\varepsilon > 0$ *such that* $\hat{C} > \varepsilon$ *with probability tending to one and that* $nT \gg p_{n,T} \gg n$, $\hat{C} = O_P(1)$*. Then* $P(\hat{K}^{IC} = K) \to 1$ *and*[10]

$$\sqrt{nT} \left((\hat{\alpha}_1^{IC}, \ldots, \hat{\alpha}_{\hat{K}^{IC}}^{IC}, (\hat{\beta}^{IC})^\top) - (\alpha_{(01)}, \ldots, \alpha_{(0K)}, \beta_0^\top) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_{2,\tau}^{-1} \Sigma_{1,\tau} \Sigma_{2,\tau}^{-1}),$$

$$\sqrt{nT} \left((\hat{\alpha}_{(1)}^{(OR)}, \ldots, \hat{\alpha}_{(K)}^{(OR)}, (\hat{\beta}^{(OR)})^\top) - (\alpha_{(01)}, \ldots, \alpha_{(0K)}, \beta_0^\top) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_{2,\tau}^{-1} \Sigma_{1,\tau} \Sigma_{2,\tau}^{-1}).$$

[9] Appropriate forms of asymptotic normality of the oracle and IC estimators continue to hold without assuming that the limits exist. This assumption is made for notational convenience.

[10] Strictly speaking, $\hat{\alpha}_k^{IC}$ is not defined if $\hat{K}^{IC} < K$. Since the probability of this event tends to zero, we can simply define $\hat{\alpha}_k^{IC} = 0$ for $\hat{K}^{IC} > k \geq K$.

Remark 3.3. Theorem 3.2 and Theorem 3.1 hold point-wise in the parameter space, and we expect that deriving a similar result uniformly in the parameter space (in particular, if cluster centers are allowed to depend on n, T and if their separation is lost, see Leeb and Pötscher (2008) for such findings in the context of classical lasso penalized regression) is impossible. It is a well established fact in the Statistics and Econometrics literature that inference which is based on such 'point-wise' asymptotic results can be unreliable. Recently, several approaches to alleviate this problem and achieve uniformly valid post-regularization inference have been proposed (see, among others, Belloni et al. (2014), Lockhart et al. (2014) and van de Geer et al. (2014)). Applying similar ideas to the present setting is a very important question which we leave for future research.

4. Details on the optimization problem in Algorithm 1

To implement the proposed quantile panel data regression with group fixed effect, we need to solve the optimization problem stated in (1). A natural normalization of the objective and the penalty function leads to

$$\min_{\alpha_1,\ldots,\alpha_n,\boldsymbol{\beta}} \frac{1}{nT}\sum_{i,t}\rho_\tau(Y_{it}-X_{it}^\top\boldsymbol{\beta}-\alpha_i)+\frac{\tilde{\lambda}}{n(n-1)}\sum_{i\neq j}\frac{|\alpha_i-\alpha_j|}{|\check{\alpha}_i-\check{\alpha}_j|^2} \tag{7}$$

this is equivalent to the objective (1) except that $\tilde{\lambda}$ is adjusted according to n and T so that we can use a generic grid for $\tilde{\lambda}$ rather than letting the grid support change with (n, T). In practice, the grid support of $\tilde{\lambda} \in \{0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_\ell, \ldots, \tilde{\lambda}_L\}$ is chosen such that the number of distinct values of the solution $\{\hat{\alpha}_{1,\ell}, \ldots, \hat{\alpha}_{n,\ell}\}$ for $\ell = 1, \ldots, L$ takes all possible integer values in the set $\{1, 2, \ldots, n\}$. This is always achievable as long as the grid width of $\tilde{\lambda}$ is small enough. Since for each fixed $\tilde{\lambda}_\ell$, (7) is a linear programming problem which can be efficiently solved in any reliable solvers, this is not computationally expensive.

To make this section self contained, we provide some details of the primal problem stated in (7) and its corresponding dual problem. Define $\lambda_{ij} := |\check{\alpha}_i - \check{\alpha}_j|^{-2}$ and observe that we can re-write $|\alpha_i - \alpha_j| = 2\left(\frac{1}{2}-1\{0 \le \alpha_i-\alpha_j\}\right)\Big(0-(\alpha_i-\alpha_j)\Big)$. With this notation (7) can be equivalently expressed as follows

$$\min_{\mathbf{u},\mathbf{v},\mathbf{w_1},\mathbf{w_2},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad \frac{\tau}{nT}\sum_{i,t}u_{it}+\frac{(1-\tau)}{nT}\sum_{i,t}v_{it}+\frac{4\tilde{\lambda}}{n(n-1)}\left(\frac{1}{2}\sum_{j=1}^{n(n-1)/2}w_{1j}+\frac{1}{2}\sum_{j=1}^{n(n-1)/2}w_{2j}\right)$$

subject to

$$\begin{aligned}
u_{it} &= \max\{Y_{it}-X_{it}^\top\boldsymbol{\beta}-\alpha_i, 0\}\\
v_{it} &= \max\{X_{it}^\top\boldsymbol{\beta}+\alpha_i-Y_{it}, 0\}\\
w_{1j} &= \max\{-\theta_j, 0\}\\
w_{2j} &= \max\{\theta_j, 0\}\\
Y_{it} &= u_{it}-v_{it}+\alpha_i+X_{it}^\top\beta\\
0 &= w_{1j}-w_{2j}+\theta_j
\end{aligned}$$

with $\boldsymbol{\theta}$ being a vector of length $\frac{n(n-1)}{2}$ that consists of entries $(\alpha_i - \alpha_j)\lambda_{ij}$ for $i < j$. We can represent $\boldsymbol{\theta}$ as $A\boldsymbol{\alpha}$ where A is a $\frac{n(n-1)}{2} \times n$ matrix taking the form

$$A=\begin{pmatrix}
\lambda_{12} & -\lambda_{12} & 0 & 0 & \ldots & 0 & 0\\
\lambda_{13} & 0 & -\lambda_{13} & 0 & \ldots & 0 & 0\\
 & & & \ldots & & & \\
\lambda_{1n} & 0 & 0 & 0 & \ldots & 0 & -\lambda_{1n}\\
0 & \lambda_{23} & -\lambda_{23} & 0 & \ldots & 0 & 0\\
 & & & \ldots & & & \\
 & & & & & & \\
0 & 0 & 0 & 0 & \ldots & \lambda_{n-1,n} & -\lambda_{n-1,n}
\end{pmatrix}$$

The corresponding dual problem of (7) can be stated as:

$$\begin{aligned}
&\max_{\boldsymbol{a_1},\boldsymbol{a_2}} \ \boldsymbol{a_1}^\top Y \qquad \text{subject to}\\
&X^\top\boldsymbol{a_1} = (1-\tau)X^\top\mathbf{1}_{nT}\\
&Z^\top\boldsymbol{a_1}+\frac{4nT\tilde{\lambda}}{n(n-1)}A^\top\boldsymbol{a_2} = (1-\tau)Z^\top\mathbf{1}_{nT}+\frac{2nT\tilde{\lambda}}{n(n-1)}A^\top\mathbf{1}_{n(n-1)/2}
\end{aligned}$$

with Z being the incidence matrix that identifies the n individuals. The solution for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the primal problem is then the dual solutions of the dual problem. We implement the dual problem using the Mosek optimization software

of Andersen (2010) through the R interface **Rmosek** of Friberg (2012). We have also implemented the estimation procedure using the **quantreg** package in R and the code will be made available for public use.[11]

5. Monte Carlo simulations

5.1. Finite sample performance of the proposed estimator

To assess the finite sample performance of the proposed convex clustering panel quantile regression estimator, we apply the method to simulated data sets. In particular, we consider data generated from two models and two error distributions for a range of n and T. The responses, Y_{it}, are generated by either a location shift model

$$Y_{it} = \alpha_i + X_{it}\beta + u_{it} \tag{8}$$

or a location scale shift model

$$Y_{it} = \alpha_i + X_{it}\beta + (1 + X_{it}\gamma)u_{it} \tag{9}$$

where the individual latent effects α_i are generated from three groups taking values $\{1, 2, 3\}$ with equal proportions. The covariate X_{it} is generated such that it has a non-zero interclass correlation coefficient. In particular,

$$X_{it} = \rho\alpha_i + \gamma_i + v_{it}$$

with γ_i and v_{it} independent and identically distributed over i and i, t respectively. We conduct simulation experiments with $\rho \in \{0, 0.5\}$ to investigate both cases where the fixed effect is independent or correlated with the covariate.[12] The true parameters are $\beta = 1$ and $\gamma = 1/10$. The error terms u_{it} are i.i.d. following either a standard normal distribution or a student t distribution with three degrees of freedom. Results reported are based on 2000 repetitions.

We first investigate the performance of using the information criterion for estimating the number of groups. Tables 1 and 2 report the proportion of estimated number of groups under the two models for different combinations of n and T for $\tau = 0.5$ and $\tau = 0.75$, respectively. Throughout the simulations, we used an equally spaced grid of λ values with width $1/200$ and support $[0, 0.35]$. This grid was chosen to ensure that the number of groups estimated for different λ values covers the integers in range $[1, n]$. For the IC criteria, the sparsity function $\hat{s}(\tau)$ is estimated with bandwidth chosen based on the Hall and Sheather (1988) rule implemented in the **quantreg** package (see discussion in Koenker (2005)).

The results suggest that the probability of getting the correct number of groups for $\tau = 0.5$ is slightly better than for $\tau = 0.75$. When $T \geq 30$, the estimates for the number of groups for both error distributions and for different n are mostly satisfactory. The performance for t error deteriorates compared to those with normal error, especially for higher quantiles. For $T = 15$ and quantiles other than the median the proposed method should be used with caution. Including correlation between individual effects and predictors does not lead to dramatic changes in the accuracy for estimating the number of groups and group membership.

Tables 3 and 4 summarize the finite sample properties of $\hat{\beta}^{IC}(\tau)$ for $\tau = 0.5$ and $\tau = 0.75$, respectively and compare with the QRFE estimator where no penalization on the individual fixed effect is used (i.e. $\lambda = 0$).

The standard errors used for constructing confidence intervals (nominal coverage 95%) are based on the `nid` option with Hall–Sheather bandwidth in the package quantreg. Results based on the Bofinger bandwidth selection are similar and not reported here. For the QRFE, the Bofinger bandwidth rule was used since the Hall–Sheather rule resulted in substantial under-coverage with $T = 30$ for some of the models.

When covariates and fixed effects are independent, the RMSE of the PQR-FEgroup estimator, $\hat{\beta}(\tau)$, is smaller than that of the QRFE for all settings considered. This shows that our penalization gains efficiency for estimating β when there is group structure in the fixed effects. The results do not change much from normal error to t error and from median to higher quantiles.

Introducing correlation between predictors and group membership leads to a bias for the grouped effect estimator, while the fixed effects estimator does not suffer from additional bias. This bias can be quite noticeable for small values of T, especially at the 75% quantile. The bias becomes negligible as T increases, so there is no contradiction to our asymptotic theory. An intuitive explanation for this behavior is that for smaller T it is difficult to get a perfect grouping, and a wrong grouping leads to bias since there is dependence between predictors and group structure.

Last, we report in Tables 5 and 6 the proportion of perfect classification of individual effects and the average value of the percentage of correct classification together with their standard errors. Since the comparison of the estimated membership and the true membership only makes sense when $\hat{K} = K_0$, the estimated membership are based on λ for which $\hat{K} = K_0$ (see Su et al. (2016) for a similar approach). Results suggest that for $T \geq 30$ and $\tau = 0.5$, the group membership estimation is quite satisfactory. While the proportion of perfect matches is low even for $T = 30$, the average

[11] Mosek is a commercial state-of-the-art convex optimization solver that provides a free academic license. We use its interior point algorithm to solve our linear programming problem. The estimation procedure implemented using the **quantreg** package calls the sfn method, which uses the Frisch–Newton algorithm and exploits the sparse algebra to compute iterates.

[12] Koenker (2004) used a similar data generating process with $\rho = 0$ for X and pointed out that the interclass correlation induced by γ_i is crucial for the penalized quantile regression fixed effect estimator to have superior performance than the unpenalized QRFE estimator.

Table 1
Frequency of estimated number of groups as $k = 1, \dots K$. We aggregate the frequency for $K \geq 5$ since the occurrence is not very often. True $K_0 = 3$. Results are based on 2000 simulation repetitions for quantile level $\tau = 0.5$.

n	T	Normal error					t_3 error				
		1	2	3	4	≥5	1	2	3	4	≥5
DGP1: Independence between α_i and x_{it}.											
Model 1: Location shift model											
30	15	0	0.074	0.504	0.324	0.098	0	0.168	0.490	0.266	0.076
30	30	0	0.002	0.803	0.167	0.028	0	0.007	0.744	0.202	0.048
30	60	0	0.000	0.984	0.016	0.000	0	0.000	0.966	0.032	0.002
60	15	0	0.056	0.502	0.315	0.128	0	0.122	0.428	0.319	0.130
60	30	0	0.000	0.856	0.127	0.018	0	0.002	0.767	0.186	0.046
60	60	0	0.000	0.992	0.008	0.000	0	0.000	0.978	0.021	0.000
90	15	0	0.040	0.465	0.339	0.156	0	0.113	0.392	0.322	0.172
90	30	0	0.000	0.872	0.114	0.013	0	0.000	0.778	0.182	0.038
90	60	0	0.000	0.996	0.004	0.000	0	0.000	0.984	0.016	0.000
Model 2: Location-scale shift model											
30	15	0	0.059	0.512	0.336	0.092	0	0.158	0.496	0.266	0.081
30	30	0	0.000	0.814	0.159	0.026	0	0.006	0.759	0.198	0.037
30	60	0	0.000	0.982	0.017	0.000	0	0.000	0.964	0.034	0.002
60	15	0	0.038	0.520	0.324	0.118	0	0.112	0.437	0.330	0.121
60	30	0	0.000	0.857	0.126	0.016	0	0.002	0.776	0.182	0.038
60	60	0	0.000	0.994	0.006	0.000	0	0.000	0.980	0.020	0.000
90	15	0	0.032	0.506	0.318	0.144	0	0.108	0.372	0.328	0.192
90	30	0	0.000	0.876	0.110	0.013	0	0.001	0.794	0.170	0.034
90	60	0	0.000	0.992	0.008	0.000	0	0.000	0.979	0.020	0.000
DGP2: Correlation between α_i and x_{it}.											
Model 1: Location shift model											
30	15	0	0.080	0.496	0.320	0.104	0	0.173	0.478	0.268	0.082
30	30	0	0.003	0.788	0.178	0.031	0	0.012	0.722	0.229	0.038
30	60	0	0.000	0.986	0.014	0.000	0	0.000	0.970	0.028	0.002
60	15	0	0.062	0.492	0.314	0.132	0	0.128	0.426	0.315	0.131
60	30	0	0.000	0.852	0.135	0.013	0	0.002	0.736	0.217	0.046
60	60	0	0.000	0.992	0.008	0.000	0	0.000	0.980	0.020	0.000
90	15	0	0.045	0.463	0.332	0.160	0	0.116	0.378	0.327	0.179
90	30	0	0.000	0.854	0.133	0.012	0	0.002	0.732	0.220	0.046
90	60	0	0.000	0.994	0.006	0.000	0	0.000	0.976	0.024	0.000
Model 2: Location-scale shift model											
30	15	0	0.138	0.453	0.330	0.078	0	0.228	0.456	0.249	0.066
30	30	0	0.008	0.724	0.224	0.044	0	0.040	0.666	0.251	0.042
30	60	0	0.000	0.972	0.025	0.003	0	0.000	0.923	0.068	0.008
60	15	0	0.099	0.442	0.312	0.147	0	0.176	0.367	0.313	0.144
60	30	0	0.001	0.748	0.210	0.042	0	0.016	0.631	0.272	0.081
60	60	0	0.000	0.976	0.024	0.000	0	0.000	0.960	0.036	0.003
90	15	0	0.098	0.389	0.324	0.189	0	0.145	0.330	0.316	0.210
90	30	0	0.001	0.756	0.207	0.036	0	0.016	0.612	0.288	0.085
90	60	0	0.000	0.978	0.022	0.000	0	0.000	0.950	0.050	0.000

proportion of correct classification shows that those effects are typically due to very few misclassified individuals. For $T = 15$ perfect classification is almost impossible while average correct classification rates remain reasonable. Adding correlation between individual effects and covariates leads to a deterioration of the probability for achieving a perfect grouping for location-scale models, especially at higher quantiles, but does not have a strong impact on other results. Overall the simulations suggest that for small T, there is just not enough information available for each individual to hope for perfect classification.

5.2. Further analysis of tuning parameters in the IC criteria

The discussion at the end of Section 2 provides a motivation for the tuning parameters $\hat{C}$ in the IC criteria. Here we further investigate the impact of rescaling $p_{n,T}$ by different factors. For illustration we consider DGP1 in the previous section where data are generated based on the model (9). We use a grid of constants $c \in [0.01, 0.3]$ with width 0.01 and plot the associated performance of the estimated number of groups, the RMSE of $\hat{\beta}^{IC}(\tau)$ and the coverage rate for $p_{n,T} = cnT^{1/4}$. Fig. 1 and Fig. 2 contain corresponding results for the location-scale shift model with t errors and $\tau = 0.5, 0.75$, respectively. For T as small as 15, the performance is quite sensitive to the chosen constant. As predicted by the theory this dependence becomes somewhat less prominent as T increases. Overall the choice $p_{n,T} = nT^{1/4}/10$ shows good performance for settings that we tried in this simulation. The patterns are similar for those with the normal

Table 2
Frequency of estimated number of groups as $k = 1, \ldots K$. We aggregate the frequency for $K \geq 5$ since the occurrence is not very often. True $K_0 = 3$. Results are based on 2000 simulation repetitions for quantile level $\tau = 0.75$.

n	T	Normal error					t_3 error				
		1	2	3	4	≥5	1	2	3	4	≥5
DGP1: Independence between α_i and x_{it}.											
Model 1: Location shift model											
30	15	0	0.159	0.484	0.284	0.072	0.002	0.330	0.408	0.210	0.048
30	30	0	0.011	0.734	0.207	0.048	0.000	0.200	0.560	0.207	0.033
30	60	0	0.000	0.966	0.032	0.002	0.000	0.011	0.866	0.112	0.012
60	15	0	0.129	0.398	0.325	0.148	0.000	0.188	0.337	0.306	0.168
60	30	0	0.002	0.760	0.198	0.040	0.000	0.105	0.472	0.328	0.094
60	60	0	0.000	0.970	0.029	0.000	0.000	0.000	0.861	0.128	0.012
90	15	0	0.110	0.364	0.331	0.196	0.001	0.142	0.314	0.293	0.251
90	30	0	0.000	0.768	0.194	0.038	0.000	0.072	0.450	0.341	0.136
90	60	0	0.000	0.966	0.034	0.000	0.000	0.000	0.852	0.136	0.012
Model 2: Location-scale shift model											
30	15	0	0.158	0.472	0.300	0.070	0.002	0.334	0.390	0.214	0.060
30	30	0	0.007	0.754	0.200	0.039	0.000	0.179	0.581	0.216	0.024
30	60	0	0.000	0.968	0.031	0.001	0.000	0.007	0.868	0.118	0.008
60	15	0	0.114	0.404	0.334	0.147	0.000	0.178	0.342	0.306	0.174
60	30	0	0.002	0.771	0.188	0.038	0.000	0.082	0.474	0.353	0.091
60	60	0	0.000	0.974	0.025	0.001	0.000	0.000	0.870	0.123	0.007
90	15	0	0.093	0.354	0.346	0.208	0.000	0.134	0.292	0.310	0.264
90	30	0	0.000	0.766	0.198	0.036	0.000	0.058	0.456	0.342	0.144
90	60	0	0.000	0.955	0.044	0.000	0.000	0.000	0.854	0.132	0.014
DGP2: Correlation between α_i and x_{it}.											
Model 1: Location shift model											
30	15	0	0.172	0.494	0.265	0.070	0.006	0.327	0.416	0.200	0.052
30	30	0	0.014	0.716	0.222	0.050	0.000	0.214	0.552	0.208	0.027
30	60	0	0.000	0.954	0.044	0.002	0.000	0.014	0.844	0.132	0.009
60	15	0	0.137	0.414	0.308	0.141	0.001	0.178	0.342	0.304	0.175
60	30	0	0.003	0.758	0.202	0.037	0.000	0.109	0.464	0.328	0.099
60	60	0	0.000	0.974	0.026	0.000	0.000	0.000	0.869	0.118	0.013
90	15	0	0.110	0.349	0.336	0.205	0.002	0.139	0.320	0.294	0.246
90	30	0	0.000	0.768	0.198	0.034	0.000	0.074	0.447	0.349	0.130
90	60	0	0.000	0.964	0.034	0.002	0.000	0.000	0.836	0.149	0.015
Model 2: Location-scale shift model											
30	15	0	0.228	0.446	0.264	0.062	0.014	0.326	0.414	0.203	0.042
30	30	0	0.038	0.660	0.257	0.045	0.000	0.276	0.502	0.196	0.027
30	60	0	0.000	0.932	0.064	0.004	0.000	0.039	0.769	0.180	0.012
60	15	0	0.176	0.377	0.304	0.144	0.004	0.169	0.338	0.303	0.185
60	30	0	0.018	0.652	0.257	0.073	0.000	0.157	0.420	0.311	0.112
60	60	0	0.000	0.954	0.044	0.002	0.000	0.004	0.775	0.192	0.030
90	15	0	0.146	0.309	0.342	0.204	0.006	0.108	0.296	0.312	0.277
90	30	0	0.010	0.660	0.258	0.072	0.000	0.118	0.373	0.342	0.167
90	60	0	0.000	0.948	0.050	0.002	0.000	0.002	0.753	0.212	0.032

error and the location-shift models and likewise for DGP2, results are reported in Figs. 1–6 in the online Appendix for the sake of completeness.

6. Conclusions and future extensions

The present paper suggests a simple and computationally efficient way to incorporate group fixed effects into a panel data quantile regression by means of a convex clustering penalty. We develop theoretical results on consistent group structure estimation and discuss the asymptotic properties of the resulting joint and group-specific estimators.

There are several directions that we plan to explore in the future. First, our theory focused on individual fixed effects while assuming common slope coefficients. It is equally interesting to allow for group structure in some of the slope coefficients while keeping other slope coefficients common across individuals, perhaps even allowing for individual fixed effects. This can be achieved by straightforward modifications of the penalization approach which we explored so far, but a more detailed theoretical analysis of this approach remains beyond the scope of the present paper.

Second, one can take the standpoint that in many applications there is no exact group structure. In such settings, an alternative interpretation of the penalty which we investigated is as a way of regularizing problems that have too many parameters. Such an interpretation is in the spirit of the proposals of Koenker (2004) and Lamarche (2010), and a detailed investigation of the resulting bias–variance trade-off warrants further research.

Table 3
Comparison of bias and root mean squared error of $\hat{\beta}(\tau)$ based on the group fixed effect quantile regression (PQR-FEgroup)and the fixed effect quantile regression estimator (QRFE). Results are based on 2000 simulation repetitions for quantile level $\tau = 0.5$. DGP1 assumes that x_{it} is independent of the fixed effect α_i. DGP2 assumes that $x_{it} = 0.5\alpha_i + \gamma_i + v_{it}$.

n	T	Normal error						t_3 error					
		PQR-FEgroup			QRFE			PQR-FEgroup			QRFE		
		Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
DGP1: Independence between α_i and x_{it}.													
Model 1: Location shift model													
30	15	−0.001	0.060	0.792	−0.001	0.062	0.924	0.001	0.069	0.797	0.000	0.069	0.931
30	30	0.001	0.036	0.900	0.001	0.042	0.906	−0.001	0.042	0.874	−0.001	0.047	0.908
30	60	0.000	0.022	0.942	0.000	0.030	0.932	0.000	0.025	0.933	0.000	0.033	0.934
60	15	0.000	0.040	0.832	0.001	0.042	0.798	0.001	0.046	0.819	0.001	0.048	0.780
60	30	0.001	0.024	0.902	0.000	0.030	0.942	−0.001	0.028	0.884	−0.001	0.033	0.942
60	60	0.001	0.016	0.936	0.001	0.022	0.936	0.000	0.017	0.940	0.000	0.023	0.942
90	15	−0.001	0.033	0.834	0.000	0.035	0.836	−0.001	0.037	0.824	−0.001	0.039	0.828
90	30	0.000	0.020	0.914	0.001	0.024	0.955	0.001	0.023	0.886	0.000	0.027	0.968
90	60	0.000	0.012	0.942	0.000	0.018	0.905	0.000	0.014	0.937	0.000	0.019	0.913
Model 2: Location-scale shift model													
30	15	0.000	0.059	0.803	−0.001	0.060	0.923	0.001	0.067	0.800	0.000	0.067	0.932
30	30	0.002	0.035	0.893	0.001	0.041	0.894	0.000	0.041	0.870	0.000	0.047	0.890
30	60	0.000	0.022	0.936	0.000	0.030	0.924	0.000	0.025	0.929	0.000	0.032	0.935
60	15	0.001	0.039	0.834	0.002	0.041	0.792	0.001	0.045	0.820	0.001	0.047	0.774
60	30	0.001	0.023	0.903	0.000	0.029	0.940	0.000	0.028	0.882	0.000	0.032	0.946
60	60	0.001	0.015	0.932	0.001	0.021	0.935	0.000	0.017	0.932	0.000	0.022	0.938
90	15	0.000	0.032	0.841	0.000	0.034	0.824	−0.001	0.037	0.818	−0.001	0.038	0.830
90	30	0.000	0.019	0.913	0.000	0.024	0.952	0.002	0.022	0.884	0.001	0.026	0.966
90	60	0.000	0.012	0.940	0.000	0.017	0.904	0.001	0.014	0.935	0.000	0.019	0.906
DGP2: Correlation between α_i and x_{it}.													
Model 1: Location shift model													
30	15	0.016	0.063	0.782	−0.001	0.062	0.924	0.023	0.073	0.762	0.000	0.069	0.931
30	30	0.007	0.037	0.889	0.001	0.042	0.906	0.007	0.044	0.868	−0.001	0.047	0.908
30	60	0.001	0.022	0.944	0.000	0.030	0.932	0.001	0.025	0.928	0.000	0.033	0.934
60	15	0.015	0.044	0.783	0.001	0.042	0.798	0.021	0.052	0.768	0.001	0.048	0.780
60	30	0.005	0.026	0.889	0.000	0.030	0.942	0.006	0.029	0.872	−0.001	0.033	0.942
60	60	0.001	0.016	0.934	0.001	0.022	0.936	0.001	0.017	0.938	0.000	0.023	0.942
90	15	0.014	0.037	0.774	0.000	0.035	0.836	0.019	0.043	0.756	−0.001	0.039	0.828
90	30	0.004	0.020	0.894	0.001	0.024	0.955	0.007	0.024	0.871	0.000	0.027	0.968
90	60	0.000	0.013	0.941	0.000	0.018	0.905	0.001	0.014	0.934	0.000	0.019	0.913
Model 2: Location-scale shift model													
30	15	0.020	0.071	0.752	−0.001	0.066	0.926	0.027	0.080	0.751	0.000	0.074	0.933
30	30	0.010	0.042	0.861	0.001	0.045	0.898	0.011	0.051	0.828	−0.001	0.051	0.892
30	60	0.002	0.025	0.924	0.000	0.033	0.926	0.002	0.029	0.912	0.000	0.035	0.938
60	15	0.020	0.050	0.750	0.002	0.045	0.792	0.026	0.058	0.729	0.001	0.051	0.775
60	30	0.007	0.029	0.860	0.000	0.032	0.940	0.009	0.034	0.832	0.000	0.035	0.943
60	60	0.002	0.018	0.924	0.001	0.023	0.934	0.002	0.019	0.920	0.000	0.025	0.938
90	15	0.018	0.041	0.752	0.000	0.037	0.824	0.024	0.049	0.706	−0.001	0.042	0.831
90	30	0.006	0.024	0.861	0.001	0.026	0.954	0.009	0.028	0.831	0.001	0.029	0.966
90	60	0.001	0.014	0.932	0.000	0.019	0.905	0.002	0.016	0.923	0.000	0.020	0.906

Finally, a deeper analysis of issues that are related to uniformity of distributional approximation in the entire parameter space was not addressed here, but remains an important theoretical and practical question which we hope to address in the future.

7. Proofs

We begin by collecting some useful facts and defining additional notation. We will repeatedly make use of Knight's identity (see Koenker (2005), p. 121)) which holds for $u \neq 0$:

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v \mathbf{I}\{u \leq s\} - \mathbf{I}\{u \leq 0\}ds. \tag{10}$$

Additionally, let $\gamma_{0i} := (\alpha_{0i}, \beta_0^\top)^\top$. The symbols $a_n \lesssim b_n, a_n \gtrsim b_n$ will mean that there exists a non-random constant $C \in (0, \infty)$ which is independent of n, T, τ such that $P(a_n \leq Cb_n) = 1$ and $P(a_n \geq Cb_n) = 1$, respectively. Define $\varepsilon_{it}^\tau := Y_{it} - Z_{it}^\top \gamma_{i0}(\tau)$ and let $F_{\varepsilon_{it}^\tau|X_{it}}(u|X_{it}) = F_{Y_{it}|X_{it}}(Z_{it}^\top \gamma_{i0}(\tau) + u|X_{it})$ denote the conditional cdf of ε_{it}^τ given X_{it}. When there is no risk of confusion, we will also write ε_{it} instead of ε_{it}^τ. Define $\psi_\tau(x) := (\mathbf{I}\{x \leq 0\} - \tau)$.

Table 4
Comparison of bias and root mean squared error of $\hat{\beta}(\tau)$ based on the group fixed effect quantile regression (PQR-FEgroup) and the fixed effect quantile regression estimator (QRFE). Results are based on 2000 simulation repetitions for quantile level $\tau = 0.75$. DGP1 assumes that x_{it} is independent of the fixed effect α_i. DGP2 assumes that $x_{it} = 0.5\alpha_i + \gamma_i + v_{it}$.

n	T	Normal error						t_3 error					
		PQR-FEgroup			QRFE			PQR-FEgroup			QRFE		
		Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
DGP1: Independence between α_i and x_{it}.													
Model 1: Location shift model: normal error													
30	15	−0.001	0.066	0.784	−0.001	0.066	0.921	0.003	0.130	0.812	0.000	0.082	0.944
30	30	0.000	0.041	0.856	−0.001	0.045	0.895	−0.001	0.057	0.816	−0.001	0.057	0.890
30	60	0.000	0.025	0.930	0.000	0.033	0.922	0.000	0.035	0.898	−0.001	0.040	0.938
60	15	0.001	0.046	0.790	0.000	0.047	0.787	0.001	0.073	0.784	−0.001	0.059	0.786
60	30	0.000	0.029	0.874	0.000	0.033	0.939	−0.001	0.038	0.846	0.000	0.039	0.945
60	60	0.000	0.017	0.932	0.000	0.023	0.890	0.000	0.023	0.923	0.001	0.028	0.902
90	15	−0.001	0.037	0.812	0.000	0.038	0.827	0.002	0.080	0.790	0.001	0.048	0.825
90	30	0.000	0.023	0.884	0.001	0.026	0.957	0.001	0.030	0.850	0.001	0.033	0.963
90	60	0.000	0.014	0.936	0.000	0.018	0.930	0.000	0.018	0.917	0.000	0.023	0.936
Model 2: Location-scale shift model: normal error													
30	15	−0.009	0.065	0.772	−0.007	0.064	0.924	−0.009	0.126	0.796	−0.009	0.081	0.942
30	30	−0.001	0.040	0.852	−0.003	0.045	0.886	−0.005	0.056	0.814	−0.005	0.056	0.886
30	60	0.000	0.025	0.928	−0.001	0.032	0.924	−0.001	0.034	0.896	−0.002	0.039	0.933
60	15	−0.006	0.045	0.792	−0.005	0.046	0.794	−0.012	0.057	0.773	−0.009	0.058	0.784
60	30	−0.001	0.027	0.884	−0.002	0.032	0.926	−0.005	0.037	0.842	−0.004	0.038	0.936
60	60	0.000	0.017	0.930	−0.001	0.023	0.884	−0.001	0.022	0.906	−0.001	0.028	0.892
90	15	−0.007	0.037	0.781	−0.005	0.037	0.831	−0.012	0.048	0.776	−0.007	0.047	0.824
90	30	−0.001	0.022	0.884	−0.002	0.026	0.946	−0.003	0.029	0.854	−0.003	0.032	0.958
90	60	0.000	0.014	0.936	−0.001	0.018	0.926	−0.001	0.018	0.918	−0.002	0.023	0.934
DGP2: Correlation between α_i and x_{it}.													
Model 1: Location shift model													
30	15	0.021	0.071	0.756	−0.001	0.066	0.921	0.044	0.170	0.757	0.000	0.082	0.944
30	30	0.008	0.042	0.853	−0.001	0.045	0.895	0.019	0.062	0.781	−0.001	0.057	0.890
30	60	0.001	0.026	0.925	0.000	0.033	0.922	0.005	0.036	0.883	−0.001	0.040	0.938
60	15	0.021	0.052	0.746	0.000	0.047	0.787	0.033	0.086	0.708	−0.001	0.059	0.786
60	30	0.006	0.030	0.854	0.000	0.033	0.939	0.015	0.043	0.788	0.000	0.039	0.945
60	60	0.001	0.018	0.934	0.000	0.023	0.890	0.003	0.023	0.909	0.001	0.028	0.902
90	15	0.019	0.043	0.726	0.000	0.038	0.827	0.035	0.092	0.673	0.001	0.048	0.825
90	30	0.006	0.024	0.868	0.001	0.026	0.957	0.015	0.036	0.787	0.001	0.033	0.963
90	60	0.001	0.014	0.935	0.000	0.018	0.930	0.003	0.019	0.912	0.000	0.023	0.936
Model 2: Location-scale shift model													
30	15	0.017	0.074	0.752	−0.007	0.071	0.924	0.052	0.228	0.760	−0.009	0.089	0.942
30	30	0.009	0.048	0.830	−0.003	0.049	0.889	0.019	0.067	0.778	−0.005	0.062	0.888
30	60	0.002	0.029	0.912	−0.001	0.035	0.922	0.007	0.041	0.871	−0.002	0.043	0.935
60	15	0.018	0.054	0.747	−0.005	0.051	0.794	0.031	0.135	0.739	−0.009	0.064	0.784
60	30	0.007	0.034	0.834	−0.002	0.035	0.928	0.015	0.047	0.785	−0.005	0.042	0.936
60	60	0.002	0.020	0.914	−0.001	0.025	0.886	0.004	0.027	0.874	−0.001	0.031	0.896
90	15	0.016	0.045	0.744	−0.005	0.041	0.830	0.034	0.145	0.718	−0.007	0.052	0.828
90	30	0.007	0.027	0.840	−0.002	0.028	0.946	0.015	0.039	0.762	−0.003	0.036	0.960
90	60	0.002	0.016	0.916	−0.001	0.020	0.926	0.003	0.021	0.896	−0.002	0.025	0.932

7.1. Proof of Theorem 3.1

We begin by stating some useful technical results which will be proved at the end of this section.

Lemma 7.1. *For any fixed $\beta \in \mathbb{R}^p$ define $\varepsilon^{\tau}_{it,\beta} := Y_{it} - X_{it}^{\top}\beta - \alpha_{i0}(\tau)$. Then we have under assumptions (A1)–(A3)*

$$\sum_{t=1}^{T} \rho_\tau(\varepsilon^{\tau}_{it,\beta} - a_1) - \rho_\tau(\varepsilon^{\tau}_{it,\beta} - a_2) = (a_2 - a_1)\sum_{t=1}^{T} \psi_\tau(\varepsilon^{\tau}_{it,\beta}) + \tilde{r}^{(1)}_{n,i}(a_1, a_2) + \tilde{r}^{(2)}_{n,i}(a_1, a_2)$$

where

$$\sup_i \tilde{r}^{(1)}_{n,i}(a_1, a_2) \lesssim T|a_1 - a_2| \max(|a_1|, |a_2|), \quad \sup_i \tilde{r}^{(2)}_{n,i}(a_1, a_2) = |a_1 - a_2| O_P(T^{1/2}(\log n)^{1/2}).$$

Lemma 7.2. *Under assumptions (A1)–(A3) there exist $\varepsilon > 0, \infty > c_1, c_0 > 0$ such that for all $i = 1, \ldots, n$*

$$c_1\|\gamma - \gamma_{0i}\|^2 \geq \mathbb{E}[\rho_\tau(Y_{it} - Z_{it}^{\top}\gamma)] - \mathbb{E}[\rho_\tau(Y_{it} - Z_{it}^{\top}\gamma_{0i})] \geq c_0(\|\gamma - \gamma_{0i}\|^2 \wedge \varepsilon^2). \tag{11}$$

Table 5
Membership estimation for $\tau = 0.5$ for two different error distributions: Perfect Match states the percentage of perfect membership estimation out of the 2000 repetitions. Average match reports the mean of the percentage of correct membership estimation and the standard error reports the associated standard deviation.

n	T	Normal error			t_3 error		
		Perfect match	Avg match	Std error	Perfect match	Avg match	Std error
DGP1: Independence between α_i and x_{it}.							
Model 1: Location shift model							
30	15	0.025	0.659	0.285	0.008	0.636	0.267
30	30	0.384	0.877	0.222	0.223	0.820	0.262
30	60	0.918	0.989	0.069	0.822	0.976	0.105
60	15	0.003	0.651	0.307	0.000	0.590	0.297
60	30	0.225	0.900	0.202	0.088	0.841	0.256
60	60	0.884	0.992	0.064	0.735	0.985	0.077
90	15	0.000	0.646	0.317	0.000	0.562	0.320
90	30	0.119	0.912	0.187	0.024	0.839	0.270
90	60	0.856	0.995	0.044	0.654	0.985	0.080
Model 2: Location-scale shift model							
30	15	0.028	0.670	0.285	0.012	0.638	0.274
30	30	0.394	0.880	0.221	0.225	0.836	0.245
30	60	0.906	0.988	0.072	0.800	0.975	0.106
60	15	0.002	0.663	0.309	0.000	0.595	0.305
60	30	0.218	0.903	0.201	0.084	0.850	0.247
60	60	0.856	0.994	0.047	0.708	0.983	0.085
90	15	0.000	0.662	0.323	0.000	0.550	0.331
90	30	0.118	0.908	0.198	0.026	0.851	0.261
90	60	0.827	0.994	0.051	0.634	0.984	0.084
DGP2: Correlation between α_i and x_{it}.							
Model 1: Location shift model							
30	15	0.024	0.655	0.289	0.008	0.624	0.270
30	30	0.365	0.865	0.234	0.199	0.817	0.257
30	60	0.923	0.991	0.061	0.824	0.978	0.101
60	15	0.002	0.650	0.305	0.000	0.590	0.297
60	30	0.202	0.895	0.212	0.076	0.829	0.267
60	60	0.884	0.993	0.052	0.738	0.985	0.076
90	15	0.000	0.643	0.318	0.000	0.554	0.320
90	30	0.104	0.900	0.208	0.025	0.817	0.285
90	60	0.850	0.994	0.053	0.646	0.984	0.078
Model 2: Location-scale shift model							
30	15	0.006	0.636	0.263	0.002	0.608	0.248
30	30	0.215	0.826	0.250	0.119	0.780	0.260
30	60	0.818	0.979	0.096	0.649	0.949	0.146
60	15	0.000	0.612	0.293	0.000	0.545	0.280
60	30	0.083	0.837	0.254	0.026	0.764	0.294
60	60	0.715	0.981	0.091	0.516	0.968	0.117
90	15	0.000	0.585	0.312	0.000	0.514	0.296
90	30	0.023	0.832	0.268	0.006	0.744	0.322
90	60	0.648	0.983	0.085	0.398	0.962	0.133

Lemma 7.3. *Under assumption (A1) define for fixed $B \in \mathbb{R}$*

$$s_{n,1}(B) := \sup_i \sup_{|\gamma| \le B} \left| \sum_t \left(\rho_\tau(Y_{it} - Z_{it}^\top \gamma) - \rho_\tau(Y_{it}) - \mathbb{E}[\rho_\tau(Y_{it} - Z_{it}^\top \gamma) - \rho_\tau(Y_{it})] \right) \right| \tag{12}$$

We have for any fixed $B < \infty$, provided that $\min(n, T) \to \infty$, $\log n = o(T)$

$$s_{n,1}(B) = O_P(T^{1/2}(\log n)^{1/2}). \tag{13}$$

Proof of Theorem 3.1.

Step 1: first bounds In this step we shall prove that

$$\|\hat\beta - \beta_0\|^2 + \sup_i |\hat\alpha_i - \alpha_{i0}|^2 = O_P(T^{-1/2}(\log n)^{1/2} + \Lambda_D n/T). \tag{14}$$

Combine the results in Lemmas 7.2 and 7.3 to find that any minimizer of $\Theta(\alpha_1, \ldots, \alpha_n, \beta)$ must satisfy

$$c_0 T \sum_i \left\{ (\|\beta - \beta_0\|^2 + |\alpha_i - \alpha_{i0}|^2) \wedge \varepsilon^2 \right\} \lesssim n s_{n,1} + \Lambda_D n^2.$$

Table 6
Membership estimation for $\tau = 0.75$ for two different error distributions: Perfect Match states the percentage of perfect membership estimation out of the 2000 repetitions. Average match reports the mean of the percentage of correct membership estimation and the standard error reports the associated standard deviation.

n	T	Normal error			t_3 error		
		Perfect match	Avg match	Std error	Perfect match	Avg match	Std error
DGP1: Independence between α_i and x_{it}.							
Model 1: Location shift model							
30	15	0.012	0.666	0.240	0.000	0.574	0.181
30	30	0.242	0.849	0.220	0.034	0.737	0.228
30	60	0.846	0.982	0.082	0.471	0.923	0.160
60	15	0.000	0.634	0.258	0.000	0.558	0.197
60	30	0.098	0.873	0.198	0.002	0.732	0.248
60	60	0.754	0.984	0.072	0.260	0.934	0.142
90	15	0.000	0.633	0.267	0.000	0.530	0.210
90	30	0.041	0.880	0.201	0.000	0.734	0.253
90	60	0.690	0.984	0.071	0.182	0.939	0.132
Model 2: Location-scale shift model							
30	15	0.014	0.668	0.240	0.000	0.573	0.186
30	30	0.248	0.858	0.215	0.040	0.751	0.225
30	60	0.824	0.980	0.086	0.451	0.926	0.158
60	15	0.000	0.640	0.260	0.000	0.566	0.205
60	30	0.100	0.875	0.204	0.002	0.747	0.243
60	60	0.729	0.985	0.069	0.270	0.938	0.136
90	15	0.000	0.630	0.278	0.000	0.531	0.219
90	30	0.045	0.885	0.195	0.001	0.741	0.261
90	60	0.646	0.978	0.084	0.164	0.938	0.134
DGP2: Correlation between α_i and x_{it}.							
Model 1: Location shift model							
30	15	0.012	0.674	0.187	0.000	0.584	0.158
30	30	0.255	0.873	0.163	0.052	0.758	0.190
30	60	0.772	0.977	0.078	0.368	0.908	0.148
60	15	0.000	0.727	0.189	0.000	0.523	0.155
60	30	0.188	0.921	0.124	0.008	0.699	0.212
60	60	0.855	0.994	0.032	0.258	0.911	0.160
90	15	0.000	0.624	0.207	0.000	0.482	0.153
90	30	0.048	0.868	0.177	0.002	0.667	0.229
90	60	0.665	0.989	0.044	0.172	0.924	0.157
Model 2: Location-scale shift model							
30	15	0.002	0.618	0.161	0.000	0.538	0.141
30	30	0.072	0.776	0.183	0.008	0.665	0.171
30	60	0.468	0.945	0.103	0.140	0.819	0.187
60	15	0.000	0.556	0.162	0.000	0.476	0.134
60	30	0.000	0.761	0.188	0.000	0.591	0.179
60	60	0.355	0.954	0.094	0.055	0.825	0.198
90	15	0.000	0.527	0.172	0.000	0.428	0.113
90	30	0.002	0.748	0.212	0.000	0.554	0.191
90	60	0.240	0.953	0.103	0.010	0.805	0.211

Let $N_\Delta := \#\{i : \|\beta - \beta_0\|^2 + |\alpha_i - \alpha_{0i}|^2 \geq \Delta\}$. Then for any $0 < \Delta < \varepsilon^2$

$$TN_\Delta \Delta = O_P(ns_{n,1} + \Lambda_D n^2),$$

i.e. by Lemma 7.3

$$N_\Delta = nO_P((T^{-1/2}(\log n)^{1/2}) + \Lambda_D n/T)\Delta^{-1}$$

and in particular $N_\Delta = o_P(n)$ as long as $\Delta \gg T^{-1/2}(\log n)^{1/2} + \Lambda_D n/T$. Provided that $\Lambda_D n/T = o_P(1)$ we obtain

$$\|\hat{\beta} - \beta_0\|^2 = O_P(T^{-1/2}(\log n)^{1/2} + \Lambda_D n/T). \tag{15}$$

Define $D_{n,T} := T^{-1/2}(\log n)^{1/2} + \Lambda_D n/T$. Next we will prove that, provided $n\Lambda_D = o_P(T^{1/2})$, $T^{3/4}(\log n)^{3/4}/(n\Lambda_S) = o_P(1)$ also

$$\sup_i |\hat{\alpha}_i - \alpha_{i0}|^2 = O_P(D_{n,T}).$$

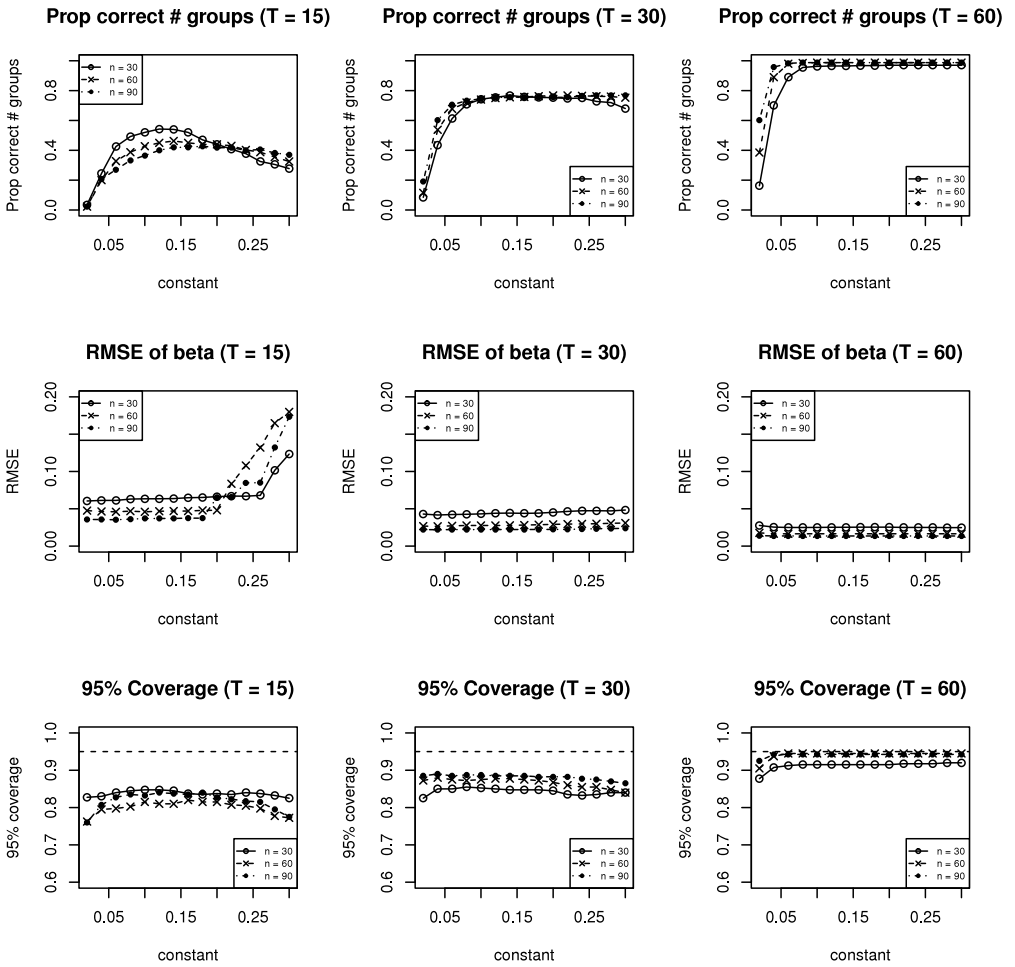

Fig. 1. Use different constants in $p_{n,T}$ for the IC criteria for location-scale shift model with t error on DGP1: For an equally spaced grid on $[0.01, 0.3]$ with width 0.01, the three columns represent different magnitudes of T while each figures in the row overlays the curves for $n \in \{30, 60, 90\}$ for various performance measures. The first row plots the proportion of correctly estimated number of groups. The second row plots the RMSE of $\hat{\beta}^{IC}(\tau)$ where $\tau = 0.5$ and the third plots the coverage rate for nominal size 5%. Results are based on 400 repetitions.

To this end, it suffices to prove that $\sup_i |\hat{\alpha}_i - \alpha_{i0}|^2 = O_P(d_{n,T})$ for any $n\Lambda_S \gg d_{n,T} \gg D_{n,T}$. Define

$$\tilde{\alpha}_i = \begin{cases} \hat{\alpha}_i & \text{if} \quad |\hat{\alpha}_i - \alpha_{0i}|^2 \le d_n, \\ \alpha_{i0} + d_n^{1/2}\mathrm{sgn}(\hat{\alpha}_i - \alpha_{i0}) & \text{if} \quad |\hat{\alpha}_i - \alpha_{0i}|^2 > d_n. \end{cases}$$

Define the set $E := \{i : \tilde{\alpha}_i = \hat{\alpha}_i\}$. Observe that

$$\begin{aligned} &\left| |\tilde{\alpha}_i - \tilde{\alpha}_j| - |\hat{\alpha}_i - \hat{\alpha}_j| \right| \le |\hat{\alpha}_i - \tilde{\alpha}_i| + |\hat{\alpha}_j - \tilde{\alpha}_j| && \forall i, j, \\ &|\tilde{\alpha}_i - \tilde{\alpha}_j| - |\hat{\alpha}_i - \hat{\alpha}_j| < -|\hat{\alpha}_i - \alpha_{0i}| + d_{n,T}^{1/2} && \exists k : i \in I_k \cap E^C, j \in I_k \cap E, \\ &|\tilde{\alpha}_i - \tilde{\alpha}_j| - |\hat{\alpha}_i - \hat{\alpha}_j| \le 0 && \exists k : i, j \in I_k. \end{aligned}$$

Thus

$$\begin{aligned} &\sum_{i,j} \lambda_{i,j} \left\{ |\hat{\alpha}_i - \hat{\alpha}_j| - |\tilde{\alpha}_i - \tilde{\alpha}_j| \right\} \\ = \; &\left(2\sum_{i\in E^C}\sum_{j\in E} + \sum_{i\in E^C}\sum_{j\in E^C} \right) \lambda_{i,j} \left\{ |\hat{\alpha}_i - \hat{\alpha}_j| - |\tilde{\alpha}_i - \tilde{\alpha}_j| \right\} \\ = \; &2\sum_k \sum_{i\in I_k\cap E^C} \left(\sum_{j\in I_k\cap E} + \sum_{j\in I_k^C\cap E} \right) \lambda_{i,j} \left\{ |\hat{\alpha}_i - \hat{\alpha}_j| - |\tilde{\alpha}_i - \tilde{\alpha}_j| \right\} \end{aligned}$$

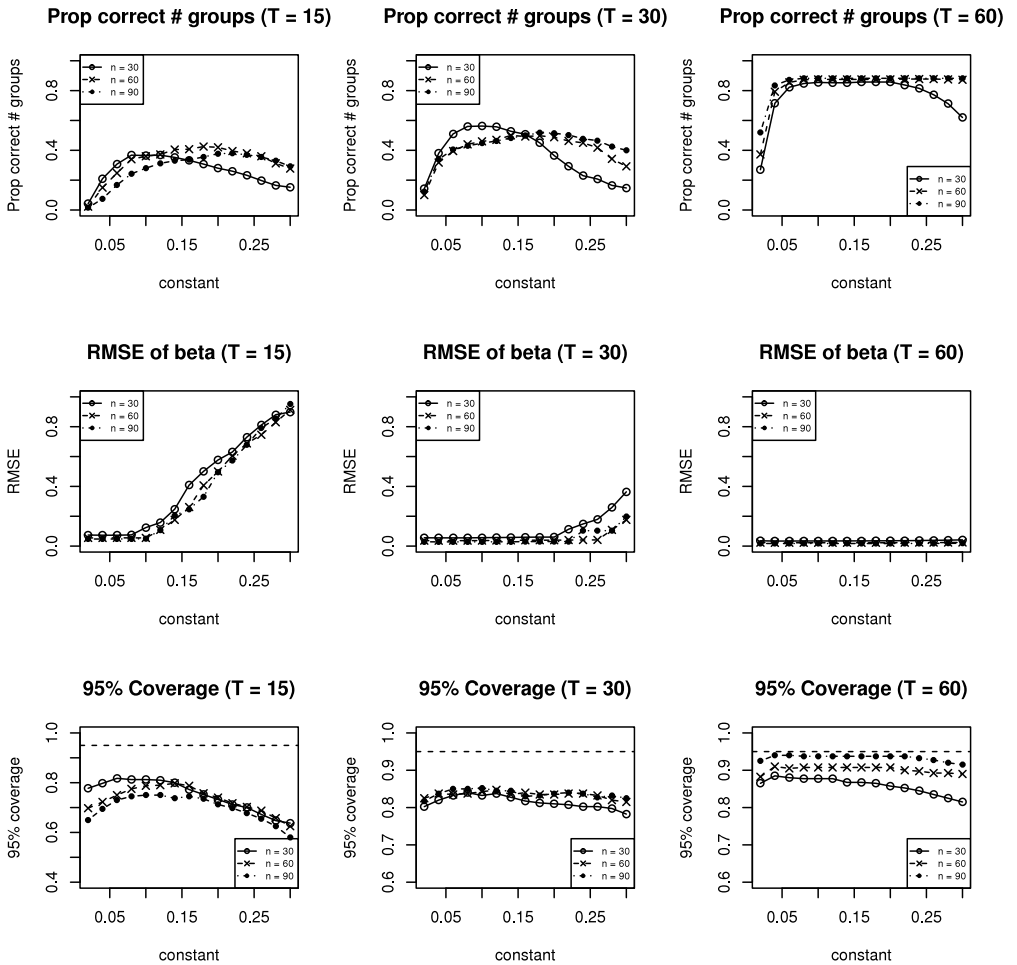

Fig. 2. Use different constants in $p_{n,T}$ for the IC criteria for location-scale shift model with t error on DGP1: For an equally spaced grid on $[0.01, 0.3]$ with width 0.01, the three columns represent different magnitudes of T while each figures in the row overlays the curves for $n \in \{30, 60, 90\}$ for various performance measures. The first row plots the proportion of correctly estimated number of groups. The second row plots the RMSE of $\hat{\beta}^{IC}(\tau)$ where $\tau = 0.75$ and the third plots the coverage rate for nominal size 5%. Results are based on 400 repetitions.

$$
\begin{aligned}
&+ \sum_k \sum_{i \in I_k \cap E^C} \Big(\sum_{j \in I_k \cap E^C} + \sum_{j \in I_k^C \cap E^C} \Big) \lambda_{i,j} \Big\{ |\hat{\alpha}_i - \hat{\alpha}_j| - |\tilde{\alpha}_i - \tilde{\alpha}_j| \Big\} \\
\geq\ & 2 \sum_k \sum_{i \in I_k \cap E^C} \Big(\Lambda_S |I_k \cap E| - n\Lambda_D \Big) \{ |\hat{\alpha}_i - \alpha_{0i}| - d_{n,T}^{1/2} \} \\
&- \sum_k \sum_{i \in I_k \cap E^C} \sum_{j \in I_k^C \cap E^C} \lambda_{i,j} \Big\{ |\hat{\alpha}_i - \tilde{\alpha}_i| + |\hat{\alpha}_j - \tilde{\alpha}_j| \Big\} \\
\geq\ & 2 \sum_k \sum_{i \in I_k \cap E^C} \Big(\Lambda_S |I_k \cap E| - n\Lambda_D \Big) \{ |\hat{\alpha}_i - \alpha_{0i}| - d_{n,T}^{1/2} \} \\
&+ \sum_k \sum_{i \in I_k \cap E^C} \Big\{ -n\Lambda_D \{ |\hat{\alpha}_i - \alpha_{0i}| - d_{n,T}^{1/2} \} - \Lambda_D \sum_{j \in I_k^C \cap E^C} \{ |\hat{\alpha}_j - \alpha_{0j}| - d_{n,T}^{1/2} \} \Big\} \\
\geq\ & 2 \sum_k \sum_{i \in I_k \cap E^C} \Big(\Lambda_S |I_k \cap E| - 2n\Lambda_D \Big) \{ |\hat{\alpha}_i - \alpha_{0i}| - d_{n,T}^{1/2} \}.
\end{aligned}
$$

Now since $N_{d_n} = o_P(n)$ and by definition of $\tilde{\alpha}_n$ it follows that under (C)

$$
\max_k \left| \frac{|I_k \cap E|}{n\mu_k} - 1 \right| = o_P(1),
$$

and since by assumption $\Lambda_D/\Lambda_S = o_P(1)$ we obtain

$$\sum_{i,j} \lambda_{i,j} \left\{ |\hat\alpha_i - \hat\alpha_j| - |\tilde\alpha_i - \tilde\alpha_j| \right\} \gtrsim n\Lambda_S \sum_{i \in E^C} \{|\hat\alpha_i - \alpha_{0i}| - d_{n,T}^{1/2}\}.$$

Next we note that for any i with $|\hat\alpha_i - \alpha_{0i}|^2 \geq (2 + c_1/c_0) d_{n,T}$ we have

$$\begin{aligned}
&\frac{1}{T}\sum_t \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \hat\alpha_i) - \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \tilde\alpha_i) \\
\geq\ & \int \rho_\tau(y - x^\top \hat\beta - \hat\alpha_i) - \rho_\tau(y - x^\top \hat\beta - \tilde\alpha_i) dP^{Y_{i1}, X_{i1}}(x,y) - 2s_{n,1}/T \\
=\ & \int \rho_\tau(y - x^\top \hat\beta - \hat\alpha_i) - \rho_\tau(y - x^\top \beta_0 - \alpha_{i0}) dP^{Y_{i1}, X_{i1}}(x,y) \\
& \quad - \int \rho_\tau(y - x^\top \hat\beta - \tilde\alpha_i) - \rho_\tau(y - x^\top \beta_0 - \alpha_{i0}) dP^{Y_{i1}, X_{i1}}(x,y) - 2s_{n,1}/T \\
\geq\ & c_0(\{\|\hat\beta - \beta_0\|^2 + |\hat\alpha_i - \alpha_{0i}|^2\} \wedge \varepsilon^2) - c_1(\|\hat\beta - \beta_0\|^2 + |\tilde\alpha_i - \alpha_{0i}|^2) - 2s_{n,1}/T \\
>\ & 0
\end{aligned}$$

with probability tending to one by (15) and the definition of $d_{n,T}$. For i with $|\hat\alpha_i - \alpha_{0i}|^2 < (2 + c_1/c_0) d_{n,T}$ note that by Lemma 7.1

$$\begin{aligned}
&\left| \sum_t \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \hat\alpha_i) - \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \tilde\alpha_i) \right| \\
\lesssim\ & |\hat\alpha_i - \tilde\alpha_i| \left(\sum_t \psi_\tau(\varepsilon^\tau_{it,\hat\beta}) + T d_{n,T}^{1/2} + O_P(T^{1/2}(\log n)^{1/2}) \right) \\
\lesssim\ & \{|\hat\alpha_i - \alpha_{0i}| - d_{n,T}^{1/2}\} \left(T\|\hat\beta - \beta_0\| + T d_{n,T}^{1/2} + O_P(T^{1/2}(\log n)^{1/2}) \right)
\end{aligned}$$

where the O_P terms are uniform in i. Thus

$$\begin{aligned}
&\sum_i \sum_t \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \hat\alpha_i) - \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \tilde\alpha_i) \\
\gtrsim\ & - \left(T d_{n,T}^{1/2} + O_P(T^{1/2}(\log n)^{1/2}) \right) \sum_{i \in E^C} \{|\hat\alpha_i - \alpha_{0i}| - d_{n,T}^{1/2}\}
\end{aligned}$$

Summarizing we have proved that

$$\begin{aligned}
&\Theta(\hat\alpha_1, \ldots, \hat\alpha_n, \hat\beta) - \Theta(\tilde\alpha_1, \ldots, \tilde\alpha_n, \hat\beta) \\
\gtrsim\ & \left[n\Lambda_S - \left(T d_{n,T}^{1/2} + O_P(T^{1/2}(\log n)^{1/2}) \right) \right] \sum_{i \in E^C} \{|\hat\alpha_i - \alpha_{0i}| - d_{n,T}^{1/2}\}.
\end{aligned}$$

Under the conditions $n\Lambda_D = o_P(T^{1/2})$, $T^{3/4}(\log n)^{3/4}/(n\Lambda_S) = o_P(1)$ the last line is strictly positive with probability tending to one unless $E^C = \emptyset$ with probability tending to one. Thus the proof of (14) is complete.

Step 2: recovery of clusters with probability to one

To simplify notation, assume that individual $1, \ldots, N_1$ belongs to cluster 1, individual $N_1 + 1, \ldots, N_1 + N_2$ to cluster 2 and so on. Since all clusters can be handled by similar arguments we only consider the first cluster. Let $\hat\alpha_{(1)}, \ldots, \hat\alpha_{(L)}$ denote the distinct values of $\hat\alpha_1, \ldots, \hat\alpha_{N_1}$, ordered in increasing order, and let $n_{1,k} := \#\{i : \hat\alpha_i = \hat\alpha_{(k)}\}$. Again, to simplify notation assume w.o.l.g. that $\hat\alpha_1 = \cdots = \hat\alpha_{n_{1,1}} = \hat\alpha_{(1)}$. To prove the result, we proceed in an iterative way. We will prove by contradiction that $L = 1$, i.e. all estimators of individuals from cluster 1 take the same value. Assume that $L \geq 2$.

We will now prove by contradiction that $n_{1,1} > N_1/2$. Assume that $n_{1,1} < N_1/2$. Define $\tilde\alpha_i = \hat\alpha_{(2)}$ for $i = 1, \ldots, n_{1,1}$ and $\tilde\alpha_i = \hat\alpha_i$ for $i > n_{1,1}$. By (14) Lemma 7.1 we find that

$$\begin{aligned}
&\left| \sum_i \sum_{t=1}^T \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \hat\alpha_i) - \rho_\tau(Y_{it} - X_{it}^\top \hat\beta - \tilde\alpha_i) \right| \\
\lesssim\ & n_{1,1}(\hat\alpha_{(2)} - \hat\alpha_{(1)}) \left\{ \left| \sum_t \psi_\tau(\varepsilon^\tau_{it,\hat\beta}) \right| + O_P(T^{3/4}(\log n)^{1/4}) + O_P(T^{1/2}(\log n)^{1/2}) \right\} \\
\lesssim\ & n_{1,1}(\hat\alpha_{(2)} - \hat\alpha_{(1)}) O_P(T^{3/4}(\log n)^{1/4})
\end{aligned}$$

Next, observe that by construction, under (C) and using the fact that $\sup_i |\hat\alpha_i - \alpha_{i0}| = o_P(1)$,

$$\begin{aligned}
&|\tilde\alpha_i - \tilde\alpha_j| - |\hat\alpha_i - \hat\alpha_j| = -|\hat\alpha_{(2)} - \hat\alpha_{(1)}|, && 1 \le i \le n_{1,1},\, n_{1,1} < j \le N_1 \\
& && \text{or } 1 \le j \le n_{1,1},\, n_{1,1} < i \le N_1, \\
&\Big||\tilde\alpha_i - \tilde\alpha_j| - |\hat\alpha_i - \hat\alpha_j|\Big| \le |\hat\alpha_{(2)} - \hat\alpha_{(1)}|, && 1 \le i \le n_{1,1}, N_1 < j \text{ or } 1 \le j \le n_{1,1}, N_1 < i, \\
&|\tilde\alpha_i - \tilde\alpha_j| - |\hat\alpha_i - \hat\alpha_j| = 0, && \text{else.}
\end{aligned}$$

From this we obtain

$$\begin{aligned}
&\Theta(\tilde\alpha_1, \dots, \tilde\alpha_n, \hat\beta) - \Theta(\hat\alpha_1, \dots, \hat\alpha_n, \hat\beta) \\
\lesssim\ & n_{1,1}(\hat\alpha_{(2)} - \hat\alpha_{(1)}) O_P(T^{3/4}(\log n)^{1/4}) - (\hat\alpha_{(2)} - \hat\alpha_{(1)}) \Lambda_S n_{1,1}(N_1 - n_{1,1}) \\
& + (\hat\alpha_{(2)} - \hat\alpha_{(1)}) n_{1,1} \Lambda_D O_P(n) \\
<\ & 0
\end{aligned}$$

where the last inequality holds for sufficiently large n, T since by assumption $\Lambda_D/\Lambda_S = o_P(1)$, $n\Lambda_S \gg T^{3/4}(\log n)^{1/4}$ and since we assumed $n_{1,1} < N_1/2$ so that $N_1 - n_{1,1} \ge N_1/2 \gtrsim n$. However, this is a contradiction to the fact that $\hat\alpha_1, \dots, \hat\alpha_n, \hat\beta$ minimizes Θ.

In a similar fashion, one can prove that $n_{1,L} > N_1/2$. Just define $\tilde\alpha_{N_1}, \dots, \tilde\alpha_{N_1 - n_{1,L}+1} = \hat\alpha_{(L-1)}$ and proceed as above. Since $n_{1,L} + n_{1,1} \le N_1$ and we have already proved that $n_{1,1} > N_1/2$ this leads to a contradiction with $L \ge 2$, and hence $L = 1$. All other clusters can be handled in a similar fashion and that completes the proof of the second step. □

Proof of Lemma 7.1. Apply Knight's identity (10) to find that

$$\begin{aligned}
&\sum_{t=1}^T \rho_\tau(\varepsilon^\tau_{it,\beta} - \delta) - \rho_\tau(\varepsilon^\tau_{it,\beta}) \\
=\ & -\delta \sum_t \psi_\tau(\varepsilon^\tau_{it,\beta}) + \sum_t \int_0^\delta \mathbb{E}\Big[\, \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} \,\Big]\, ds \\
& + \int_0^\delta \sum_t \Big\{ \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} - \mathbb{E}\Big[\, \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} \,\Big]\Big\}\, ds.
\end{aligned}$$

Hence it follows that

$$\begin{aligned}
&\sum_{t=1}^T \rho_\tau(\varepsilon^\tau_{it,\beta} - a_1) - \rho_\tau(\varepsilon^\tau_{it,\beta} - a_2) \\
=\ & (a_2 - a_1) \sum_t \psi_\tau(\varepsilon^\tau_{it,\beta}) + \int_{a_2}^{a_1} \sum_t \mathbb{E}\Big[\, \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} \,\Big]\, ds \\
& + \int_{a_2}^{a_1} \sum_t \Big\{ \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} - \mathbb{E}\Big[\, \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} \,\Big]\Big\}\, ds \\
=:\ & (a_2 - a_1) \sum_{t=1}^T \psi_\tau(\varepsilon^\tau_{it,\beta}) + \tilde r^{(1)}_{n,i}(a_1, a_2) + \tilde r^{(2)}_{n,i}(a_1, a_2).
\end{aligned}$$

Now by a Taylor expansion

$$\begin{aligned}
&\sup_i \Big| \sum_t \int_{a_2}^{a_1} \mathbb{E}\Big[\, \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \le 0\} \,\Big]\, ds \Big| \\
&= \sup_i \Big| \int_{a_2}^{a_1} \sum_t \mathbb{E}[F_{Y_{i1}|X_{i1}}(\beta^\top X_{it} + s | X_{it}) - F_{Y_{i1}|X_{i1}}(\beta^\top X_{it} | X_{it})] ds \Big| \lesssim T|a_1 - a_2| \max(|a_1|, |a_2|),
\end{aligned}$$

so the bound on $\tilde r^{(1)}_{n,i}(a_1, a_2)$ is established. Next define the classes of functions

$$\begin{aligned}
\mathcal{G}_1 &:= \Big\{ (y, x) \mapsto \mathbf{I}\{y - \beta^\top x \le s\} - \mathbf{I}\{y - \beta^\top x \le 0\} \;\Big|\; s \in \mathbb{R}, \beta \in \mathbb{R}^d \le B \Big\}, \\
\mathcal{G}_2 &:= \Big\{ (y, x) \mapsto \mathbf{I}\{y - \beta^\top x \le s\} \;\Big|\; s \in \mathbb{R}, \beta \in \mathbb{R}^d \Big\}.
\end{aligned}$$

Note that the class of functions $\mathcal{G}_2$ has envelope function $F \equiv 1$. Thus by Lemma 2.6.15 and Theorem 2.6.7 of van der Vaart and Wellner (1996) the class of functions $\mathcal{G}_2$ satisfies, for any probability measure Q, $N(\varepsilon, \mathcal{G}_2, L_2(Q)) \le K(1/\varepsilon)^V$

for some finite constants K, V(here, $N(\varepsilon, \mathcal{G}_2, L_2(Q))$ denotes the covering number, see Section 2.1 of van der Vaart and Wellner (1996). Moreover, $\mathcal{G}_1 \subseteq \{g_1 - g_2 | g_1, g_2 \in \mathcal{G}_2\}$, and elementary computations with covering numbers show that $N(\varepsilon, \mathcal{G}_1, L_2(Q)) \leq \tilde{K}(1/\varepsilon)^{\tilde{V}}$ for some finite constants $\tilde{V}, \tilde{K}$. Hence we find that by Theorem 2.14.9 of van der Vaart and Wellner (1996), for any $h > 0$,

$$P^*\left(\sup_{g\in\mathcal{G}_1}\frac{1}{\sqrt{T}}\left|\sum_t g(Y_{it}, X_{it}) - \mathbb{E}[g(Y_{it}, X_{it})]\right| \geq h\right) \leq \left(\frac{Dh}{\sqrt{\tilde{V}}}\right)^{\tilde{V}} e^{-2h^2}$$

for some constant D that depends only on $\tilde{K}$ (here, P^* denotes outer probability). Letting $h = \sqrt{\log n}$ and applying the union bound for probabilities we obtain

$$\begin{aligned}&\sup_i \sup_{\beta\in\mathbb{R}^d, s\in\mathbb{R}} \left|\sum_t \left\{\mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq 0\} - \mathbb{E}\left[\mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq 0\}\right]\right\}\right| \\ =\ & O_P(T^{1/2}(\log n)^{1/2}).\end{aligned}$$

Hence

$$\begin{aligned}&\sup_i \sup_{\beta\in\mathbb{R}^p} \left|\int_{a_2}^{a_1} \sum_t \left\{\mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq 0\} - \mathbb{E}\left[\mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq s\} - \mathbf{I}\{\varepsilon^\tau_{it,\beta} \leq 0\}\right]\right\} ds\right| \\ =\ & O_P(T^{1/2}(\log n)^{1/2})|a_1 - a_2|.\end{aligned}$$

Thus the bound on $\tilde{r}^{(2)}_{n,i}(a_1, a_2)$ follows and the proof is complete. □

Proof of Lemma 7.2. Observe that by Knight's identity (10)

$$\begin{aligned}&\mathbb{E}\left[\rho_\tau(Y_{it} - Z_{it}^\top\gamma) - \rho_\tau(Y_{it} - Z_{it}^\top\gamma_{0i})\right] \\ =\ & \mathbb{E}\left[\rho_\tau(Y_{it} - Z_{it}^\top\gamma_{0i} - Z_{it}^\top(\gamma - \gamma_{0i})) - \rho_\tau(Y_{it} - Z_{it}^\top\gamma_{0i})\right] \\ =\ & \mathbb{E}\left[-(\gamma - \gamma_{0i})^\top Z_{it}\psi_\tau(\varepsilon_{it}) + \int_0^{(\gamma-\gamma_{0i})^\top Z_{it}} \mathbf{I}\{\varepsilon_{it} \leq s\} - \mathbf{I}\{\varepsilon_{it} \leq 0\} ds\right] \\ =\ & \mathbb{E}\left[\int_0^{(\gamma-\gamma_{0i})^\top Z_{it}} F_{\varepsilon_{it}|X_{it}}(s|X_{it}) - F_{\varepsilon_{it}|X_{it}}(0|X_{it}) ds\right].\end{aligned}$$

Now under assumption (A2) $|F_{\varepsilon_{it}|X_{it}}(s|X_{it}) - F_{\varepsilon_{it}|X_{it}}(0|X_{it})| \leq s\overline{f'}$ a.s., and thus given (A1)

$$\mathbb{E}\left|\int_0^{(\gamma-\gamma_{0i})^\top Z_{it}} F_{\varepsilon_{it}|X_{it}}(s|X_{it}) - F_{\varepsilon_{it}|X_{it}}(0|X_{it}) ds\right| \leq \frac{\overline{f'}}{2}\mathbb{E}\left[((\gamma - \gamma_{0i})^\top Z_{it})^2\right] \leq \frac{M^2\overline{f'}}{2}\|\gamma - \gamma_{0i}\|^2.$$

This shows the upper bound in (11). For the lower bound, note that $s \mapsto F_{\varepsilon_{it}|X_{it}}(s|X_{it})$ is non-decreasing almost surely. Moreover, $f_{\varepsilon_{it}|X_{it}}(0|X_{it}) \geq f_{min}$ a.s. by (A3) and thus by (A2) and (A3) we have almost surely

$$\inf_{|s|\leq f_{min}/2\overline{f'}} f_{\varepsilon_{it}|X_{it}}(s|X_{it}) \geq \frac{f_{min}}{2}.$$

Define $\delta_i := (\gamma - \gamma_{0i})\min\{1, f_{min}/(2M\overline{f'}\|\gamma - \gamma_{0i}\|)\}$. Noting that $s \mapsto F_{\varepsilon_{it}|X_{it}}(s|X_{it})$ is non-decreasing almost surely, it follows that a.s.

$$\begin{aligned}&\int_0^{(\gamma-\gamma_{0i})^\top Z_{it}} F_{\varepsilon_{it}|X_{it}}(s|X_{it}) - F_{\varepsilon_{it}|X_{it}}(0|X_{it}) ds \\ &\qquad \geq \int_0^{\delta_i^\top Z_{it}} F_{\varepsilon_{it}|X_{it}}(s|X_{it}) - F_{\varepsilon_{it}|X_{it}}(0|X_{it}) ds \geq \frac{f_{min}}{4}(\delta_i^\top Z_{it})^2\end{aligned}$$

where the last inequality follows since by definition $|\delta_i^\top Z_{it}| \leq f_{min}/(2\overline{f'})$ a.s. Finally, under assumption (A1), $\mathbb{E}[(\delta_i^\top Z_{it})^2] \geq \|\delta_i\|^2 c_\lambda$. Summarizing, we find

$$\mathbb{E}\left[\rho_\tau(Y_{it} - Z_{it}^\top\gamma) - \rho_\tau(Y_{it} - Z_{it}^\top\gamma_{0i})\right] \geq \frac{f_{min}c_\lambda}{4}\|\delta_i\|^2 = \frac{f_{min}c_\lambda}{4}\left(\|\gamma - \gamma_{0i}\| \wedge \frac{f_{min}}{2M\overline{f'}}\right)^2$$

which proves the lower bound in (11). Thus the proof of the Lemma is complete. □

Proof of Lemma 7.3. Consider the class of functions

$$\mathcal{G}_B := \left\{(y, z) \mapsto g_\gamma(y, z) := \frac{(\rho_\tau(y - z^\top\gamma) - \rho_\tau(y))\mathbf{I}\{|z| \leq M\} + MB}{2MB} \;\middle|\; \|\gamma\| \leq B\right\}.$$

Note that by construction $0 \leq g_\gamma(y,z) \leq 1$ for all $\|\gamma\| \leq B$ and moreover $\sup_{y,z} |g_\gamma(y,z) - g_{\gamma'}(y,z)| \leq \|\gamma - \gamma'\|/(2B)$. This shows the existence of constants $V, K_B < \infty$ such that for all $i = 1, \ldots, n$ $N_{[\,]}(\varepsilon, \mathcal{G}_B, L_1(P_i)) \leq (K_B/\varepsilon)^V$ for $0 < \varepsilon < K_B$ where K_B depends on B only and P_i denotes the measure corresponding to (Y_{i1}, Z_{i1}). Thus we have by Theorem 2.14.9 of van der Vaart and Wellner (1996),

$$P^*\left(\sup_\gamma \frac{1}{\sqrt{T}} \left| \sum_t g_\gamma(Y_{it}, Z_{it}) - \mathbb{E}[g_\gamma(Y_{it}, Z_{it})] \right| \geq h \right) \leq \left(\frac{D_B h}{\sqrt{V}} \right)^V e^{-2h^2}$$

where the constant D_B depends only on K_B and P^* denotes outer probability. Set $h = \sqrt{\log n}$ to bound the right-hand side above by $o(n^{-1})$. Defining the events

$$E_{i,n} := \left\{ \sup_\gamma \frac{1}{\sqrt{T}} \left| \sum_t g_\gamma(Y_{it}, Z_{it}) - \mathbb{E}[g_\gamma(Y_{it}, Z_{it})] \right| \geq \sqrt{\log n} \right\}$$

we obtain

$$P^*(\cup_i E_{i,n}) \leq n \sup_i P^*(E_{i,n}) \leq n o(n^{-1}) = o(1).$$

Finally, note that under (A1) we have a.s.

$$\frac{\rho_\tau(Y_{it} - Z_{it}^\top \gamma) - \rho_\tau(Y_{it}) - \mathbb{E}[\rho_\tau(Y_{it} - Z_{it}^\top \gamma) - \rho_\tau(Y_{it})]}{2MB} = g_\gamma(Y_{it}, Z_{it}) - \mathbb{E}[g_\gamma(Y_{it}, Z_{it})] \quad \forall i, t.$$

This completes the proof. □

7.2. Proof of Theorem 3.2

We begin by stating a useful technical result that will be proved at the end of this section.

Lemma 7.4. *Under assumptions (A1)–(A3)*

$$\sum_{t=1}^T \rho_\tau(Y_{it} - Z_{it}^\top(\gamma_0 + \delta)) - \rho_\tau(Y_{it} - Z_{it}^\top \gamma_0)$$
$$= \delta^\top \sum_{t=1}^T Z_{it} \psi_\tau(\varepsilon_{it}) + \frac{1}{2} T \delta^\top \mathbb{E}[Z_{it} Z_{it}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{it})]\delta + r_{n,i}^{(1)}(\delta) + r_{n,i}^{(2)}(\delta)$$

where, defining $\ell_{n,T} := \max\{\log n, \log T\}$, *there exists a constant* C_2 *independent of* n, T, δ *such that*

$$\sup_i \sup_{T^{-1}\ell_{n,T}^2 \leq \|\delta\| \leq 1} \frac{|r_{n,i}^{(1)}(\delta)|}{\|\delta\|^{3/2}} = O_P(T^{1/2} \ell_{n,T}^{1/2}), \quad \sup_i |r_{n,i}^{(2)}(\delta)| \leq T C_2 \|\delta\|^3.$$

Proof of Theorem 3.2. The proof proceeds in several steps. First, we note that the 'oracle' estimation problem (6) corresponds to a classical, fixed-dimensional quantile regression with true parameter vector $(\alpha_{(01)}, \ldots, \alpha_{(0K)}, \beta_0^\top)$ and nT independent observations $(Y_{it}, \tilde{Z}_{it})$ where $\tilde{Z}_{it}^\top = (e_k^\top, X_{it}^\top)$, $i \in I_k$, $t = 1, \ldots, T$ where e_k denotes the k'th unit vector in $\mathbb{R}^K$. A straightforward extension of classical proof techniques in parametric quantile regression shows that under assumptions (A1)–(A3) and (C) the oracle estimator is asymptotically normal as claimed.

Second, we observe that by definition of the optimization problem the estimated group structure $\hat{I}_{1,\ell}, \ldots, \hat{I}_{K_\ell,\ell}$ is the same for all values of ℓ with λ_ℓ that give rise to the same number of groups. Since the value of $IC(\ell)$ depends only on $\hat{I}_{1,\ell}, \ldots, \hat{I}_{K_\ell,\ell}$, it suffices to minimize IC over those values of ℓ that correspond to different numbers of groups. Denote the distinct estimated numbers of groups by $\hat{K}_1, \ldots, \hat{K}_R$, the corresponding estimated groupings by $\hat{I}_{(1\hat{K}_r)}, \ldots, \hat{I}_{(\hat{K}_r\hat{K}_r)}$, and the corresponding values of IC by $IC_{\hat{K}_1}, \ldots, IC_{\hat{K}_R}$. By assumption (G) and Theorem 3.1, the probability of the event

$$P\left(\exists r : \hat{K}_r = K, \hat{I}_{(k\hat{K}_r)} = I_k, k = 1, \ldots, K \right) \to 1. \tag{16}$$

Hence it suffices to prove that

$$P\left(\arg\min_r IC_{K_r} = K \right) \to 1. \tag{17}$$

Once this result is established, we directly obtain

$$P\left((\hat{\alpha}_1^{IC}, \ldots, \hat{\alpha}_{\hat{K}^{IC}}^{IC}, (\hat{\beta}^{IC})^\top) = (\hat{\alpha}_{(1)}^{(OR)}, \ldots, \hat{\alpha}_{(K)}^{(OR)}, (\hat{\beta}^{(OR)})^\top) \right) \to 1,$$

and thus the asymptotic distribution of $(\hat{\alpha}_1^{IC}, \ldots, \hat{\alpha}_{\hat{K}^{IC}}^{IC}, (\hat{\beta}^{IC})^\top)$ matches that of the oracle estimator.

We will now prove (17). From Theorem 3.2 in Kato et al. (2012) we know that under (A1)–(A3) and the additional assumptions that $n \to \infty$ but T grows at most polynomially in n

$$\check{\beta} - \beta_0 = O_P((T/\log n)^{-3/4} \vee (nT)^{-1/2}).$$

If $n \to \infty$ and T grows at most polynomially in n it follows that $\check{\beta} - \beta_0 = o_P(T^{-1/2})$. Moreover, standard quantile regression arguments show that

$$\check{\alpha}_i - \alpha_{0i} = -\frac{1}{\mathbb{E}[f_{\varepsilon_{it}^\tau|X_{it}}(0|X_{it})]} \frac{1}{T} \sum_t \psi_\tau(\varepsilon_{it}^\tau) + R_{n,i}$$

where $\sup_i |R_{n,i}| = O_p\left(\left(\frac{\log T}{T} \right)^{3/4} \right)$. Next apply Lemma 7.4 to find that provided $\frac{(\log T)^3 (\log n)^2}{T} \to 0$,

$$\begin{aligned}
&\sum_{i,t} \rho_\tau(Y_{it} - Z_{it}^\top \check{\gamma}_i) - \rho_\tau(\varepsilon_{it}^\tau) \\
&= \sum_i (\check{\gamma}_i - \gamma_{0i})^\top \sum_t Z_{it} \psi_\tau(\varepsilon_{it}^\tau) + \frac{T}{2} \sum_i (\check{\gamma}_i - \gamma_{0i})^\top \mathbb{E}[Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}^\tau|X_{i1}}(0|X_{i1})](\check{\gamma}_i - \gamma_{0i}) + o_P(n) \\
&= \sum_i (\check{\alpha}_i - \alpha_{0i}) \sum_t \psi_\tau(\varepsilon_{it}^\tau) + \frac{T}{2} \sum_i (\check{\alpha}_i - \alpha_{0i})^2 \mathbb{E}[f_{\varepsilon_{i1}^\tau|X_{i1}}(0|X_{i1})] + o_P(n) \\
&= -\sum_i \frac{1}{2\mathbb{E}[f_{\varepsilon_{i1}^\tau|X_{i1}}(0|X_{i1})]} \left(\frac{1}{\sqrt{T}} \sum_t \psi_\tau(\varepsilon_{it}^\tau) \right)^2 + o_P(n) \\
&= -\sum_i \frac{\tau(1-\tau)}{2\mathbb{E}[f_{\varepsilon_{i1}^\tau|X_{i1}}(0|X_{i1})]} + o_P(n).
\end{aligned}$$

Next, observe that by asymptotic normality of the oracle estimator

$$\sup_{k=1,\dots,K} \|\hat{\gamma}_{(k)}^{(OR)} - \gamma_{(0k)}\| = O_P((nT)^{-1/2})$$

where we defined $\hat{\gamma}_{(k)}^{(OR)} := (\hat{\alpha}_{(k)}^{(OR)}, \hat{\beta}^{(OR)})$. Again applying Lemma 7.4 we obtain

$$\begin{aligned}
&\sum_k \sum_{i \in I_k} \sum_t \rho_\tau(Y_{it} - Z_{it}^\top \hat{\gamma}_{(k)}^{(OR)}) - \rho_\tau(\varepsilon_{it}^\tau) \\
&= \sum_k (\hat{\gamma}_{(k)}^{(OR)} - \gamma_{(0k)})^\top \sum_{i \in I_k} \sum_t Z_{it} \psi_\tau(\varepsilon_{it}) + nTO_P\left(\sup_k \|\tilde{\gamma}_k - \gamma_{(0k)}\|^2 \right) + o_P(n) \\
&= o_P(n).
\end{aligned}$$

Combining the results obtained so far we have

$$\sum_k \sum_{i \in I_k} \sum_t \rho_\tau(Y_{it} - Z_{it}^\top \hat{\gamma}_{(k)}^{(OR)}) - \inf_{\alpha_1,\dots,\alpha_n,\beta} \sum_{i,t} \rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha_i) \geq -\sum_i \frac{\tau(1-\tau)}{2\mathbb{E}[f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1})]} + o_P(n). \tag{18}$$

Next, let $V_n(L)$ denote the set of all disjoint partitions of $\{1, \dots, n\}$ into L subsets. Observe that by (11) we have under assumption (C)

$$\begin{aligned}
&\inf_{L<K} \inf_{J_1,\dots,J_L \in V_N(L)} \inf_{\alpha_1,\dots,\alpha_L,\beta} \left(\sum_{\ell=1}^L \sum_{i \in J_\ell} \sum_t \mathbb{E}[\rho_\tau(Y_{it} - \beta^\top X_{it} - \alpha_\ell)] - \sum_i \sum_t \mathbb{E}[\rho_\tau(\varepsilon_{it}^\tau)] \right) \\
&\geq \frac{Tc_0}{2} \min\{N_1, \dots, N_K\}(\varepsilon^2 \wedge \varepsilon_0^2).
\end{aligned}$$

Finally, note that by Lemma 7.3

$$\begin{aligned}
&\sup_{|\alpha_i| \leq B, \|\beta\| \leq B} \left| \sum_{i,t} \left(\rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha) - \rho_\tau(\varepsilon_{it}^\tau) - \mathbb{E}[\rho_\tau(Y_{it} - X_{it}^\top \beta - \alpha) - \rho_\tau(\varepsilon_{it}^\tau)] \right) \right| \\
&\leq ns_{n,1} = O_P(nT^{1/2}(\log n)^{1/2}).
\end{aligned}$$

Summarizing, we find that under (C)

$$\begin{aligned}
&\inf_{L<K}\inf_{J_1,\ldots,J_L\in V_N(L)}\inf_{\alpha_1,\ldots,\alpha_L,\beta}\Big(\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\sum_t \rho_\tau(Y_{it}-\beta^\top X_{it}-\alpha_\ell)-\sum_i\sum_t\rho_\tau(\varepsilon_{it}^\tau)\Big)\\
\geq\ & \frac{Tc_0}{2}\min\{N_1,\ldots,N_K\}(\varepsilon^2\wedge\varepsilon_0^2)-O_P(nT^{1/2}(\log n)^{1/2})\\
\gtrsim\ & nT-O_P(nT^{1/2}(\log n)^{1/2}).
\end{aligned}\tag{19}$$

The final result follows from a combination of (16) and (18) and (19). First, observe that for $K_r > K$ we have by (16), (18) and the assumptions on $p_{n,T}$, $\hat{C}$, with probability tending to one

$$IC_{K_r}-\inf_s IC_{K_s}\gtrsim p_{n,T}-n+o_P(n)\gg 0.$$

It follows that, with probability tending to one, $\arg\min_\ell IC(\ell)\leq K$. Moreover, for $K_r < K$ we have by (19) and the assumptions on $p_{n,T}$, $\hat{C}$, with probability tending to one

$$IC_{K_r}-\inf_s IC_{K_s}\gtrsim -Kp_{n,T}+nT-O_P(nT^{1/2}(\log n)^{1/2})\gg 0.$$

Hence, with probability tending to one, $K\geq\arg\min_r IC_{K_r}\geq K$ and thus (17) follows. □

Proof of Lemma 7.4. By Knight's identity (10) we have

$$\begin{aligned}
&\rho_\tau(Y_{it}-Z_{it}^\top(\gamma_{0i}+\delta))-\rho_\tau(Y_{it}-Z_{it}^\top\gamma_{0i})\\
=\ & -\delta^\top Z_{it}\psi_\tau(\varepsilon_{it})+\int_0^{Z_{it}^\top\delta}F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it})ds\\
&+\int_0^{Z_{it}^\top\delta}\mathbf{I}\{\varepsilon_{it}\leq s\}-\mathbf{I}\{\varepsilon_{it}\leq 0\}-(F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it}))ds.
\end{aligned}$$

Define

$$\begin{aligned}
r_{n,i}^{(1)}(\delta)&:=\sum_t\int_0^{Z_{it}^\top\delta}\mathbf{I}\{\varepsilon_{it}\leq s\}-\mathbf{I}\{\varepsilon_{it}\leq 0\}-(F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it}))ds\\
&\quad-\frac{T}{2}\delta^\top\mathbb{E}[Z_{i1}Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1})]+\sum_t\frac{1}{2}f_{\varepsilon_{i1}|X_{i1}}(0|X_{it})(Z_{it}^\top\delta)^2,\\
r_{n,i}^{(2)}(\delta)&:=\sum_t\Big\{\int_0^{Z_{it}^\top\delta}F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it})ds-\frac{1}{2}f_{\varepsilon_{i1}|X_{i1}}(0|X_{it})(Z_{it}^\top\delta)^2\Big\}.
\end{aligned}$$

By a Taylor expansion we obtain

$$\Big|\int_0^{Z_{it}^\top\delta}F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})ds-\frac{1}{2}f_{\varepsilon_{i1}|X_{i1}}(0|X_{it})(Z_{it}^\top\delta)^2\Big|\leq(Z_{it}^\top\delta)^3\overline{f'}\leq M^3\overline{f'}\|\delta\|^3,$$

and thus the bound on $r_{n,2}^{(i)}$ is established. Next we note that

$$\mathbb{E}\Big[\int_0^{Z_{it}^\top\delta}\mathbf{I}\{\varepsilon_{it}\leq s\}-\mathbf{I}\{\varepsilon_{it}\leq 0\}-(F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it}))ds\Big]=0$$

since the conditional expectation given Z_{it} equals zero almost surely and moreover

$$\begin{aligned}
|I_{it}(\delta)|&:=\Big|\int_0^{Z_{it}^\top\delta}\mathbf{I}\{\varepsilon_{it}\leq s\}-\mathbf{I}\{\varepsilon_{it}\leq 0\}-(F_{\varepsilon_{i1}|X_{i1}}(s|X_{it})-F_{\varepsilon_{i1}|X_{i1}}(0|X_{it}))ds\Big|\\
&\leq M^2\|\delta\|^2+M\|\delta\|\mathbf{I}\{|\varepsilon_{it}|\leq M\|\delta\|\},\\
|I_{it}(\delta)-I_{it}(\delta')|&\leq M\|\delta-\delta'\|.
\end{aligned}$$

Note that in particular for $\|\delta\|\leq 1$ we have

$$|I_{it}(\delta)|\leq M(M+1)\|\delta\|,\quad \mathbb{E}[I_{it}^2(\delta)]\leq 2(M^4+4M^3\overline{f'})\|\delta\|^3.$$

Define $c_{1,M} := M(M+1), c_{2,M} := 2(M^4 + 4M^3\overline{f'})$ and apply the Bernstein inequality to show that for any $1 \geq \|\delta\| \geq T^{-1}\ell_{n,T}^2$, $0 < a < \infty$

$$P\left(\left| \sum_t I_{it}(\delta) \right| > a\ell_{n,T}^{1/2} T^{1/2} \|\delta\|^{3/2} \right) \leq 2\exp\left(-\frac{a^2 \ell_{n,T} T\|\delta\|^3/2}{Tc_{2,M}\|\delta\|^3 + ac_{1,M}\ell_{n,T}^{1/2} T^{1/2}\|\delta\|^{5/2}/3} \right)$$
$$= 2\exp\left(-\frac{a^2 \ell_{n,T}/2}{c_{2,M} + ac_{1,M}\ell_{n,T}^{1/2}(T\|\delta\|)^{-1/2}/3} \right).$$

For $0 < a < 3\ell_{n,T}^{1/2} c_{2,M}/c_{1,M}$ the last line above is bounded by $2(n \vee T)^{-a^2/(4c_{2,M})}$. Denote by G_T a grid of values $\delta_1, \ldots, \delta_{|G_T|}$ such that $T^{-1} \leq \|\delta_j\| \leq 1$ for all $j \in G_T$ and

$$\sup_{\ell_{n,T}^2 T^{-1} \leq \|\delta\| \leq 1} \inf_{\tilde{\delta} \in G_T} \|\delta - \tilde{\delta}\| = o(T^{-2}).$$

Note that it is possible to find such a G_T with $|G_T| = O(T^{2(d+1)})$. It follows that

$$\sup_i \sup_{\ell_{n,T}^2 T^{-1} \leq \|\delta\| \leq 1} \frac{\left| \sum_t I_{it}(\delta) \right|}{\|\delta\|^{3/2}} \leq \sup_i \sup_{\delta \in G_T} \frac{\left| \sum_t I_{it}(\delta) \right|}{\|\delta\|^{3/2}} + T^{3/2} TMo(T^{-2})$$
$$= \sup_i \sup_{\delta \in G_T} \frac{\left| \sum_t I_{it}(\delta) \right|}{\|\delta\|^{3/2}} + o(T^{1/2}).$$

Finally, note that for $0 < a < 3\ell_{n,T}^{1/2} c_{2,M}/c_{1,M}$

$$P\left(\sup_i \sup_{\delta \in G_T} \frac{\left| \sum_t I_{it}(\delta) \right|}{\|\delta\|^{3/2}} > a\ell_{n,T}^{1/2} T^{1/2} \right)$$
$$\leq n|G_T| 2(n \vee T)^{-a^2/(4c_{2,M})} = O(nT^{2(d+1)})(n \vee T)^{-a^2/(4c_{2,M})}.$$

Since $\ell_{n,T} \to \infty$ we can pick a such that the last line above is $o(1)$, and hence

$$\sup_i \sup_{\delta \in G_T} \frac{\left| \sum_t I_{it}(\delta) \right|}{\|\delta\|^{3/2}} = O_P(\ell_{n,T}^{1/2} T^{1/2}).$$

Finally, observe that, denoting by $\|A\|_\infty$ the maximum norm of the entries of the matrix A,

$$\sup_i \left| -\frac{T}{2}\delta^\top \mathbb{E}[Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1})] + \sum_t \frac{1}{2} f_{\varepsilon_{i1}|X_{i1}}(0|X_{it})(Z_{it}^\top \delta)^2 \right|$$
$$= \sup_i \left| \frac{1}{2}\delta^\top \left\{ \sum_t Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1}) - \mathbb{E}[Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1})] \right\} \delta \right|$$
$$\lesssim \|\delta\|^2 \sup_i \left\| \sum_t Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1}) - \mathbb{E}[Z_{i1} Z_{i1}^\top f_{\varepsilon_{i1}|X_{i1}}(0|X_{i1})] \right\|$$
$$= \|\delta\|^2 O_P(\sqrt{T \log n})$$

where the last line follows by a straightforward application of the Hoeffding inequality. Thus the proof of Lemma 7.4 is complete. □

Acknowledgments

The research of Jiaying Gu was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC). The research of Stanislav olgushev was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.04.006.

References

Abrevaya, J., Dahl, C., 2008. The effects of birth inputs on birthweight. J. Bus. Econom. Statist. 26, 379–397.

Allman, E., Mathias, C., Rhodes, J., 2009. Identifiability of parameters in latent structure models with many observed variables. Ann. Statist. 37, 3099–3132.

Andersen, E.D., 2010. The Mosek Optimization Tools Manual, Version 6.0. Available from http://www.mosek.com.

Ando, T., Bai, J., 2016. Panel date models with grouped factor structure under unknown group membership. J. Appl. Econometrics 31, 163–191.

Arellano, M., Bonhomme, S., 2016. Nonlinear panel data estimation via quantile regressions. Econom. J. 19, 64–94.

Belloni, A., Chernozhukov, V., 2009. Least squares after model selection in high-dimensional sparse models.

Belloni, A., Chernozhukov, V., Kato, K., 2014. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. Biometrika (Oberwolfach 2012) 102 (1), 77–94.

Bester, A., Hansen, C., 2016. Grouped effects estimators in fixed effects models. J. Econometrics 190, 197–208.

Bonhomme, S., Manresa, E., 2015. Grouped pattern of heterogeneity in panel data. Econometrica 83, 1147–1184.

Chamberlain, G., 1982. Multivariate regression models for panel data. J. Econometrics 18, 5–46.

Chernozhukov, V., Fernández-Val, I., Hoderlein, S., Holzmann, H., Newey, W., 2015. Nonparametric identification in panels using quantiles. J. Econometrics 188, 378–392.

Chetverikov, D., Larsen, B., Palmer, C., 2016. IV Quantile regression for group-level treatments, with an application on the distributional effects of trade. Econometrica 84, 809–833.

Evdokimov, K., 2010. Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity. Princeton University, preprint.

Friberg, H.A., 2012. Users guide to the R-to-Mosek interface. Available from http://rmosek.r-forge.r-project.org.

Galvao, A.F., Kato, K., 2016. Smoothed quantile regression for panel data. J. Econometrics 193 (1), 92–112.

Galvao, A.F., Wang, L., 2015. Efficient minimum distance estimator for quantile regression fixed effects panel data. J. Multivariate Anal. 133, 1–26.

Hall, P., Sheather, S., 1988. On the distribution of a studentized quantile. J. R. Stat. Soc. Ser. B Stat. Methodol. 50, 381–391.

Harding, M., Lamarche, C., 2017. Penalized quantile regression with semiparametric correlated effects: an application with heterogeneous preferences. J. Appl. Econometrics 32 (2), 342–358.

Heckman, J., Singer, B., 1982. The identification problem in econometric models for duration data. In: Hildenbrand, W. (Ed.), Advances in Econometrics. Cambridge University Press.

Hocking, T., Vert, J., Bach, F., Joulin, A., 2011. Clusterpath: an algorithm for clustering using convex fusion penalties. In: Getoor, L., Scheffer, T. (Eds.), Proceeds of the International Conference of Machine Learning. Omnipress, Madison.

Hsiao, C., 2003. Analysis of Panel Data. Cambridge university press.

Kasahara, H., Shimotsu, K., 2009. Nonparametric identification of finite mixture models of dynamic discrete choices. Econometrica 77, 135–175.

Kato, K., Galvao, A., Montes-Rojas, G., 2012. Asymptotics for panel quantile regression models with individual effects. J. Econometrics 170, 76–91.

Kaufman, L., Rousseeuw, P., 2009. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Keane, M., Wolpin, K., 1997. The career decisions of young men. J. Political Econ. 105, 473–522.

Koenker, R., 2004. Quantile regression for longitudinal data. J. Multivariate Anal. 91, 74–89.

Koenker, R., 2005. Quantile Regression, Vol. 38. Cambridge university press.

Lamarche, C., 2010. Robust penalized quantile regression estimation for panel data. J. Econometrics 157, 396–408.

Leeb, H., Pötscher, B.M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142 (1), 201–211.

Lin, C.-C., Ng, S., 2012. Estimation of panel data models with parameter heterogeneity when group membership is unknown. J. Econom. Methods 1, 42–55.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the lasso. Ann. Stat. 42 (2), 413.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Cam, L.L., Neyman, J. (Eds.), Proceeds of the 5th Berkeley Symposium of Mathematical Statistics and Probability. University of California Press: Berkeley.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.

Radchenko, P., Mukherjee, G., 2017. Convex clustering via ℓ_1 fusion penalization. J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (5), 1527–1546.

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

Sun, Y., 2005. Estimation and Inference in Panel Structure Models, Working paper, University of California, San Diego.

Tan, K., Witten, D., 2015. Statistical properties of convex clustering. Electron. J. Stat. 9, 2324–2347.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.

van der Vaart, A.W., Wellner, J.A., 1996. Weak convergence and empirical processes. In: Weak Convergence and Empirical Processes. Springer.

Zhu, C., Xu, H., Leng, C., Yan, S., 2014. Convex optimization procedure for clustering: theoretical revisit. In: Advances in Neural Information Processing Systems. pp. 1619–1627.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101 (476), 1418–1429.