# Spectral clustering with variance information for group structure estimation in panel data *

Lu Yu† Jiaying Gu‡ Stanislav Volgushev§

February 18, 2024

**Abstract**

Consider a panel data setting where repeated observations on individuals are available. Often it is reasonable to assume that there exist groups of individuals that share similar effects of observed characteristics, but the grouping is typically unknown in advance. We first conduct a local analysis which reveals that the variances of the individual coefficient estimates contain useful information for the estimation of group structure. We then propose a method to estimate unobserved groupings for general panal data models that explicitly accounts for the variance information. Our proposed method remains computationally feasible with a large number of individuals and/or repeated measurements on each individual. The developed ideas can also be applied even when individual-level data are not available and only parameter estimates together with some quantification of estimation uncertainty are given to the researcher. A thorough simulation study demonstrates superior performance of our method than existing methods and we apply the method to two empirical applications.

*Keywords:* group structure estimation, spectral clustering, panel data models
*JEL classification:* C13, C23, C38

*We are grateful to Professors Huixia Judy Wang and Yingying Zhang for sending us the code for their simulations in Zhang et al. (2019a) . We thank Professor D. Millimet for kindly sharing this data with us. The data in the first empirical example is the same as in Millimet et al. (2003). We also thank the AE and anonymous referees for constructive comments on an earlier version of this manuscript that motivated us to consider the local analysis in section 2.2 and resulted in a greatly improved manuscript. Jiaying Gu acknowledges the financial support from the Social Sciences and Humanities Research Council of Canada (SSHRC). Stanislav Volgushev acknowledges a discovery grant from NSERC of Canada.

†Department of Statistical Sciences, University of Toronto. E-mail: stat.yu@mail.utoronto.ca
‡Department of Economics, University of Toronto. E-mail: jiaying.gu@utoronto.ca
§Department of Statistical Sciences, University of Toronto. E-mail: stanislav.volgushev@utoronto.ca

# 1 Introduction

Panel data models are a standard empirical tool in statistics, economics, marketing, and financial research. The conventional modeling approach is to assume that all individual heterogeneity can be summarized by an individual specific intercept, often known as the fixed effects, while assuming all covariates have a common effect among all the individuals, such that information can be pooled across individuals to gain efficiency for estimating these common parameters. However, heterogeneous responses towards observed control variables are often better supported by empirical evidence, especially as detailed individual level data becomes more available.

An increasingly popular approach to model unobserved heterogeneity in the effects of covariates on individual responses is to assume the existence of a finite number of homogeneous groups. Here, parameters in a potentially non-linear model[1] are assumed to take common values within groups but differ across groups. The main challenge is to learn the unobserved group structure from observed data. An alternative way to model unobserved heterogeneity is through latent factors (e.g., Bai (2009)). This approach also has discrete heterogeneity in the sense that a small number of unobserved factors drive the co-movement of a large number of time series. Both group pattern and factor structure are useful empirical tools, but they have different interpretations. In this paper, we focus on group patterns.

The existing literature can be roughly categorized into three categories. Methods from the first category rely on minimizing a loss function that incorporates different coefficients for all individuals combined with a penalty which encourages the coefficient estimates to be similar. Su et al. (2016) propose the classifier-LASSO (C-LASSO) approach, which is applicable to both linear and nonlinear models. Differences among individual parameters are penalized through a LASSO type penalty, and consistent grouping can be achieved if the penalty parameter is chosen properly. Wang et al. (2018) propose a Panel-CARDS penalty which extends the idea of homogeneity pursuit in Ke et al. (2015) from cross-sectional models to panel data models. Gu and Volgushev (2019) propose to use the convex clustering penalty of Hocking et al. (2011) in panel data quantile regression models with grouped individual intercepts and common slope parameters.

An alternative approach is to relate the group structure estimation problem to clustering; here clusters in the coefficient vectors correspond to latent groups of individuals. Estimating clusters has a long history in statistics and economics. Among the many clustering algorithms, the $k$-mean algorithm by MacQueen et al. (1967) is one of the most popular and commonly used methods. However, instead of directly applying $k$-mean methods on the estimated individual parameters, Lin and Ng (2012) and Bonhomme and Manresa (2015) propose to incorporate the regression loss function and re-estimate the group-specific co-

[1] Examples include quantile regression and discrete outcome models.

efficients in an iterative fashion. Originally proposed for linear regression models, this approach has also been extended to quantile regression models by Zhang et al. (2019a) and Leng et al. (2023). Further advancement of this literature has considered time varying group membership, for example Miao et al. (2020), Okui and Wang (2021) and Lumsdaine et al. (2023).

Both the penalization-based and clustering-based approaches described above require the repeated fitting of large regression models which involve all individuals and all individual-specific parameters in a large-scale minimization problem. This can be computationally costly especially for large scale datasets, which become more and more common in practice. In addition, the extensions of the $k$-means approach discussed above rely on iterative algorithms with random initialization which require repeated application with many different starting points. Motivated by those computational challenges, Chetverikov and Manresa (2022) propose an estimator for linear panel data models with grouped intercepts and common slope. Their approach is shown to guarantee the same theoretical properties as Bonhomme and Manresa (2015) but is computationally much faster. It should be pointed out however that their approach seems to be difficult to extend to non-linear panels. Wang and Su (2021) propose to use ordered individual-specific regression estimators to convert the grouping problem into a change-point detection setting and apply binary segmentation to learn the underlying group structure. This approach can be applied to both linear and nonlinear panel data models. It is computationally efficient because the individual-specific regressions only need to be estimated once rather than in an iterative fashion. They further show that by considering the spectral decomposition of an outer product of the individual parameter estimates and then applying binary segmentation on the leading eigenvectors can lead to improved group estimation.

In the present paper, we propose a novel approach that retains the computational advantages of working with individual-specific regressions but explicitly takes into account the uncertainty in the corresponding estimates. This information is particularly important in settings where different entries of a coefficient vector are estimated with different degrees of precision and hence carry varying amounts of information about the underlying population coefficients. To motivate the specific form of reweighting we use, we first conduct a simplified analysis in a local alternative framework. In the simplest case where there are only two groups in the population, we study the probability of classifying an individual to one of two groups when the separation between group centers tends to zero at a certain rate. This analysis targets a simplified iteration step which is the key ingredient of most existing iterative procedures for estimating group membership.

This local analysis motivates us to weigh the differences between coefficient estimates of different individuals by an estimated variance-covariance matrix. The resulting weighted differences can not be interpreted as a Euclidean distance. This renders many classical clustering approaches such as the vanilla $k$-means algorithm or extensions of homogeneity

pursuit and binary segmentation inapplicable. We handle this challenge by interpreting the weighted distances as a quantification of dissimilarity between individuals. With this interpretation, we can apply any clustering approach that works with general measures of dissimilarity. We consider two popular approaches: *k-medoids* Schubert and Rousseeuw (2019) and *spectral clustering* Ng et al. (2002). In simulation studies, we find that both approaches outperform existing proposals. In finite samples, the spectral clustering approach works better than the k-medoids approach and we provide high level assumptions which guarantee consistent group structure recovery asymptotically.

The remaining paper is organized as follows. In Section 2.2 we present the simple local analysis motivates our approach. Section 2.3 contains a detailed description of the proposed estimation procedure and illustrates it on several specific models that were previously considered in the literature. Section 3.1 contains theoretical guarantees on correct group estimation under high-level conditions. Those conditions are verified for several examples in Section 3.2. A simulation study is presented in Section 4. An empirical illustration analyzing the heterogeneous relationship between income and pollution level among different states using data from the United States is given in Section 5. We also apply our approach to the commuting zone summary statistics provided by Chetty and Hendren (2018) to analyze group patterns of intergenerational income mobility. Section 6 concludes. All proofs and some additional plots are deferred to the supplementary material.

# 2 Setting and proposed methodology

## 2.1 General setting

Assume that we have repeated observations $(\mathbf{x}_{it}, Y_{it})_{t=1,...,T}$ from individuals $i = 1, ..., n$. Our goal is to assign the individuals into $G^*$ groups such that individuals in the same group share a set of characteristics. For now, let $G^*$ be given, a data-driven choice of $G^*$ will be discussed at a later point.

Specifically, assume that the characteristics of individual $i$ are described by a vector of parameters $\boldsymbol{\gamma}_i$ and that we are interested in grouping individuals according to sub-vectors $\boldsymbol{\beta}_i \in \mathbb{R}^p$ of $\boldsymbol{\gamma}_i$. For instance, $\boldsymbol{\gamma}_i$ can be coefficients in a non-linear model linking the response $Y_{it}$ to the covariates $\mathbf{x}_{it}$ and $\boldsymbol{\beta}_i$ can be the full vector $\boldsymbol{\gamma}_i$, a sub-vector thereof, or simply the intercept term in a regression model. Specific examples are provided in Section 2.4.

A popular approach to such problems, pioneered by Lin and Ng (2012) and Bonhomme and Manresa (2015), is to interpret this as a clustering problem and apply an iterative approach in the spirit of Lloyd's k-means clustering algorithm. For concreteness, assume that we only have two groups and that the coefficient vectors $\boldsymbol{\gamma}_i = (\alpha_i, \boldsymbol{\beta}_i)$[2] can be estimated

[2]since the $\alpha_i$ will be left unrestricted, they correspond to the individual specific effects

by minimizing a loss function $\mathcal{L}$ via

$$(\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i) = \arg\min_{\alpha,\boldsymbol{\beta}} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_{it}, Y_{it}; \alpha, \boldsymbol{\beta}).$$

Roughly speaking, procedures in the spirit of Lin and Ng (2012); Bonhomme and Manresa (2015) consist of an initialization step where individuals are assigned to groups in a randomized fashion, followed by iterative re-assignments until convergence. In the k'th iteration step, denote the group centers from step $k-1$ by $\hat{\boldsymbol{\beta}}_1^{(k-1)}, \hat{\boldsymbol{\beta}}_2^{(k-1)}$. Now individual $i$ is assigned to group 1 iff[3]

$$\inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_{it}, Y_{it}; \alpha, \hat{\boldsymbol{\beta}}_1^{(k-1)}) < \inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_{it}, Y_{it}; \alpha, \hat{\boldsymbol{\beta}}_2^{(k-1)}). \tag{1}$$

This approach has been adopted to quantile regression by Zhang et al. (2019a). In practice, it has two potential drawbacks. First, for large $n, T$ the cost of each iteration step can be expensive. Second and more importantly, if only initial estimators $\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i$ but not individual level data are available, this approach is infeasible to implement.

Assuming that we only have access to estimators $\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i$ and covariance estimates $\hat{\Sigma}_i$ for $\hat{\boldsymbol{\beta}}_i$, a natural alternative to the iteration step is to assign individual $i$ to group 1 iff

$$\|\hat{\Sigma}_i^{-1/2}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_1^{(k-1)})\|_2 < \|\hat{\Sigma}_i^{-1/2}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_2^{(k-1)})\|_2. \tag{2}$$

For a motivation, note that the problem of assigning individual $i$ to group 1 or 2 reduces to classifying an individual into one of two classes. The rule in (2) can now be viewed as an approximate Bayes rule in classification: if $\hat{\Sigma}_i$ are fixed and $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^* \sim N(0, \hat{\Sigma}_i)$ and the population parameters $\boldsymbol{\beta}_i^*$ satisfy $\boldsymbol{\beta}_i^* \in \{\hat{\boldsymbol{\beta}}_1^{(k-1)}, \hat{\boldsymbol{\beta}}_2^{(k-1)}\}$, (2) reduces to the Bayes rule which is known to be optimal for minimizing classification error.

At this point, it is natural to wonder whether the rule in (1) or in (2) should be used. We next argue that, in a simplified but general setting, the classification error of rule (2) is (asymptotically) always at least as good as that of (1).

## 2.2 Loss functions versus weighted distances of estimators: a local analysis

To keep the presentation focused and notation simple, consider a single individual and drop the index $i$ throughout this section. Assume that the true parameter that generated the data is $\boldsymbol{\gamma}^* = (\alpha^*, \boldsymbol{\beta}^*)$ and that we want to decide based on observations $(\mathbf{x}_t, Y_t)_{t=1,\dots,T}$ whether the data are generated from parameter $(\alpha_1, \boldsymbol{\beta}_1)$ or $(\alpha_2, \boldsymbol{\beta}_2)$ where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are given

[3]Bonhomme and Manresa (2015) consider linear least squares models where the individual-specific intercepts $\alpha_i$ can be differenced out. The method presented here is a canonical generalization of their approach to non-linear models where differencing out individual effects may not be possible.

and $\alpha_1, \alpha_2 \in \mathbb{R}$ are unspecified. Let $\Gamma$ denote the parameter space and define

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) := \underset{(\alpha,\boldsymbol{\beta})\in\Gamma}{\arg\min} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha, \boldsymbol{\beta}).$$

Denote by $\hat{\Sigma}$ a consistent estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$. Define

$$\hat{k}^{BM} = 1 \iff \inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(Y_t - \alpha - \mathbf{x}_t^\top \boldsymbol{\beta}_1) < \inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(Y_t - \alpha - \mathbf{x}_t^\top \boldsymbol{\beta}_2)$$

and

$$\hat{k}^{PAM} = 1 \iff \|\hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)\|_2 < \|\hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_2)\|_2.$$

We also consider a more general approach for a general weight matrix $K_T$ that can depend on the sample size and on the available data

$$\hat{k}^{PAM,K_T} = 1 \iff \|K_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)\|_2 < \|K_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_2)\|_2.$$

This includes the case of no weighting by setting $K_T$ to be the identity matrix. We will now compare those rules in a local alternative regime where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}^*, \boldsymbol{\beta}_2 = \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}$. Assume that the loss function $\mathcal{L}$ has the following properties.

**Assumption 2.1.** *Assume that $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_T, Y_T)$ are i.i.d. and that further*

*(i) The map $m : \boldsymbol{\gamma} \mapsto \mathbb{E}[\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma})]$ is twice continuously differentiable in a neighbourhood of $\boldsymbol{\gamma}^*$ with symmetric Hessian matrix $A_{\boldsymbol{\gamma}}$ of full rank.*

*(ii) The map $g : \boldsymbol{\gamma} \mapsto \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma})$ is differentiable at $\boldsymbol{\gamma}^*$ on a set $\mathcal{Z}$ such that $\mathbb{P}((\mathbf{x}_t, Y_t) \in \mathcal{Z}) = 1$ and there exists a measurable function $\dot{g}$ such that almost surely $|\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}_1) - \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}_2)| \leq \dot{g}(\mathbf{x}_t, Y_t)\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|$ for all $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ in a neighborhood of $\boldsymbol{\gamma}^*$ and $\mathbb{E}[\dot{g}(\mathbf{x}_t, Y_t)^2] < \infty$.*

*(iii) For any $\boldsymbol{\beta}$ in a neighbourhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ the function $\alpha \mapsto m(\alpha, \boldsymbol{\beta})$ has a well separated (uniformly in $\boldsymbol{\beta}$) global minimizer $\alpha^*_{\boldsymbol{\beta}}$, i.e. for every $\varepsilon > 0$ we have*

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}} \inf_{|\alpha - \alpha^*_{\boldsymbol{\beta}}| > \varepsilon} \big(m(\alpha, \boldsymbol{\beta}) - m(\alpha^*_{\boldsymbol{\beta}}, \boldsymbol{\beta})\big) > 0.$$

*(iv) The value $\boldsymbol{\gamma}^*$ is in the interior of the parameter space $\Gamma$. Either the parameter space $\Gamma$ is compact or the parameter space is convex and the function $\boldsymbol{\gamma} \mapsto \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma})$ is convex almost surely.*

It is routine to verify that all of the above conditions hold for two important examples that we will discuss throughout this paper: quantile regression and logistic regression. More

generally, parts (i) and (ii) of the assumptions are fairly mild and standard conditions for establishing asymptotic normality and expansions for m-estimators, see for instance Theorem 5.23 and the discussion around it in van der Vaart (2000). Conditions (iii) and (iv) are added because the proof relies not only on expansions for the original estimator but also for the minimizer of the perturbed objective $\sum_{t=1}^{T} \mathcal{L}(Y_t - \alpha - \mathbf{x}_t^\top \boldsymbol{\beta})$ where $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$. We have opted for simple to state and verify conditions rather than the most general possible ones. The proof of Theorem 2.1 reveals that it is the expansions (25)–(28) in the proof rather than the specific conditions we state above that are needed to establish this result. Such expansions can also be established for data with serial dependence but we do not pursue this direction here as it does not add any insights to our main message.

To state the next result introduce some additional notation. For square matrices $M$ of dimension $p+1$ consider the following block structures

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

with $M_{11} \in \mathbb{R}$.

**Theorem 2.1.** *Assume that Assumption 2.1 holds and that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}^*, \boldsymbol{\beta}_2 = \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}$, $\boldsymbol{\Delta} \neq 0$. Let $A = A_{\boldsymbol{\gamma}^*}$, $B = Var(\nabla_{\boldsymbol{\gamma}} \mathcal{L}(\mathbf{x}, Y; \boldsymbol{\gamma}^*))$ and assume that $B$ is of full rank. Then $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, \Sigma_{\boldsymbol{\beta}})$ where $\Sigma_{\boldsymbol{\beta}} = [A^{-1}BA^{-1}]_{22}$. Assume that $\hat{\Sigma} = \Sigma_{\boldsymbol{\beta}} + o_{\mathbb{P}}(1)$. Then*

$$\lim_{T\to\infty} \mathbb{P}\Big(\hat{k}^{PAM} = 1\Big) \geq \lim_{T\to\infty} \mathbb{P}\Big(\hat{k}^{BM} = 1\Big). \tag{3}$$

*Define $D = [[A^{-1}]_{22}]^{-1}, C := B_{22} + \frac{B_{11}}{A_{11}^2} A_{21}A_{21}^\top - 2\frac{A_{21}B_{21}^\top}{A_{11}}$. Equality in (3) holds if and only if $C^{1/2}\boldsymbol{\Delta}$ is a scalar multiple of $C^{-1/2}D\boldsymbol{\Delta}$. Further, if $K_T \to K$ entry-wise in probability for a fixed matrix $K$ with finite entries*

$$\lim_{T\to\infty} \mathbb{P}\Big(\hat{k}^{PAM} = 1\Big) \geq \lim_{T\to\infty} \mathbb{P}\Big(\hat{k}^{PAM,K_T} = 1\Big).$$

A similar result under even weaker conditions continues to hold if there is no individual-specific $\alpha$ and all parameters are estimated globally. The proof of this result is similar in spirit but even simpler and we omit the details for the sake of brevity.

Note that when $\mathcal{L}$ is a correctly specified negative log-likelihood function, standard regularity conditions yield $A = B$ which further implies $C = D$ by the block matrix inversion formula. In this case $C^{-1/2}D = C^{1/2}$ so the asymptotic probabilities for rules (1) and (2) selecting the correct center are equal for any $\boldsymbol{\Delta}$. Correct specification of $\mathcal{L}$ is sufficient but not necessarry. The equality $C = D$ continues to hold in the case where $A$ is a scalar multiple of $B$ which is the case in least squares or quantile regression with homoscedastic errors, for instance. However, in general models such as quantile regression or ordinary least squares estimation with heteroscedasticity or in the presence of temporal

dependence, $A$ is not a scalar multiple of $B$ in general and thus also $C \neq D$. Since rule (2) is always at least as good as (1) asymptotically, this suggests that (2) would be preferable whenever the asymptotic covariance matrix can be estimated consistently, even when (1) is feasible.

The second statement of Theorem 2.1 implies that the proposed scaling with $\hat{\Sigma}^{-1/2}$ is asymptotically optimal among all possible choices of scale matrix that converge to a fixed matrix.

Although the results presented above only work in a very idealized setting and can not be directly utilized to analyze the performance of rules (1) and (2) when applied inside an iterative procedure, the findings strongly suggest that using the objective function in iteration for group centers might not be optimal from a statistical perspective. Instead, using information on the (asymptotic) variance of the estimators $\hat{\boldsymbol{\beta}}_i$ can lead to more efficient procedures. This motivates the ideas in the following section.

**Remark 2.1.** *The key to proving the first statement of Theorem 2.1 is an asymptotic expansion for the probabilities appearing in* (3)*. Specifically, we derive the following limits*

$$P(\hat{k}^{PAM} = 1) \to \Phi(\|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta_{\boldsymbol{\beta}}\|_2/2).$$

*in equation* (30) *and*

$$P(\hat{k}^{BM} = 1) \to \Phi\Big(\frac{\Delta\Big[[A^{-1}]_{22}\Big]^{-1}\Delta}{2(\Delta^\top C\Delta)^{1/2}}\Big)$$

*in equation* (31) *in the proof of Theorem 2.1. This is where the matrix $C$ comes into play. Given those expansions,* (3) *follows by an application of the Cauchy-Schwarz inequality as follows*

$$\frac{\Delta^\top D\Delta}{(\Delta^\top C\Delta)^{1/2}} = \frac{\Delta^\top C^{1/2}C^{-1/2}D\Delta}{(\Delta^\top C\Delta)^{1/2}} \leq \frac{\|\Delta^\top C^{1/2}\|_2\|C^{-1/2}D\Delta\|_2}{(\Delta^\top C\Delta)^{1/2}} = (\Delta^\top\Sigma_{\boldsymbol{\beta}}^{-1}\Delta)^{1/2}.$$

*This inequality is strict unless $C^{1/2}\Delta$ is a scalar multiple of $C^{-1/2}D\Delta$.*

## 2.3 Proposed methodology through the lens of clustering

The discussion up to this point focused on variants of the k-means algorithm for grouping individuals. However, k-means is not the only clustering method which is available and other approaches have been observed to have superior performance in certain settings. Many methods of this type work with general measures of dissimilarity between units and attempt to cluster units that are most similar to each other. Given the developments in the previous sections, a natural measure of dissimilarity is given by

$$\hat{V}_{ij} := \|\hat{\Sigma}_{i,j}^{-1/2}(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j)\|_2\,, \tag{4}$$

where typically $\hat{\Sigma}_{i,j} = \hat{\Sigma}_i + \hat{\Sigma}_j$ and $\hat{\Sigma}_i$ estimates the variance of $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}$. Note that for consistent estimators $\hat{\boldsymbol{\beta}}_i$, $\hat{\Sigma}_i$ will typically converge to zero. This measure of dissimilarity can be computed based on summary statistics and variance estimates and does not require individual level data. The importance of taking variance information into account was illustrated in a simplified setting in Theorem 2.1 and is also confirmed in our simulations. As pointed out by the Associate Editor, using covariance estimates or diagonal versions thereof for $\hat{\Sigma}_i$ has the added benefit of making the procedure scale invariant.

Two popular clustering approaches in the literature that work with general measures of dissimilarity are *k-medoids* Schubert and Rousseeuw (2019) and *spectral clustering* Ng et al. (2002); Chung and Graham (1997); von Luxburg (2007). Similarly to $k$-means clustering, the $k$-medoids problem is NP-hard to solve exactly. In practice, approximate solutions to this problem are obtained by employing the algorithm *Partitioning Around Medoids* (PAM) Reynolds et al. (2006); Schubert and Rousseeuw (2019); Kaufman and Rousseeuw (2005). We refer to (Kaufman and Rousseeuw, 2005, Section 4.1, Chapter 2) for more details about the PAM algorithm. As we observe in simulations, using the PAM algorithm with dissimilarity measure (4) can already lead to substantial gains relative to the iterative k-means style approaches of Lin and Ng (2012); Bonhomme and Manresa (2015); Zhang et al. (2019a). However, extensive simulations showed that in all settings considered spectral clustering leads to even more accurate group estimation than PAM, and hence we focus on spectral clustering in the theoretical developments that follow. Simulation evidence for the superiority of spectral clustering over to PAM is presented in Section 4.

Since there are many variations of spectral clustering that are available in the literature, a detailed description of the specific version we use is given in Algorithm 1[4].

**Algorithm 1** Spectral Clustering

**Input**: Number of clusters $G^*$, dissimilarity matrix $\hat{V} := (\hat{V}_{ij})$ computed in (4).
**Output**: Clusters $\hat{I}_1, \ldots, \hat{I}_{G^*}$.

1: Compute the empirical adjacency matrix $\hat{A} \in \mathbb{R}^{n \times n}$ with entries $\hat{A}_{ij} := e^{-\hat{V}_{ij}}$ for $i \neq j$ and $\hat{A}_{ij} = 1$ for $i = j$.
2: Compute the empirical degree matrix $\hat{D} := \text{diag}(\hat{D}_1, \ldots, \hat{D}_n)$, where $\hat{D}_i := \sum_{j=1}^{n} \hat{A}_{ij}, i = 1, \ldots, n$.
3: Calculate the normalized graph Laplacian $\hat{L} := \hat{D}^{-1/2}(\hat{D} - \hat{A})\hat{D}^{-1/2}$.
4: Find $G^*$ orthonormal eigenvectors corresponding to the $G^*$ smallest eigenvalues of $\hat{L}$, and form the matrix $\hat{Z} \in \mathbb{R}^{n \times G^*}$ by stacking those vectors in columns. Normalize the rows of $\hat{Z}$, to have $\ell^2$-norm 1 and denote the resulting matrix by $\hat{U}$.
5: Apply standard $k$-means clustering with $G^*$ clusters taking the rows of $\hat{U}$ as input vectors, and return the clusters $\hat{I}_1, \ldots, \hat{I}_{G^*}$.

To intuitively understand the motivation behind the above algorithm observe that the dissimilarities $\hat{V}_{ij}$ can be expected to be large if individuals $i, j$ are from different groups.

[4]We do not claim any novel contributions to this specific algorithm, the details and explanation are presented here for the reader's convenience.

In the limit $T \to \infty$ those distances will tend to infinity, and thus $\hat{A}_{ij} \approx 0$ whenever $i, j$ are from different groups. Similarly, $\hat{V}_{ij}$ can be expected to be bounded when $i, j$ are in the group, and thus $\hat{A}_{ij}$ will usually be bounded away from zero for such pairs. Thus after rearranging the order of individuals we see that $\hat{V}_{ij}$ will be approximately block diagonal with non-zero entries in the blocks. It is now straightforward to see that $\hat{L}$ will have exactly $G^*$ zero eigenvalues if there are $G^*$ such blocks and all other eigenvalues will be strictly positive. Moreover, the eigen-space corresponding to zero eigenvalues will have an orthogonal basis consisting of vectors that have non-zero entries in the exact components corresponding to different groups, see also the discussion surrounding equation (32) and Lemma 9.1 in the supplementary material. For a more detailed discussion of the intuition and alternative formulations of the spectral clustering algorithm see von Luxburg (2007) and the literature cited therein. Although the last step of the algorithm uses the standard $k$-means algorithm, we note that it is applied on the rows of $\hat{U}$ which is a standard clustering problem with $n$ data points in Euclidean space. No refitting of models on individual level or large scale models as in Bonhomme and Manresa (2015) is required.

Some additional comments on specific choices that we made in Algorithm 1 are in order. First, in step (1), we apply an exponential kernel to the dissimilarity matrix. Other monotone transformations can be used, for instance the Gaussian kernel is another popular choice. Our simulation exercise confirms that both the exponential kernel and the Gaussian kernel perform similarly. Second, in step (3), we apply a normalization to the graph Laplacian for the spectral clustering analysis. A line of seminal works (von Luxburg et al. (2004) and von Luxburg et al. (2008)) investigate the convergence of the normalized and unnormalized versions of the popular spectral clustering algorithm. They demonstrate that the normalized spectral clustering converges under very general conditions, while the unnormalized spectral clustering is only consistent under strong additional assumptions, which are not always satisfied in real data. These works give strong evidence for the superiority of normalized spectral clustering.

**Remark 2.2.** *Wang and Su (2021) also observe that the spectral decomposition of a certain matrix that is derived from individual-specific estimators contains information on the latent group structure. However, there are several crucial differences between their and our approach. Most importantly, we explicitly take into account the uncertainty that is associated with individual-specific estimators while Wang and Su (2021) work directly with raw estimators. Moreover, Wang and Su (2021) do not apply spectral clustering directly but rather use certain eigenvectors as input to a binary segmentation algorithm. For a simulation-based comparison with that method, see section 4.1.*

*The idea to use spectral clustering for grouping different entities also appeared in van Delft and Dette (2021). The setting in the latter paper is very different from ours since van Delft and Dette (2021) consider grouping locally stationary functional time series and do not take into account estimation uncertainty when constructing their dissimilarity mea-*

*sure between observations. Still, some parts of our theoretical analysis under high-level assumptions are related to theirs, additional comments on this can be found in Remark 3.1.*

So far we discussed an algorithm for assigning individuals to $G^*$ groups for any given $G^*$. In some settings, $G^*$ will be chosen based on domain knowledge about the problem at hand. If no such knowledge is available, we propose to select the $G^*$ that maximizes the relative eigen–gap (von Luxburg (2007)) of a modified graph Laplacian $\tilde{L}$. More precisely, consider the scaled dissimilarity $\tilde{V}_{ij} := \frac{2}{\sqrt{\log n \log T}} \hat{V}_{ij}$. Use $\tilde{V}_{ij}$ as input to Algorithm 1 and obtain $\tilde{L}$ as output from step 3 of that algorithm. Consider the values $\tilde{\lambda}_i := 1 - \hat{\lambda}_i, i = 1, \ldots, n$, with $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_n$ denoting the ordered eigenvalues of $\tilde{L}$. The estimated number of groups is

$$\hat{G} = \underset{g=1,\ldots,n-1}{\arg\max} \frac{|\tilde{\lambda}_{g+1} - \tilde{\lambda}_g|}{\tilde{\lambda}_{g+1}}, \tag{5}$$

The motivation for using the scaling in $\tilde{V}_{ij}$ is that, under technical assumptions made later, this scaling ensures $\tilde{V}_{ij} \to 0$ for all $i, j$ in the same group. Without this scaling, the heuristic tends to have a small probability of not selecting a correct number of groups as $T$ increases.

Similar heuristic eigen-gap methods for estimating the number of groups can also be found in van Delft and Dette (2021); John et al. (2020); Little et al. (2020), among many others.

**Remark 2.3.** *There are at least two other popular approaches to selecting the number of groups or equivalently the number of clusters. The first type of method combines cross-validation with the idea that "true" cluster assignment should be stable under small perturbations of the data. This idea was exploited in Wang (2010) for selecting the number of clusters in a general setting and adapted by Zhang et al. (2019a) to selecting the number of groups for panel data quantile regression. However, as pointed out in Ben-David et al. (2006), methods that select the number of clusters based on stability can fail for certain cluster configurations. One such example will be given in the simulation section, see Model 2 in section 3.2.2 . The second drawback of such methods is that clustering stability can only be defined when there are at least two clusters. Hence, by construction, stability methods always select at least two clusters and fail if there is only a single cluster in the data.*

*The second method uses information criteria which select the number of clusters that maximize a sum of objective function plus penalty, see for instance Su et al. (2016); Gu and Volgushev (2019); Wang and Su (2021) among many others. The main drawback of such approaches is that information criteria need to be derived case by case as they differ depending on the specific form of the objective function making them difficult to use for applied researchers. We note that this is different from the classical setting involving AIC and BIC in a maximum likelihood framework where only the number of parameters in the model matters. Moreover, computation of such information criteria typically requires access to raw data which might not always be available as in our second application. The infor-*

*mation criteria method also involves the heaviest computation burden because to construct the information criteria statistics, all candidate models with varying values of G need to be estimated which can be costly (See computation time comparison in Section 4.1).*

*We also conduct an extensive simulation comparing different methods of selecting the number of groups in Section 4.1 and 4.3. Results show that our heuristic approach works reasonably well in all settings considered. Unsurprisingly, we also find that there is no universally dominating method.*

### 2.4 Examples

The setting above is generic and so far we did not assume anything about the specific structure of the estimators. In the remainder of this section, we provide several illustrative examples of model specifications that were considered previously and show how those examples fit into the proposed framework.

**Example 2.1** (*Logistic regression regression with individual-specific intercepts and grouping on slopes*)**.** *Consider binary responses* $Y_{it} \in \{0, 1\}$ *and assume that*

$$\mathbb{P}(Y_{it} = 1) = \frac{\exp(\alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta}_i)}{1 + \exp(\alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta}_i)} = \frac{\exp(\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i)}{1 + \exp(\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i)} \,,$$

*where* $\mathbf{z}_{it}^\top = (1, \mathbf{x}_{it}^\top)$ *and* $\boldsymbol{\gamma}_i^\top = (\alpha_i, \boldsymbol{\beta}_i^\top)$*. We leave the* $\alpha_i \in \mathbb{R}$ *unrestricted and assume that certain sub-vectors of* $\boldsymbol{\beta}_i \in \mathbb{R}^p$ *have a group structure.*

Su et al. (2016) considers a similar model; they assume a Gaussian link function for the binary response. Ando et al. (2022) also considers the logit model with individual specific slope coefficients and a factor structure on the individual fixed effects. Their way of modeling unobserved heterogeneity is different from ours as we focus on group patterns.

**Example 2.2** (*Quantile regression with individual-specific intercepts and grouping on slopes*)**.** *Given the observations are* $(\mathbf{x}_{it}, Y_{it})$*, assume that the conditional quantile function of the response* $Y_{it}$ *given covariates* $\mathbf{x}_{it}$ *for individual* $i$ *satisfies*

$$q_{i,\tau}(\mathbf{z}_{it}) = \alpha_i(\tau) + \mathbf{x}_{it}^\top \boldsymbol{\beta}_i(\tau) = \mathbf{z}_{it}^\top \boldsymbol{\gamma}_i(\tau) \,,$$

*where* $\alpha_i(\tau) \in \mathbb{R}$ *are unrestricted and we search for a group structure on* $\boldsymbol{\beta}_i(\tau) \in \mathbb{R}^p$*.*

This setting was also considered in Zhang et al. (2019a), Leng et al. (2023). Zhang et al. (2019a) propose an iterative algorithm based on the $k$-mean algorithm in Bonhomme and Manresa (2015) to learn group structure. Leng et al. (2023) use a $k$-means type of iterative algorithm, but allow for time fixed effect while grouping both the individual fixed effects and the slope coefficients. This model will be considered in Section 5 where coefficients of the panel quantile regression model will be utilized to analyze heterogeneous relationship between income and pollution level among different states in the US.

**Example 2.3** (*Quantile regression with joint slope and grouping on intercepts*)**.** *Assume that the conditional quantile function of response $Y_{it}$ given covariates $\mathbf{x}_{it}$ for individual $i$ is*

$$q_{i,\tau}(\mathbf{x}_{it}) = \alpha_i(\tau) + \mathbf{x}_{it}^\top \boldsymbol{\beta}(\tau)\,,$$

*where the vector of slope coefficients $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$ is assumed to be the same across individuals.*

This model was first considered in Koenker (2004), who proposed to regularize the individual fixed effects via $\ell_1$ penalization. Lamarche (2010) considers the optimal choice of the penalty parameters in this approach. There has been an active literature on panel data quantile regression, mainly focusing on estimation of the common parameters $\beta(\tau)$ (e.g., Kato et al. (2012), Galvao and Kato (2016), Harding and Lamarche (2017) and Galvao et al. (2020)). Zhang et al. (2019b) and Gu and Volgushev (2019) consider group structure on $\alpha_i(\tau) \in \mathbb{R}$.

# 3 Theoretical Analysis

## 3.1 Generic spectral clustering results

In this section, we provide high-level conditions on the estimators $\hat{\boldsymbol{\beta}}_i \in \mathbb{R}^p$ and $\hat{\Sigma}_{i,j} \in \mathbb{R}^{p\times p}$ which ensure that the correct group structure is recovered with probability tending to one as $n, T$ tend to infinity. Formally, assume that the true coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n$ take $G^*$ different values, say $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_{G^*}^*$ and the true group membership is given by

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_k^* \quad \Leftrightarrow \quad i \in I_k^*, \quad k = 1, ..., G^*\,,$$

where $I_k^* \subseteq \{1, \dots, n\}, k = 1, \dots, G^*$ denote the true underlying groups. Naturally, we assume $I_k^* \cap I_\ell^* = \emptyset$ for $k \neq \ell$. We begin by providing an analytical non-asymptotic result which guarantees perfect classification in terms of certain abstract quantities. More precisely, define

$$\begin{aligned} A_{1,max} &:= \max_{i,j \text{ in different groups}} \hat{A}_{ij}\,, \\ A_{0,min} &:= \min_{i,j \text{ in same group}} \hat{A}_{ij} \\ A_{0,max} &:= \max_{i,j \text{ in same group}} \hat{A}_{ij}\,. \end{aligned}$$

**Theorem 3.1.** *A sufficient condition for perfect classification is*

$$\frac{A_{1,max}}{A_{0,min}} \sqrt{\frac{A_{0,max}^3}{A_{0,min}^3}} \leq 2^{-8.5} (nG^*)^{-1/2} \sqrt{\frac{\min_k |I_k^*|^3}{n \max_k |I_k^*|^2}} \tag{6}$$

Theorem 3.1 holds for fixed $n, T$ and is proved in a purely analytic way. The result

does not assume anything about temporal or cross–sectional dependence. On a high level, this result corresponds to intuition as the inequality in (6) becomes more difficult to satisfy for a larger number of groups $G^*$ or when groups have more unbalanced sizes leading to a larger ratio $\max_k |I_k^*| / \min_k |I_k^*|$. Having more individuals (larger $n$) also intuitively makes the problem harder. In order to achieve perfect classification, a large minimal dissimilarity between individuals from different groups, i.e. a small $A_{1,max}$, relative to $A_{0,min}$, is required. The ratio $\frac{A_{0,max}}{A_{0,min}}$ describes the spread of similarity measures among individuals that belong to the same group. Having a large spread here makes the problem harder, which again corresponds to intuition. Note that this is only a sufficient condition, and sharper results might be possible. However, we are not aware of any necessary and sufficient conditions guaranteeing the success of spectral clustering or sharp expansions for the proportion of correctly grouped units.

**Remark 3.1.** *The proof relies on the type of arguments that appeared in earlier work on spectral clustering, in particular Ng et al. (2002), von Luxburg (2007) and van Delft and Dette (2021). However, the setting we consider is different from any of the works mentioned above and the arguments need to be modified accordingly. The work of van Delft and Dette (2021) is closest in spirit, but our analysis is complicated by the fact that we allow the number of individuals $n$ to diverge while the number of entities to be clustered was fixed in van Delft and Dette (2021). In order to deal with this complication, we leverage the fact that our construction of the similarity matrix gives rise to the different order for the diagonal blocks and off-diagonal blocks of the empirical Laplacian matrix. Taking advantage of this difference in order together with spectral information contained in the diagonal blocks of the empirical Laplacian matrix allows us to handle a diverging number of individuals.*

Below, we will provide more specific assumptions on the minimal separation of group centers and quality of initial estimators which guarantee that the probability of the events in (6) tend to one. In the assumptions below, we allow for data from triangular arrays where the values of $\boldsymbol{\beta}_i$ and $\Sigma_{i,j}$ change with $n, T$. To keep the presentation simple this is not emphasized in the notation. We also allow the number of groups $G^*$ to grow with $n$.

**Assumption 3.1.** *The estimators $\hat{\boldsymbol{\beta}}_i$ are uniformly consistent with rate $a_{n,T}$, i.e.*

$$a_{n,T} := \sup_{i \in \{1,\dots,n\}} \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|_2 = o_{\mathbb{P}}(1).$$

**Assumption 3.2.** *There exists a sequence $b_T \to \infty$ and matrices $\Sigma_{i,j}$ (which may depend on $n, T$) such that*

$$\sup_{i,j} \left| \left| \left| b_T \hat{\Sigma}_{i,j} - \Sigma_{i,j} \right| \right|_2 = o_{\mathbb{P}}(1)\,,$$

*where $|||\cdot|||_2$ denotes the spectral norm. Moreover, assume*

$$0 < m < \lambda_{\min}(\Sigma_{i,j}) \leq \lambda_{\max}(\Sigma_{i,j}) < M < \infty \quad \forall i \neq j \in \{1, \dots, n\} \tag{7}$$

*with some fixed constants* $0 < m \leq M < \infty$ *that do not depend on* $n, T$.

Assumptions 3.1 and 3.2 impose minimal restrictions on the quality of the initial estimates $\hat{\boldsymbol{\beta}}_i$ and $\hat{\Sigma}_{i,j}$. We emphasize that the matrices $\Sigma_{i,j}$ in Assumption 3.2 are not required to be equal to the true asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$ for the theory to work. This is a useful result because in some environments researchers only have access to individual estimates and the associated coordinate-by-coordinate standard deviation, but the covariances estimate are missing. In such cases, our method can still be used by setting the off-diagonal elements of $\hat{\Sigma}_{i,j}$ to zero. Assumption 3.2 will hold provided that the variance estimators on the diagonal converge to non-negative values. While setting off-diagonal entries to zero might not be optimal in the asymptotic setting of Theorem 2.1, simulations indicate that in finite samples the performance can be close to using consistent estimates of the covariance. When covariances are difficult to estimate, using only the diagonal entries can even enhance finite sample performance as we will see in Section 4.1. Similarly, this assumption can be satisfied if there is dependence across individuals but this dependence is ignored when estimating the covariance of $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$. Again, ignoring this dependence will not lead to procedures with best possible performance but might work reasonably well if the dependence across individuals is mild.

In all examples we consider later the individuals will be assumed independent and the estimators $\hat{\boldsymbol{\beta}}_i$ will satisfy $\sqrt{T}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, \Sigma_i), i = 1, \ldots, n$. By independence among individuals, the weak convergence above holds jointly for any given pair of individuals and the corresponding limits will be independent. In this case, we will set $b_T = T$, $\Sigma_{i,j} := \Sigma_i + \Sigma_j$, $\hat{\Sigma}_{i,j} := \hat{\Sigma}_i + \hat{\Sigma}_j$ where $T\hat{\Sigma}_i$ will be consistent estimators of $\Sigma_i$.

The bound in $a_{n,T}$ is uniform over a potentially growing number of individuals $n$ and typically be of the form $a_{n,T} = \mathcal{O}_{\mathbb{P}}(\sqrt{T^{-1} \log n})$ where the additional $\sqrt{\log n}$ factor is to ensure uniformity.

We now have the following result

**Theorem 3.2.** *Under Assumptions 3.1, 3.2 let* $\Delta_{min} := \min_{k \neq \ell} \|\boldsymbol{\beta}^*_k - \boldsymbol{\beta}^*_\ell\|_2$. *Assume that* $a_{n,T} = o_{\mathbb{P}}(\Delta_{min})$, $n \geq 3$ *and*

$$\log n = o(b_T^{1/2} \Delta_{min}). \tag{8}$$

*Then the true group structure is recovered with probability tending to one as* $T \to \infty$.

In order to achieve perfect classification with probability going to one, Theorem 3.2 requires lower bounds on the minimal separation $\Delta_{min}$ which is required to grow faster than the uniform estimation error and than $b_T^{-1/2} \log n$. In the special setting discussed above the Theorem where $b_T = T, a_{n,T} = \mathcal{O}_{\mathbb{P}}(\sqrt{T^{-1} \log n})$, this corresponds to assuming that $\Delta_{min} \gg T^{-1/2} \log n$. For groups with fixed separation across centers where $\Delta_{min}$ is a constant, this leads to the requirement $\log n = o(T^{1/2})$ which allows the number of individuals to grow very quickly with $n$. If the minimal separation tends to zero, the requirements on $\log n$ relative to $\sqrt{T}$ become more stringent.

Given the non-asymptotic bound in (6), it would also be possible to conduct a more detailed analysis in the case where the orders of $\Delta_{min}$ and $a_{n,T}$ match but $\Delta_{min}$ is sufficiently large so as to dominate a constant multiple of $a_{n,T}$ with a certain probability. Such an analysis would reveal more nuanced view on the role of $G^*$ and $\max_k |I_k^*|, \min_k |I_k^*|$ but does not lead to any specific insights except that large $G^*$ and imbalanced groups make the problem harder.

### 3.2 Verification of high level conditions for specific examples

In this section, we provide specific conditions in Example 2.1–Example 2.2 which guarantee that the high-level conditions 3.1 and 3.2 are satisfied. The set of examples that we consider is by no means exhaustive for the possible applications of our methodology. Rather, it is intended as a demonstration that our high-level conditions can be verified in several different settings including the presence of individual-specific and joint parameters, binary outcomes, and non-smooth objective functions.

#### 3.2.1 Logistic regression with individual-specific intercepts and grouping on the slopes (Example 2.1)

The coefficient vector $\boldsymbol{\gamma}_i^\top := (\alpha_i, \boldsymbol{\beta}_i^\top)$ is estimated via maximum likelihood, i.e.

$$\hat{\boldsymbol{\gamma}}_i := \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}}{\arg\max} \frac{1}{T} \sum_{t=1}^{T} \Big[ Y_{it} \mathbf{z}_{it}^\top \boldsymbol{\gamma} - \log(1 + \exp(\mathbf{z}_{it}^\top \boldsymbol{\gamma})) \Big], \quad i = 1, \dots, n\,.$$

The exact form of the asymptotic variance differs depending on whether the data exhibit temporal dependence. We begin by discussing the case that the observations $(\mathbf{x}_{it}, Y_{it})$ are i.i.d. across $t$ and independent across $i$ and discuss the case with temporal dependence across $t$ later in this section. Throughout, the values of $\boldsymbol{\gamma}_i^*$ are allowed to depend on $n, T$.

Recall that in the i.i.d. case under standard assumptions the estimator $\hat{\gamma}_i$ is asymptotically normal with asymptotic variance given by

$$\widetilde{\Sigma}_i = \left( \mathbb{E}\Big[ \frac{e^{\mathbf{z}_{i1}^\top \boldsymbol{\gamma}_i^*}}{(1 + e^{\mathbf{z}_{i1}^\top \boldsymbol{\gamma}_i^*})^2} \mathbf{z}_{i1} \mathbf{z}_{i1}^\top \Big] \right)^{-1}.$$

The canonical plug-in estimator of $\widetilde{\Sigma}_i$ takes the form

$$\hat{\widetilde{\Sigma}}_i = \left( \frac{1}{T} \sum_{t=1}^{T} \frac{e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i}}{(1 + e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i})^2} \mathbf{z}_{it} \mathbf{z}_{it}^\top \right)^{-1}.$$

Denote by $\check{\Sigma}_i$ the lower $p \times p$ sub-matrix of $\hat{\widetilde{\Sigma}}_i$. Then we set

$$\hat{\Sigma}_{i,j} := T^{-1} (\check{\Sigma}_i + \check{\Sigma}_j)\,. \tag{9}$$

Consider the following assumptions.

**Assumption 3.3.** *Assume that for a constant $L > 0$ independent of $i, n, T$*

$$1/L < \left\{\lambda_{\min}\big(\mathbb{E}\big[\mathbf{z}_{i1}\mathbf{z}_{i1}^\top\big]\big)\right\} < \left\{\lambda_{\max}\big(\mathbb{E}\big[\mathbf{z}_{i1}\mathbf{z}_{i1}^\top\big]\big)\right\} < L$$

*and there exists $\kappa_1 < \infty$ independent of $n, T$ such that $\sup_i \|\boldsymbol{\gamma}_i^*\| \leq \kappa_1$.*

**Assumption 3.4.** *Assume $\sup_{i,t}\{\|\mathbf{z}_{it}\|_2\} < \kappa < \infty$ a.s. for a constant $\kappa$ that does not depend on $n, T$.*

Assumption 3.3 places mild restrictions on the design matrix. The boundedness condition in Assumption 3.4 is made for the sake of simplicity; it can be relaxed to designs with bounded moments at the cost of additional technicalities in the proofs.

**Theorem 3.3.** *Assume Assumptions 3.3 and 3.4 hold, that data are i.i.d. across $t$ and independent across $i$, and $T \to \infty, \log n/T \to 0$.*
*(i) It holds that*

$$\sup_{i\in\{1,\dots,n\}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{10}$$

*(ii) Under the same assumptions the estimators $\hat{\Sigma}_{i,j}$ in (9) satisfy*

$$\sup_{i\neq j} \left|\left|\left| T\hat{\Sigma}_{i,j} - \Sigma_{i,j} \right|\right|\right|_2 = o_{\mathbb{P}}(1)\,, \tag{11}$$

*where $\Sigma_{i,j}$ denotes the lower $p \times p$ submatrix of $\widetilde{\Sigma}_i + \widetilde{\Sigma}_j$. Furthermore $\Sigma_{i,j}$ satisfy (7).*

Theorem 3.3 implies that Assumptions 3.1 and 3.2 hold with $a_{n,T} = \sqrt{T^{-1}\log n}$, $b_T = T$ and $\Sigma_{i,j}$ corresponding to the scaled asymptotic variance matrix of $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$. In particular, (8) is satisfied provided that $\Delta_{min} \gg (\log n)/\sqrt{T}$.

Note that the results directly imply that Assumptions 3.1 and 3.2 continue to hold for any sub-vectors of $\hat{\boldsymbol{\gamma}}_i$. This covers settings where we want to leave some coefficients individual-specific and only perform grouping on a part of the full coefficient vector.

We now proceed to consider the case of temporal dependence.

**Assumption 3.5.** *For each $i \geq 1$, the process $(\mathbf{x}_{it}, Y_{it})_{t\in\mathbb{Z}}$ is strictly stationary and $\beta$-mixing. Let $\beta_i(j)$ denote the $\beta$-mixing coefficient of the process $(\mathbf{x}_{it}, Y_{it})_{t\in\mathbb{Z}}$. Assume that there exist constants $b_\beta \in (0,1), C_\beta > 0$ independent of $i, n, T$ such that*

$$\sup_i \beta_i(j) \leq \beta(j), \quad \forall j \geq 1,$$

*where $\beta(j) := C_\beta b_\beta^j$.*

Such exponential mixing assumptions are often made in the literature, see for instance Kato et al. (2012) and Galvao et al. (2020) in the context of quantile regression.

The available data $(\mathbf{x}_{it}, Y_{it})_{t=1,\dots,T}$ are an observed stretch from the strictly stationary process $(\mathbf{x}_{it}, Y_{it})_{t\in\mathbb{Z}}$. Under this assumption, the asymptotic variance of the estimator $\hat{\boldsymbol{\gamma}}_i$ is of the form

$$\widetilde{\Sigma}_i = B_i^{-1} H_i B_i^{-1}$$

with

$$\begin{aligned} B_i &:= \mathbb{E}\left[\frac{e^{\mathbf{z}_{i1}^\top \boldsymbol{\gamma}_i^*}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*})^2} \mathbf{z}_{i1}\mathbf{z}_{i1}^\top\right] \\ H_i &:= \mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i1}^\top\big] + \sum_{j=1}^{\infty} \mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i,1+j}^\top + \mathbf{w}_{i,1+j}\mathbf{w}_{i1}^\top\big]\,, \end{aligned}$$

where $\mathbf{w}_{it} := Y_{it}\mathbf{z}_{it} - \frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*}\mathbf{z}_{it}}{1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*}}$. A possible sandwich estimator of the asymptotic variance $\widetilde{\Sigma}_i$ takes the form

$$\hat{\widetilde{\Sigma}}_i = \widehat{B}_{iT}^{-1} \widehat{H}_{iT} \widehat{B}_{iT}^{-1}\,,$$

where

$$\begin{aligned} \widehat{B}_{iT} &= \frac{1}{T}\sum_{t=1}^{T} \frac{e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i}}{(1+e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i})^2} \mathbf{z}_{it}\mathbf{z}_{it}^\top \\ \widehat{H}_{iT} &= \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{it}^\top + \sum_{1\le j\le m_T} \Big(1-\frac{j}{T}\Big)\left(\frac{1}{T}\sum_{t=1}^{T-j}\big(\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{i,t+j}^\top + \widehat{\mathbf{w}}_{i,t+j}\widehat{\mathbf{w}}_{it}^\top\big)\right) \\ \widehat{\mathbf{w}}_{it} &= Y_{it}\mathbf{z}_{it} - \frac{e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i}\mathbf{z}_{it}}{1+e^{\mathbf{z}_{it}^\top \hat{\boldsymbol{\gamma}}_i}}\,, \end{aligned}$$

and $m_T > 0$ denotes the bandwidth parameter tending to be infinity as $T$ goes to infinity. Denote by $\widehat{\Sigma}_i$ the lower $p\times p$ sub-matrix of $\hat{\widetilde{\Sigma}}_i$. Then we set

$$\hat{\Sigma}_{i,j} := T^{-1}(\widehat{\Sigma}_i + \widehat{\Sigma}_j)\,. \tag{12}$$

**Theorem 3.4.** *Let Assumptions 3.3, 3.4 and 3.5 hold. Assume $T$ grows at most polynomially in $n$ and $(\log n)^3 = o(T)$. Assume that the smallest and largest eigenvalues of $H_i$ are bounded away from zero an infinity uniformly in $n, T$.*
*(i) It holds that*

$$\sup_{i\in\{1,\dots,n\}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{13}$$

*(ii) In addition, if $m_T \to \infty$ as $T\to\infty$ and $\frac{m_T^3 \log(n\vee m_T)}{T} = o(1)$, the estimators $\hat{\Sigma}_{i,j}$ satisfy*

$$\sup_{i\neq j} \Big|\Big|\Big| T\hat{\Sigma}_{i,j} - \Sigma_{i,j} \Big|\Big|\Big|_2 = o_{\mathbb{P}}(1)\,, \tag{14}$$

*where $\Sigma_{i,j}$ denotes the lower $p \times p$ submatrix of $\widetilde{\Sigma}_i + \widetilde{\Sigma}_j$. Furthermore $\Sigma_{i,j}$ satisfy* (7) .

Theorem 3.4 implies that Assumptions 3.1 and 3.2 hold with $a_{n,T} = \sqrt{T^{-1} \log n}$, $b_T = T$ and $\Sigma_{i,j}$ corresponding to the scaled asymptotic variance matrix of $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$. In particular, (8) is satisfied provided that $\Delta_{min} \gg (\log n)/\sqrt{T}$ and we need the additional condition $(\log n)^3 = o(T)$.

### 3.2.2 Quantile regression with individual-specific intercepts and grouping on the slopes (Example 2.2)

Consider the quantile regression panel data model

$$q_{i,\tau}(\mathbf{z}_{it}) = \mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*(\tau), \quad t = 1, \ldots, T, i = 1, \ldots, n \,,$$

where $q_{i,\tau}(\mathbf{z}_{it}) := \inf\{y : \mathbb{P}(Y_{it} < y | \mathbf{z}_{it}) \geq \tau\}$ is the conditional $\tau$-quantile of $Y_{it}$ given $\mathbf{z}_{it}$.

We will first assume that $(\mathbf{z}_{it}, Y_{it})$ are i.i.d. across $t$ for each $i$ and independent across $i$. An extension to temporal dependence as in Assumption 3.5 will be considered below. The distribution of $(\mathbf{z}_{it}, Y_{it})$ and the values of $\boldsymbol{\gamma}_i$ are allowed to vary with $n, T$.

Consider the quantile regression estimator $\hat{\boldsymbol{\gamma}}_i^\top = (\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i^\top)$ :

$$\hat{\boldsymbol{\gamma}}_i := \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \rho_\tau(Y_{it} - \mathbf{z}_{it}^\top \boldsymbol{\gamma}) \,,$$

where $\rho_\tau(u) := \{\tau - \mathbb{1}(u \leq 0)\}u$ denotes the check function.

Under mild regularity assumptions (in particular, this is true under Assumptions 3.6–3.8 given below) this estimator is asymptotically normal with asymptotic covariance matrix of the form $\widetilde{\Sigma}_i = B_i^{-1} H_i B_i^{-1}$ where

$$H_i := \tau(1-\tau)\mathbb{E}[\mathbf{z}_{i1}\mathbf{z}_{i1}^\top], \quad B_i = \mathbb{E}[f_{Y_{i1}|\mathbf{z}_{i1}}(q_{i,\tau}(\mathbf{z}_{i1}) \mid \mathbf{z}_{i1})\mathbf{z}_{i1}\mathbf{z}_{i1}^\top] \,, \tag{15}$$

with $f_{Y_{i1}|\mathbf{z}_{i1}}(y|\mathbf{z})$ as the density function of the conditional distribution $F_{Y_{i1}|\mathbf{z}_{i1}}(y|\mathbf{z})$.

A common way to estimate $\tilde{\Sigma}_i$ uses the Hendricks-Koenker sandwich covariance matrix estimator (Hendricks and Koenker (1992)) which takes the following form

$$\hat{\tilde{\Sigma}}_{iT} := \widehat{B}_{iT}^{-1} \widehat{H}_{iT} \widehat{B}_{iT}^{-1} \,, \quad \text{with} \tag{16}$$

$$\widehat{B}_{iT} := \frac{1}{T} \sum_{t=1}^{T} \widehat{f}_{it} \mathbf{z}_{it} \mathbf{z}_{it}^\top \,, \; \widehat{H}_{iT} := \tau(1-\tau) \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_{it} \mathbf{z}_{it}^\top \,, \; \widehat{f}_{it} := \frac{2d_T}{\mathbf{z}_{it}^\top (\hat{\boldsymbol{\gamma}}_i(\tau + d_T) - \hat{\boldsymbol{\gamma}}_i(\tau - d_T))} \,.$$

Here $d_T$ denotes a smoothing parameter that should converge to zero at an appropriate

rate. Let $\hat{\Sigma}_{iT}$ denote the lower $p \times p$ submatrix of $\hat{\widetilde{\Sigma}}_{iT}$ and set

$$\hat{\Sigma}_{i,j} := T^{-1}\Big(\hat{\Sigma}_{iT} + \hat{\Sigma}_{jT}\Big). \tag{17}$$

We now verify Assumptions 3.1 and 3.2, under the following conditions.

**Assumption 3.6.** *Assume that* $\|\mathbf{z}_{it}\|_2 \leq \kappa < \infty$, *and that* $c_\lambda \leq \lambda_{\min}(\mathbb{E}[\mathbf{z}_{it}\mathbf{z}_{it}^\top]) \leq \lambda_{\max}(\mathbb{E}[\mathbf{z}_{it}\mathbf{z}_{it}^\top]) \leq C_\lambda$ *holds uniformly in* $i$ *for some fixed constants* $c_\lambda > 0$ *and* $\kappa, C_\lambda < \infty$ *that are independent of* $n, T$.

**Assumption 3.7.** *Define* $\mathcal{Z} := [-\kappa, \kappa]^{p+1}$. *The conditional distribution* $F_{Y_{i1}|\mathbf{z}_{i1}}(y|\mathbf{z})$ *is twice differentiable w.r.t.* $y$, *with the corresponding derivatives* $f_{Y_{i1}|\mathbf{z}_{i1}}(y|\mathbf{z})$ *and* $f'_{Y_{i1}|\mathbf{z}_{i1}}(y|\mathbf{z})$. *Assume that*

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

*where* $f_max, \overline{f'}$ *are independent of* $n, T$.

**Assumption 3.8.** *Denote by* $\mathcal{T}$ *an open neighborhood of* $\tau$. *Assume that uniformly across* $i$, *there exists a constant* $f_{\min} < f_{\max}$ *independent of* $n, T$ *such that*

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

**Assumption 3.9.** *Assume that* $d_T = o(1)$, *as* $T \to \infty$ *and*

$$\frac{\log(nT)}{T d_T^{4/3}} = o(1).$$

Assumptions 3.6-3.9 are fairly standard in the quantile regression literature and have been imposed in Kato et al. (2012) and Galvao et al. (2020) among many others.

**Theorem 3.5.** *Let Assumptions 3.6-3.8 hold. Assume* $\log n = o(T)$ *and* $\min(n, T) \to \infty$. *Assume that the data are i.i.d. across* $t$ *and independent across* $i$.
*(i) It holds that*

$$\sup_{i\in\{1,\ldots,n\}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

*In particular, Assumption 3.1 holds with* $a_{n,T} = \sqrt{\frac{\log n}{T}}$ *provided that* $\log n = o(T)$.
*(ii) If in addition to the above Assumption 3.9 holds, then Assumption 3.2 is also satisfied with* $b_T := T$, $\Sigma_{i,j}$ *denoting the lower* $p \times p$ *sub-matrix of* $\widetilde{\Sigma}_i + \widetilde{\Sigma}_j$, *and* $\hat{\Sigma}_{i,j}$ *defined in* (20).

Theorem 3.3 implies that Assumptions 3.1 and 3.2 hold with $a_{n,T} = \sqrt{T^{-1}\log n}$, $b_T = T$ and $\Sigma_{i,j}$ corresponding to the scaled asymptotic variance matrix of $\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j$.

Similarly to the discussion in Section 3.2.1, the results directly imply that Assumptions 3.1 and 3.2 continue to hold for any sub-vectors of $\hat{\boldsymbol{\gamma}}_i$.

We now consider the dependent case. Since we need to account for the temporal dependence structure, the asymptotic covariance matrix for the estimator $\hat{\boldsymbol{\gamma}}_i$ is now of the form $\widetilde{\Sigma}_i = B_i^{-1}\widetilde{H}_i B_i^{-1}$ where the matrix $B_i$ is defined in (15) as in the independent case, whereas the matrix $\widetilde{H}_i$ is defined in the following way incorporating the dependence

$$\widetilde{H}_i := \tau(1-\tau)\mathbb{E}[\mathbf{z}_{i1}\mathbf{z}_{i1}^\top] + \sum_{j=1}^{\infty}\mathbb{E}[\mathbf{w}_{i1}\mathbf{w}_{i,1+j}^\top + \mathbf{w}_{i,1+j}\mathbf{w}_{i1}^\top]\,,$$

with $\mathbf{w}_{i1} = \mathbf{z}_{i1}\big(\tau - \mathbb{1}(Y_{it} \le q_{i,\tau}(\mathbf{z}_{i1}))\big)$. This motivates the following generalized version of the Hendricks-Koenker sandwich covariance matrix estimator $\hat{\widetilde{\Sigma}}_{iT}$

$$\hat{\widetilde{\Sigma}}_{iT} := \widehat{B}_{iT}^{-1}\widehat{H}'_{iT}\widehat{B}_{iT}^{-1}\,, \tag{18}$$

where $\widehat{B}_{iT} := \frac{1}{T}\sum_{t=1}^{T}\widehat{f}_{it}\mathbf{z}_{it}\mathbf{z}_{it}^\top$ is defined in the same way as in the independent case and the estimator $\widehat{H}'_{iT}$ is defined via

$$\widehat{H}'_{iT} := \tau(1-\tau)\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^\top + \sum_{1\le j\le m_T}\Big(1-\frac{j}{T}\Big)\Big(\frac{1}{T}\sum_{t\in T_j}(\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{i,t+j}^\top + \widehat{\mathbf{w}}_{i,t+j}\widehat{\mathbf{w}}_{it}^\top)\Big)$$

Here, $T_j := \{1 \le t \le T-j\}$, $m_T > 0$ denotes the bandwidth parameter tending to be infinity as $T$ goes to infinity, and

$$\widehat{\mathbf{w}}_{it} := \mathbf{z}_{it}\Big(\tau - \mathbb{1}\big(Y_{it} \le \hat{\boldsymbol{\gamma}}_i(\tau)^\top\mathbf{z}_{it}\big)\Big)\,.$$

To establish the asymptotic consistency of the covariance estimator, we need following additional assumptions.

**Assumption 3.10.** *For each $i = 1, ..., n$ and $j > 1$, the random vector $(Y_{i1}, Y_{i,1+j})$ has a density conditional on $(\mathbf{z}_{i1}, \mathbf{z}_{i,1+j})$ and this density is bounded uniformly across $i, j$ and $n, T$.*

A similar assumption was made in Kato et al. (2012).

**Assumption 3.11.** *Assume that $d_T = o(1)$ and $m_T \to \infty$ as $T \to \infty$, and*

$$\frac{\log n}{T d_T^2} = o(1), \qquad \frac{m_T^3 \log n}{T} = o(1)\,.$$

This assumption is similar to Assumption 3.9 and imposes a restriction on the relative growth of the time dimension compared to the number of individuals.

**Theorem 3.6.** *Let Assumptions 3.6-3.8, and 3.5-3.10 hold. Assume $T$ grow at most polynomial in $n$, $(\log n)^3 = o(T)$, and $\min(n,T) \to \infty$. Assume that the smallest and largest eigenvalues of $\widetilde{H}_i$ are bounded away from zero and infinity uniformly in $i, n, T$.*
*(i) It holds that*

$$\sup_{i\in\{1,\dots,n\}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{T}}\Big). \tag{19}$$

*In particular, Assumption 3.1 holds with $a_{n,T} = \sqrt{\frac{\log n}{T}}$.*
*(ii) If in addition to the above Assumption 3.11 holds, then Assumption 3.2 is also satisfied with $b_T := T$, $\Sigma_{i,j}$ denoting the lower $p \times p$ sub-matrix of $\widetilde{\Sigma}_i + \widetilde{\Sigma}_j$ and $\hat{\Sigma}_{iT}$ denoting the lower $p \times p$ submatrix of*

$$\hat{\Sigma}_{i,j} := T^{-1}\Big(\hat{\widetilde{\Sigma}}_{iT} + \hat{\widetilde{\Sigma}}_{jT}\Big). \tag{20}$$

### 3.2.3 Quantile regression with common slope and grouping on the intercepts (Example 2.3)

Consider the quantile regression panel data model

$$q_{i,\tau}(\mathbf{x}_{it}) = \alpha_i^*(\tau) + \mathbf{x}_{it}^\top \boldsymbol{\beta}^*(\tau), \quad t = 1,\dots,T, i = 1,\dots,n\,,$$

where $q_{i,\tau}(\mathbf{x}_{it}) := \inf\big\{y : \mathbb{P}(Y_{it} < y|\mathbf{x}_{it}) \geq \tau\big\}$ denotes the conditional $\tau$-quantile of $Y_{it}$ given $\mathbf{x}_{it}$. In contrast to the setting in Section 3.2.2, we assume that the slope coefficient $\boldsymbol{\beta}^*$ is common across individuals and are only interested in grouping the intercepts. This model was considered in Gu and Volgushev (2019), who used a lasso-type penalty to enforce grouping on the intercepts $\alpha_i^*$. The latter paper also demonstrated that putting this kind of regularization on $\alpha_i^*$ can result in improved estimation of $\boldsymbol{\beta}^*$ compared to leaving $\alpha_i^*$ unrestricted.

Assume that $(\mathbf{x}_{it}, Y_{it})$ are i.i.d. across $t$ for each $i$ and independent across $i$. Since only intercepts contain the grouping information, we aim to use the estimates for $\alpha_i^*$, and their variance estimates to construct the similarity matrix. At this point, there are two possibilities for estimating $\alpha_i^*$: (1) run individual-specific quantile regressions ignoring the fact that $\boldsymbol{\beta}^*$ is common across individuals or (2) put all individuals into a single large model in order to borrow information across individuals to improve the efficiency in estimating the joint coefficient vector $\boldsymbol{\beta}^*$.

Approach (1) has computational advantages, especially if $n$ is large, but can also result in a loss of efficiency. The theoretical treatment of (1) easily follows from minor modifications of the results in Section 3.2.2, and we hence focus on the second approach. Define

$$\big(\tilde{\alpha}_1(\tau), \cdots, \tilde{\alpha}_n(\tau), \tilde{\boldsymbol{\beta}}(\tau)\big) := \underset{\alpha_1,\dots,\alpha_n,\boldsymbol{\beta}}{\arg\min} \frac{1}{nT} \sum_{i=1}^{n}\sum_{t=1}^{T} \rho_\tau(Y_{it} - \alpha_i - \mathbf{x}_{it}^\top \boldsymbol{\beta})\,. \tag{21}$$

In what follows, we assume that $n \to \infty$ which is the more relevant case for group structure detection. In this case, it is possible to obtain simplified estimators for the asymptotic variance of $\tilde{\alpha}_i$. Those estimators will be motivated next.

The main insight is that under $n \to \infty$ the estimation of $\boldsymbol{\beta}^*$ has a negligible effect of the asymptotic variance of $\tilde{\alpha}_i$ since $\boldsymbol{\beta}^*$ is estimated at a faster rate due to borrowing information across individuals. Further observe that, defining $\hat{e}_{it} = Y_{it} - \mathbf{x}_{it}^\top \tilde{\boldsymbol{\beta}}$, we have

$$\tilde{\alpha}_i = \arg\min_{\alpha \in \mathbb{R}} \frac{1}{T} \sum_{t=1}^{T} \rho_\tau(\hat{e}_{it} - \alpha), \quad i = 1, \dots, n\,. \tag{22}$$

Thus $\tilde{\alpha}_i$ is approximately the sample quantile of $\{\hat{e}_{it}, t = 1, \dots, T\}$, which should be close to the sample quantile of $\{e_{it}, t = 1, \dots, T\}$, where $e_{it} := Y_{it} - \mathbf{x}_{it}^\top \boldsymbol{\beta}^*$.

If $n \to \infty$ this idea can be formalized by applying a modification of Lemma 7 in Galvao et al. (2020) (after noting that the proof of the latter result can be modified to weaken the assumption $n(\log T)^2/T \to 0$ made in there). Denoting the sample quantile of $\{e_{it}, t = 1, \dots, T\}$ by $\hat{\alpha}_i$, the latter result implies

$$\sup_{i=1,\dots,n} |\hat{\alpha}_i - \tilde{\alpha}_i| = \mathcal{O}_{\mathbb{P}}\Big( \Big\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\Big\|_2 + T^{-1} \log(n \vee T)\Big).$$

Observing that by the proof of Theorem 3.2 in Kato et al. (2012) we have $\Big\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\Big\|_2 = o_{\mathbb{P}}(T^{-1/2})$, when $n \to \infty$ (note that this part of their result does not require the restrictive growth assumption on $n$ which is needed for unbiased asymptotic normality of $\tilde{\boldsymbol{\beta}}$), this implies $|\hat{\alpha}_i - \tilde{\alpha}_i| = o_{\mathbb{P}}(T^{-1/2})$ uniformly over $i$ and thus the asymptotic distributions of $\hat{\alpha}_i$ and $\tilde{\alpha}_i$ coincide. Now classical results on the distribution of sample quantiles imply that the asymptotic variance of $\hat{\alpha}_i$ is

$$\Sigma_i = \tau(1-\tau)/f_{e_i}^2(q_{e_i}(\tau))\,, \tag{23}$$

where $f_{e_i}, q_{e_i}$ denote the (unconditional) density and quantile function of $e_{i1}$, respectively.

This motivates the following version of $\hat{\Sigma}_{i,j}$: for a bandwidth parameter $d_T$ define

$$\hat{\tilde{\Sigma}}_{iT} := \tau(1-\tau)\left(\frac{\tilde{\alpha}_i(\tau + d_T) - \tilde{\alpha}_i(\tau - d_T)}{2d_T}\right)^2, \quad i = 1, \dots, n$$

and compute

$$\hat{\Sigma}_{i,j} := T^{-1}\Big(\hat{\tilde{\Sigma}}_{iT} + \hat{\tilde{\Sigma}}_{jT}\Big)\,. \tag{24}$$

**Theorem 3.7.** *Let Assumptions 3.6-3.8 with* $\mathbf{z}_{it} = \mathbf{x}_{it}$ *hold. Assume* $\log n = o(T)$, $\min(n, T) \to \infty$.

*(i) It holds that*

$$\sup_{i\in\{1,\dots,n\}} |\tilde{\alpha}_i - \alpha_i^*| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log(n\vee T)}{T}}\right).$$

*In particular, Assumption 3.1 holds with* $a_{n,T} = \sqrt{\frac{\log(n\vee T)}{T}}$.
*(ii) If in addition to the above* $\log(n\vee T)/(nd_T^2) = o(1)$*, then Assumption 3.2 is also satisfied with* $b_T := T$, $\Sigma_{i,j} = \Sigma_i + \Sigma_j$ *where* $\Sigma_i, \Sigma_j$ *are defined in* (23)*, and* $\hat{\Sigma}_{i,j}$ *defined in* (24)*.*

We next consider the case of temporal dependence. Under Assumptions 3.5–3.10 the asymptotic variance takes he form

$$\Sigma_i = \frac{1}{f_{e_i}^2(q_{e_i}(\tau))} \sum_{t\in\mathbb{Z}} \mathrm{Cov}\left(\mathbb{1}\{e_{i0} \le q_{e_i}(\tau)\}, \mathbb{1}\{e_{it} \le q_{e_i}(\tau)\}\right).$$

This can be estimated consistently by

$$\hat{\Sigma}_{iT} := \left(\frac{\tilde{\alpha}_i(\tau+d_T) - \tilde{\alpha}_i(\tau-d_T)}{2d_T}\right)^2 \Big(\tau(1-\tau) + \sum_{1\le j\le m_T} (1-j/T) \sum_{t\in T_j} (\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{i,t+j}^\top + \widehat{\mathbf{w}}_{i,t+j}\widehat{\mathbf{w}}_{it}^\top)\Big)$$

where $T_j := \{1 \le t \le T-j\}$, $m_T > 0$ denotes the bandwidth parameter tending to be infinity as $T$ goes to infinity, and

$$\widetilde{\mathbf{w}}_{it} := \tau - \mathbb{1}\big\{Y_{it} \le \tilde{\boldsymbol{\beta}}_i(\tau)^\top \mathbf{x}_{it} + \hat{\alpha}_i(\tau)\big\}.$$

**Theorem 3.8.** *Let Assumptions 3.6-3.8, and 3.5-3.10 with* $\mathbf{z}_{it} = \mathbf{x}_{it}$ *hold. Assume* $T$ *grows at most polynomially in* $n$*,* $(\log n)^3 = o(T)$*, and* $\min(n,T) \to \infty$*.*
*(i) It holds that*

$$\sup_{i\in\{1,\dots,n\}} |\tilde{\alpha}_i - \alpha_i^*| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

*where* $\tilde{\alpha}_i$ *is defied in* (22)*. In particular, Assumption 3.1 holds with* $a_{n,T} = \sqrt{\frac{\log n}{T}}$.
*(ii) If in addition to the above Assumption 3.11 holds, then Assumption 3.2 is also satisfied with* $b_T := T$, $\Sigma_{i,j} = \Sigma_i + \Sigma_j$*, where* $\Sigma_i, \Sigma_j$ *are defined in* (23)*, and* $\hat{\Sigma}_{i,j}$ *defined in* (24)*.*

# 4 Numerical experiments

In Section 4.1 and Section 4.2, we report the performance of different algorithms in terms of assigning individuals to groups when the true number of groups is specified. We consider two performance metrics: perfect matching, which corresponds to the proportion of times that the exact group assignment is found, and average matching. The latter is computed as follows. Define the true cluster assignment as a set $\omega^* := \{\omega_1^*, \dots, \omega_n^*\}$, where $\omega_i^* \in \{1, \dots, G^*\}$ denotes the $\omega_i^*$-th group to which the individual $i$ belongs. Define the set of permutations of the labels $\Phi := \{\phi : \phi \text{ is a bijection from } \{1, \dots, G^*\} \text{ to } \{1, \dots, G^*\}\}$.

Define the estimated membership as a set $\hat{\omega} := \{\hat{\omega}_1, \ldots, \hat{\omega}_n\}$, where $\hat{\omega}_i \in \{1, \ldots, G^*\}$ denotes the estimated group number of the $i$-th individual. We define the average percentage of correct classification of the estimated membership $\hat{\omega}$ as

$$\max_{\phi \in \Phi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\phi(\omega_i^*) = \hat{\omega}_i\}\,.$$

A similar approach was taken in Su et al. (2016); Gu and Volgushev (2019); Leng et al. (2023). The performance of the heuristic (5) for selecting the number of groups is considered in Section 4.1.2 for logistic regression and in Section 4.3 for quantile regression. Additional models and simulation settings are discussed in the supplement.

### 4.1 Logistic regression

In this section, we consider logistic regression with individual-specific intercepts and groupings on the slopes specified as

$$Y_{it} = \mathbb{1}\{\alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta}_{g_i} \geq \epsilon_{it}\}\,,$$

where $\epsilon_{it}$ follows a logistic distribution, $\alpha_i = 1$ for all $i$ and $g_i \in \{1, 2, 3\}$ with equal proportions, and $\mathbf{x}_{it}^\top := (x_{1it}, x_{2it})$. Moreover,

$$\boldsymbol{\beta}_1 = \begin{pmatrix} -4 \\ 1 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \boldsymbol{\beta}_3 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$

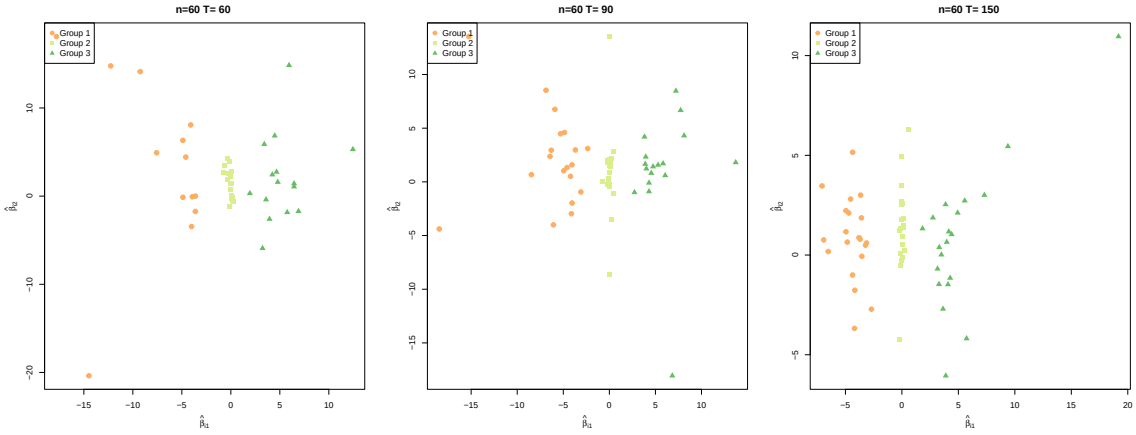
We consider two different data generating processes for the covariates $\mathbf{x}_{it}$. For Model 1,

$$x_{1it} = 0.5\alpha_i + \eta_i + z_{1it}, \quad \text{and} \quad x_{2it} = 0.5\alpha_i + \eta_i + z_{2it}\,,$$

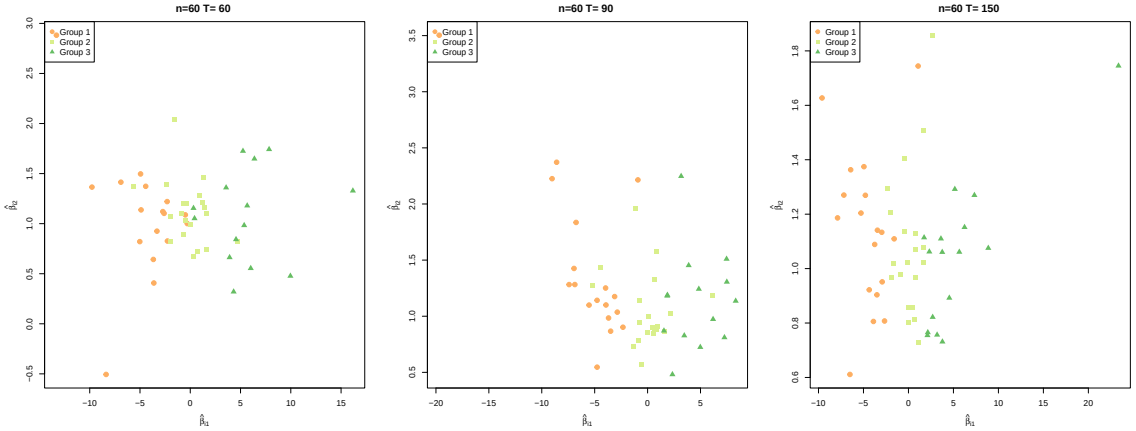
where $\eta_i \sim N(0, 1)$ and $z_{1it} \sim N(0, 4)$ and $z_{2it} \sim N(0, 0.04)$.

Here, the data generating process is constructed such that the coefficient of $x_2$ is not informative on the group structure while at the same time it is estimated less precisely. On the contrary, the coefficient of $x_1$ is informative on group structure and also precisely estimated.

For model 2, we switch the labels of $x_1$ and $x_2$. This is a more challenging DGP because the coordinate of $\boldsymbol{\beta}$ that contains group information is estimated with a lot of noise; see the scatter plot of $\{\hat{\boldsymbol{\beta}}_i\}_{i=1,\ldots,n}$ in Figure 1 for a data realization from Model 1 versus Model 2.




(a)




(b)

Figure 1: Scatter plots of $\{\hat{\boldsymbol{\beta}}_i\}_{i=1}^n$ for Model 1 (Figure (a)) and Model 2 (Figure (b)).

#### 4.1.1 Clustering with a known number of groups

We first compare our method with the C-LASSO proposed by Su et al. (2016). The C-LASSO approach proposed in Su et al. (2016) considers minimizing the following objective:

$$\frac{1}{nT}\sum_i\sum_t \psi(Y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}_i, \hat{\alpha}_i(\boldsymbol{\beta}_i)) + \frac{\lambda}{n}\sum_i \prod_{k=1}^{K_0} \|\boldsymbol{\beta}_i - \eta_k\|\,.$$

This itself is not a convex optimization, but at each $k$, we can focus on only the $k$-th element in the product term in the penalty, resulting in a convex program. For details of the implementation, we refer to our supplement or Su et al. (2016). In addition, we also consider an interative $k$-means approach in the spirit of Bonhomme and Manresa (2015). In particular, in each iterative step, we re-estimate group membership based on the logit likelihood function, and then refit the model until coefficients converge. We also compare to

the Sequential Binary Segmentation Algorithm (SBSA) in Wang and Su (2021) (labeled as SBSA in Table 1. The SBSA method applies the binary segmentation algorithm to detect break-points in eigenvectors from the spectral decomposition of the outer product of $\hat{\boldsymbol{\beta}}$ that corresponds to the $\min(p, G)$ largest eigenvalues where $p$ is the number of covariates.

**Model 1**

| | | Perfect Match | | | | | | | | | | Average Match | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | T | S | PAM | S-Diag | S-Iden | | C-LASSO | | | $k$-means | SBSA | S | PAM | S-Diag | S-Iden | | C-LASSO | | | $k$-means | SBSA |
| 30 | 60 | 0.83 | 0.53 | 0.88 | 0.00 | 0.32 | 0.67 | 0.78 | 0.40 | 0.03 | 0.07 | 0.99 | 0.92 | 0.99 | 0.48 | 0.95 | 0.98 | 0.99 | 0.96 | 0.86 | 0.63 |
| 30 | 90 | 0.93 | 0.70 | 0.99 | 0.00 | 0.44 | 0.80 | 0.86 | 0.55 | 0.05 | 0.16 | 1.00 | 0.96 | 1.00 | 0.45 | 0.97 | 0.99 | 0.99 | 0.98 | 0.88 | 0.74 |
| 30 | 150 | 0.99 | 0.98 | 0.99 | 0.09 | 0.77 | 0.93 | 0.98 | 0.93 | 0.04 | 0.70 | 1.00 | 1.00 | 1.00 | 0.59 | 0.99 | 1.00 | 1.00 | 1.00 | 0.88 | 0.92 |
| 60 | 60 | 0.77 | 0.40 | 0.82 | 0.01 | 0.14 | 0.47 | 0.55 | 0.21 | 0.00 | 0.00 | 1.00 | 0.91 | 1.00 | 0.43 | 0.95 | 0.98 | 0.98 | 0.97 | 0.86 | 0.63 |
| 60 | 90 | 0.80 | 0.57 | 0.88 | 0.01 | 0.21 | 0.57 | 0.70 | 0.54 | 0.00 | 0.16 | 1.00 | 0.96 | 1.00 | 0.41 | 0.96 | 0.99 | 0.99 | 0.99 | 0.87 | 0.73 |
| 60 | 150 | 0.92 | 0.88 | 0.96 | 0.12 | 0.62 | 0.88 | 0.91 | 0.85 | 0.00 | 0.66 | 1.00 | 1.00 | 1.00 | 0.58 | 0.99 | 1.00 | 1.00 | 1.00 | 0.88 | 0.91 |

**Model 2**

| | | Perfect Match | | | | | | | | | | Average Match | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | T | S | PAM | S-Diag | S-Iden | | C-LASSO | | | $k$-means | SBSA | S | PAM | S-Diag | S-Iden | | C-LASSO | | | $k$-means | SBSA |
| 30 | 60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.69 | 0.64 | 0.68 | 0.54 | 0.58 | 0.73 | 0.68 | 0.58 | 0.70 | 0.49 |
| 30 | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.75 | 0.68 | 0.74 | 0.54 | 0.63 | 0.79 | 0.75 | 0.63 | 0.72 | 0.52 |
| 30 | 150 | 0.01 | 0.01 | 0.02 | 0 | 0 | 0.02 | 0 | 0 | 0 | 0 | 0.87 | 0.76 | 0.86 | 0.58 | 0.73 | 0.86 | 0.84 | 0.73 | 0.74 | 0.56 |
| 60 | 60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.70 | 0.63 | 0.68 | 0.48 | 0.58 | 0.73 | 0.66 | 0.58 | 0.68 | 0.47 |
| 60 | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.79 | 0.67 | 0.77 | 0.49 | 0.65 | 0.80 | 0.75 | 0.65 | 0.72 | 0.50 |
| 60 | 150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.88 | 0.73 | 0.86 | 0.50 | 0.75 | 0.87 | 0.84 | 0.75 | 0.71 | 0.55 |

Table 1: Comparison of group membership estimation. $S$ refers to our proposed spectral clustering method, PAM refers to the PAM method applied on the dissimilarity measure $V$. S-Diag refers to the spectral clustering approach but we plug in the diagnonal of the variance-covariance matrix estimate and likewise, S-Iden is similar but the variance-covariance matrix of individual estimate is taken to be the identity matrix. For C-LASSO, the four columns of the results are based on tuning parameter constants $c = 0.05 \times \{1, \frac{1}{4}, \frac{1}{8}, \frac{1}{32}\}$. The k-means approach is adapted from Bonhomme and Manresa (2015) where we iteratively cluster individuals with a refit to update group coefficients until convergence. For the iterative k-means method, we use 20 random starting groupings and a maximum of 100 iterations for each starting grouping. Then we take the grouping that minimizes the loss function.

The first few rows in Table 1 report the performance of four different grouping methods for several combinations of $n$ and $T$ based on Model 1. We evaluate the performance by the proportion of perfect matches out of 100 simulation repetitions and the average matches described at the beginning of Section 4. The spectral clustering method works consistently better than the PAM approach (labeled PAM).

From local analysis in Section 2.2, one might expect that PAM and the iterative k-mean method should perform similarly. However, that analysis is asymptotic and inspecting the scatter plot of $\{\hat{\boldsymbol{\beta}}_i\}_{i=1,\ldots,n}$ in Figure 1 suggests that the coefficient estimates are not yet approximately Gaussian around their true values. Hence the asymptotic analysis may not provide a sufficiently accurate description of finite sample performance at the sample sizes considered in this simulation. We also note that there are non-convergence issues with the iterative k-means method. For $n = 30, L = 60$, in 12% of the cases none of the random initialization lead to convergence after 100 iterations and in 35% of all cases the initialization that led to the best likelihood function did not correspond to convergence after 100 iterations.

In addition, the non-Gaussian shape of the point clouds may also provide an explanation for the superior performance of the spectral clustering method over PAM, since spectral

clustering is known to have an advantage for clusters with non-elliptical shapes.

For small $T$, using the diagonalized estimated variance-covariance matrix (labeled S-Diag) actually performs slightly better than using the full estimated variance-covariance matrix (labeled S). In finite samples, the off-diagonal terms of the variance-covariance matrix can be poorly estimated and using just the diagonal variance information seems to provide a small margin of better performance. However, discarding variance information completely (labeled S-Iden) clearly shows much worse performance. The iterative $k$-mean method performs much worse than spectral clustering with variance information in terms of perfect match. Its average match performance is in fact better than the spectral method when the variance information is not accounted for. This shows that spectral clustering needs to be applied together with variance information for good performance. We also note that the iteration k-means method sometimes does not converge after 100 iterations, which may explain why performance is not improving monotonically as $T$ increases. The SBSA approach in Wang and Su (2021) has better performance than the k-mean method for perfect match proportion in Model 1, but is still inferior to CLASSO, PAM, and spectral clustering with variance information.

For C-LASSO, the penalty tuning parameter $\lambda$ is set at $cT^{-1/3}\mathbb{V}\text{ar}\left(Y_{it}\right)$ as recommended by the authors with a few different values of $c$ specified in the caption of Table 1. We see that the C-LASSO can perform very well for a suitably chosen constant, and our method matches that or overperforms sometimes. However, it can perform poorly if the tuning parameter constant is not chosen carefully. This imposes challenges for its practical usage.

Performance for Model 2 is reported in the last few rows in Table 1. We clearly see that this is a much more challenging DGP with almost all methods failing to recover perfect match for group membership. In terms of average matches, our method still performs comparable or sometimes better than all other methods for all combinations of $n$ and $T$.

### 4.1.2 Estimating the number of groups

Simulation results in Table 1 assume the researchers know the correct number of groups $G$. We report in Table 4.1.2 the performance of the proposed method for estimation of $G$. The comparison is made with the information criteria proposed in Su et al. (2016). For Model 1, both methods work very well while for Model 2, the information criteria of Su et al. (2016) works much better across most combinations of $n, T$. The information criteria relies on the whole sample to estimate $G$ while our heuristic approach only requires information on individual based estimates.

### 4.1.3 Computation times

In what follows, Table 3 reports the computation times for our method versus C-LASSO, the iterative k-means method which requires iteration with the whole sample as well as SBSA proposed in Wang and Su (2021). The run time of SBSA is very similar to our

**Model 1**

| | | Heuristic | | | | | IC-CLASSO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | T | 1 | 2 | 3 | 4 | ≥5 | 1 | 2 | 3 | 4 | ≥5 |
| 30 | 60 | 0.10 | 0 | **0.89** | 0 | 0.01 | 0 | 0.02 | **0.98** | 0 | 0 |
| 30 | 90 | 0 | 0 | **0.98** | 0.01 | 0.01 | 0 | 0 | **0.99** | 0.01 | 0 |
| 30 | 150 | 0 | 0 | **1.00** | 0 | 0 | 0 | 0 | **0.94** | 0.06 | 0 |
| 60 | 60 | 0.02 | 0 | **0.98** | 0 | 0 | 0 | 0 | **1.00** | 0 | 0 |
| 60 | 90 | 0 | 0 | **1.00** | 0 | 0 | 0 | 0 | **1.00** | 0 | 0 |
| 60 | 150 | 0 | 0 | **1.00** | 0 | 0 | 0 | 0 | **0.94** | 0.06 | 0 |

**Model 2**

| | | Heuristic | | | | | IC-CLASSO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | T | 1 | 2 | 3 | 4 | ≥5 | 1 | 2 | 3 | 4 | ≥5 |
| 30 | 60 | 0.93 | 0.07 | **0.00** | 0 | 0 | 0 | 0.85 | **0.15** | 0 | 0 |
| 30 | 90 | 0.76 | 0.20 | **0.04** | 0 | 0 | 0 | 0.79 | **0.20** | 0.01 | 0 |
| 30 | 150 | 0.61 | 0.16 | **0.22** | 0.01 | 0 | 0 | 0.52 | **0.40** | 0.07 | 0.01 |
| 60 | 60 | 0.92 | 0.08 | **0.00** | 0 | 0 | 0 | 0.99 | **0.01** | 0 | 0 |
| 60 | 90 | 0.87 | 0.11 | **0.02** | 0 | 0 | 0 | 0.92 | **0.08** | 0 | 0 |
| 60 | 150 | 0.47 | 0.16 | **0.36** | 0.01 | 0 | 0 | 0.62 | **0.36** | 0.02 | 0 |

Table 2: Estimation of G for Model 1 and Model 2 with true $G = 3$. IC-CLASSO is based on a combination of the log-likelihood evaluation and a penalty term that depends on a turning parameter, group size and $n$ and $T$.

proposed method because it also only required the use of individual coefficients. For C-LASSO the reported times are based on maximum 20 iterations for optimization. The final estimates are obtained when the objective function differs less than 0.001 and when the $\ell_2$ norm of the estimates group centers differ by less than 0.1% or when the maximum iterations are reached. For the iterative k-mean method, we take 20 random start of group membership and pick the best estimates that minimizes the loss function criteria. For each random starting, the maximum number of iteration is 100. With known $G$, our method and the SBSA method has the least computational time while the iterative k-means has the largest. This is because quite often the k-mean algorithm does not converge before the maximum 100 iterations is reached. Our algorithm spends most of its computation time on individual based estimates. For the C-LASSO method, the individual estimates are computed as initial estimates before applying a re-optimization with the penalty terms for group center estimates. Because the optimization problem is only convex for optimizing over one group center while fixing the others, it has to optimize group by group, which increases computation times. The iterative k-means method takes the most computation time, as it requires individual loops to decide group membership until convergence as well as refitting to obtain group center estimates. In practice, we observe that it may take a very large number of iterations to converge. When $G$ is not known, the computation time of our method does

not increase because the heuristic method recycles already computed similarity measures to estimate $G$. The SBSA method uses a IC criteria to estimate $G$. Because grouping is obtained very fast for each candidate model and the IC criteria just needs to evaluate the likelihood of each estimated candidate model, the increase in computation time is also very minimal. Both the C-LASSO and k-means rely on information criteria to estimate $G$ which requires fitting of all candidate models with varying $G$. Hence computation times grow at least linearly with the number of candidate models. The reported times in Table 3 are based on candidate models with $G = \{1, 2, 3, 4, 5\}$.

| | | Known G | | | | Estimate G | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | T | Spectral | C-LASSO | Kmeans | SBSA | Spectral | C-LASSO | Kmeans | SBSA |
| 30 | 60 | 0.32 | 5.88 | 13.23 | 0.38 | 0.32 | 31.75 | 86.27 | 0.44 |
| 30 | 90 | 0.38 | 6.55 | 24.76 | 0.44 | 0.38 | 39.94 | 122.75 | 0.50 |
| 30 | 150 | 0.42 | 6.87 | 52.69 | 0.43 | 0.43 | 48.09 | 245.97 | 0.52 |
| 60 | 60 | 1.26 | 10.15 | 48.27 | 0.77 | 1.28 | 59.48 | 445.17 | 0.94 |
| 60 | 90 | 1.47 | 11.10 | 197.70 | 0.80 | 1.52 | 79.38 | 742.79 | 1.12 |
| 60 | 150 | 1.61 | 12.30 | 225.92 | 0.85 | 1.61 | 123.48 | 801.23 | 1.34 |

Table 3: Comparison of computation time in seconds for Model 1: the left panel includes computation times when we assume $G$ is known. The right panel includes computation times when we have to estimate $G$. For our proposed method, we use the heuristic method to estimate $G$ and for all other methods, we use some form of information criteria to estimate $G$ from the set $\{1, 2, 3, 4, 5\}$. Timings are averages of 5 data realizations.

## 4.2 Quantile regression

In this section, we consider quantile regression with individual-specific intercepts and grouping on the slopes as in Example 2.2, and with joint slope and grouping of intercepts from Example 2.3. We focus on the clustering performance with a given (correctly specified) number of groups, the performance of the proposed heuristic, and several other methods for selecting the number of groups is considered in Section 4.3.

### 4.2.1 Quantile regression individual–specific intercepts and grouping on slopes

Recall the model specification in Example 2.2: $q_{it}(\tau) = \alpha_i(\tau) + \mathbf{x}_{it}^\top \boldsymbol{\beta}_i(\tau) = \mathbf{z}_{it}^\top \boldsymbol{\gamma}_i(\tau)$ . This setting was also considered in Zhang et al. (2019a) and we will compare the performance of the proposed method with theirs. Simulations are done in the **quantreg** package in R. Covariance estimates are computed using the function `summary.rq()` with option `se="nid"` and default bandwidth choice `hs=true`.

We consider three models. Model 1 corresponds to Model 3 from Zhang et al. (2019a).

**Model 1:**

$$y_{it} = \alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta}_{g_i} + 0.5 x_{2it} e_{it} \,,$$

where

$\alpha_i \overset{iid}{\sim} U(0,1), i = 1, \ldots, n,$ $g_i$ are sampled randomly with equal probablities from $\{1, 2, 3\}$.

Set

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix}, \boldsymbol{\beta}_3 = \begin{pmatrix} 0.3 \\ 0.3 \end{pmatrix},$$

$e_{it} \overset{iid}{\sim} N(0,1)$ or $e_{it} \overset{iid}{\sim} t(3)$, and

$$\mathbf{x}_{it}^{\top} = (x_{1it}, x_{2it}), \text{ with } x_{1it} = 0.3\alpha_i + z_{1it}, \text{ with } z_{1it} \overset{iid}{\sim} N(0,1), \text{and } x_{2it} \sim U(0,1).$$

Results for $\tau = 0.5$ are reported in Table 4. We considered the PAM method as well as several variants of spectral clustering. In particular, $S^g$ refers to spectral clustering when we apply the Gaussian kernel instead of the exponential kernel in Algorithm 1. S-Diag is spectral clustering when we use only the diagonal entries in the variance-covariance matrix and sets the off–diagonal entries to zero. S-Iden is spectral clustering when we do not use the variance covariance information of the coefficient estimates. k-mean$^{\circ}$ applies the k-mean clustering algorithm on $\hat{\boldsymbol{\beta}}$ and ZWZ19 is the method in Zhang et al. (2019a) which adapts the iterative method of Bonhomme and Manresa (2015) to the quantile regression case.

Spectral clustering shows uniformly best performance in terms of average and perfect matching across all settings considered. The approach of Zhang et al. (2019a) comes close in terms of average matching and is better than both methods which ignore variance information (S-Iden and k-means$^{\circ}$) but is slightly worse than PAM and the spectral method. This agrees with the theoretical analysis in Section 2.2 which suggests that for a heteroscedastic model as in Model 1, the loss function based approach implicitly takes into account variance information, but is not as efficient as using the dissimilarity measure as in Algorithm 1. Surprisingly, Zhang et al. (2019a) shows much worse performance in terms of perfect matching. A closer look at the results revealed that in this model the method of Zhang et al. (2019a) often assigns one individual to the wrong group, resulting in good average matching but inferior perfect matching performance. Despite our best efforts at varying various parameters of Zhang et al. (2019a) (e.g. criteria for termination and number of random starting points), we were not able to alleviate this issue. Among spectral methods, using the Gaussian kernel to transform the dissimilarity measure does not lead to improvements in terms of performance. Using just the diagonal of the variance-covariance matrix also yields almost identical performance than using the estimated full variance-covariance matrix. We do note that the PAM method is slightly worse than the corresponding spectral clustering method. This seems to be a persistent phenomenon we observe in all the simulations for quantile regression. Moreover, we provide the scatter plot of $\{\hat{\boldsymbol{\beta}}_i\}_{i=1,\ldots,n}$ in the online Appendix (Section 8 Figure 5) for a data realization where the proposed method

achieves perfect matching but all the other methods fail. The figure suggests that there seems to be clear separation along the first coordinate, but not in the second coordinate of $\hat{\beta}$, which is driven by the fact that the second coordinate is estimated with more noise. However, this information is not available to the researcher. Accounting for the variance information improves the performance compared to methods that do not account for this. The loss function based method of Zhang et al. (2019a) implicitly accounts for this to some extent, but less well than reweighting.

The simulation findings suggest that the main improvements in our proposal are due to using variance information, while using spectral clustering instead of PAM only leads to modest additional gains. The results here are also consistent with the local analysis in Theorem 2.1 since this is a model with heteroscedastic errors.

The second model we consider has four groups, with pairs of group centres being close together. Both entries of the coefficient vector carry information about the group structure, but one of them is estimated more precisely than the other one.

**Model 2:**

$$y_{it} = \alpha_i + \mathbf{x}_{it}^\top \boldsymbol{\beta}_{g_i} + 0.5 x_{2it} e_{it}\,,$$

where

$$\alpha_i = 1, i = 1, \ldots, n, \quad g_i \text{ are sampled randomly with equal probabilities from} \{1, 2, 3, 4\}.$$

Set

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix}, \boldsymbol{\beta}_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \boldsymbol{\beta}_4 = \begin{pmatrix} 3.1 \\ 3.1 \end{pmatrix},$$

$e_{it} \overset{iid}{\sim} N(0,1)$ or $t(3)$, and set

$$\mathbf{x}_{it}^\top = (x_{1it}, x_{2it})\,, \text{ with } x_{1it} = 0.3\alpha_i + z_{1it}, \text{ where } z_{1it} \overset{iid}{\sim} N(0,1), \text{ and } x_{2it} \sim U(0,1)\,.$$

Results for $\tau = 0.5$ are reported in Table 5. The results are fairly similar to those of Model 1, the proposed method has the best performance with respect to perfect and average match. The design of this DGP is also used later for estimation of $G$ to demonstrate the drawback of stability based method proposed in Wang (2010). Again, the main performance boost comes from using reweighting and using spectral clustering instead of PAM only leads to small additional accuracy gains.

**Model 3:** The last model we consider has the same specification as Model 1, except that we allow individuals to have varying time period lengths and individuals with shorter panel length are expected to be estimated with larger standard error. This resembles many macroeconomic settings where individual units have varying panel length and hence individual based estimates are of very different quality. In the simulation, the panel lengths are a random draw from $\{30, 60, 90\}$ with equal probabilities. Results are summarized in Table 6.

| n | T | Perfect Match | | | | | | | Average Match | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 |
| $N(0,1), \tau = 0.5$ | | | | | | | | | | | | | | | |
| 30 | 60 | 0.23 | 0.23 | 0.17 | 0.22 | 0 | 0 | 0.09 | 0.95 | 0.95 | 0.93 | 0.95 | 0.64 | 0.63 | 0.90 |
| 30 | 90 | 0.62 | 0.62 | 0.56 | 0.61 | 0 | 0 | 0.39 | 0.98 | 0.98 | 0.98 | 0.98 | 0.70 | 0.70 | 0.97 |
| 30 | 120 | 0.81 | 0.81 | 0.77 | 0.80 | 0 | 0 | 0.67 | 0.99 | 0.99 | 0.99 | 0.99 | 0.75 | 0.75 | 0.98 |
| 60 | 60 | 0.06 | 0.06 | 0.04 | 0.05 | 0 | 0 | 0.01 | 0.95 | 0.95 | 0.94 | 0.95 | 0.63 | 0.61 | 0.92 |
| 60 | 90 | 0.40 | 0.40 | 0.34 | 0.39 | 0 | 0 | 0.16 | 0.98 | 0.98 | 0.98 | 0.98 | 0.70 | 0.69 | 0.97 |
| 60 | 120 | 0.72 | 0.72 | 0.65 | 0.71 | 0 | 0 | 0.45 | 0.99 | 1.00 | 0.99 | 0.99 | 0.76 | 0.76 | 0.99 |
| 90 | 60 | 0.02 | 0.02 | 0.01 | 0.02 | 0 | 0 | 0.00 | 0.96 | 0.96 | 0.94 | 0.95 | 0.63 | 0.61 | 0.92 |
| 90 | 90 | 0.27 | 0.26 | 0.20 | 0.26 | 0 | 0 | 0.07 | 0.99 | 0.99 | 0.98 | 0.98 | 0.70 | 0.69 | 0.97 |
| 90 | 120 | 0.62 | 0.62 | 0.57 | 0.62 | 0 | 0 | 0.34 | 1.00 | 1.00 | 0.99 | 1.00 | 0.77 | 0.76 | 0.99 |
| $t(3), \tau = 0.5$ | | | | | | | | | | | | | | | |
| 30 | 60 | 0.12 | 0.11 | 0.07 | 0.10 | 0 | 0 | 0.04 | 0.92 | 0.92 | 0.89 | 0.92 | 0.61 | 0.60 | 0.86 |
| 30 | 90 | 0.42 | 0.42 | 0.38 | 0.41 | 0 | 0 | 0.23 | 0.97 | 0.97 | 0.96 | 0.97 | 0.67 | 0.67 | 0.94 |
| 30 | 120 | 0.73 | 0.73 | 0.69 | 0.73 | 0 | 0 | 0.51 | 0.99 | 0.99 | 0.99 | 0.99 | 0.72 | 0.72 | 0.98 |
| 60 | 60 | 0.01 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0.00 | 0.93 | 0.93 | 0.91 | 0.93 | 0.60 | 0.58 | 0.88 |
| 60 | 90 | 0.22 | 0.23 | 0.16 | 0.21 | 0 | 0 | 0.06 | 0.97 | 0.97 | 0.97 | 0.97 | 0.67 | 0.66 | 0.95 |
| 60 | 120 | 0.53 | 0.52 | 0.45 | 0.52 | 0 | 0 | 0.26 | 0.99 | 0.99 | 0.99 | 0.99 | 0.72 | 0.72 | 0.98 |
| 90 | 60 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.94 | 0.93 | 0.91 | 0.93 | 0.60 | 0.58 | 0.89 |
| 90 | 90 | 0.10 | 0.09 | 0.06 | 0.09 | 0 | 0 | 0.02 | 0.97 | 0.97 | 0.97 | 0.97 | 0.66 | 0.66 | 0.95 |
| 90 | 120 | 0.39 | 0.39 | 0.34 | 0.38 | 0 | 0 | 0.13 | 0.99 | 0.99 | 0.99 | 0.99 | 0.73 | 0.72 | 0.98 |

Table 4: Membership estimation based on Spectral (the proposed method), ZWZ19 Zhang et al. (2019a), and the vanilla $k-$means method without variance information for Model 1 with $\tau = 0.5$ and two error distributions.

The overall performance deteriorates in comparison to Table 4 since some individuals with shorter panel length are estimated with more noise. The spectral clustering methods ($S^g$ and S) perform comparably. Using just the diagonal information of the covariance matrix yields equally good performance, but not using the covariance information at all clearly performs worse. The vanilla k-means method again performs very similarly to spectral clustering without accounting for variance information. The method proposed by Zhang et al. (2019a) is competitive, improves upon estimates not using variance information, but is slightly inferior to PAM and our proposed spectral clustering methods.

### 4.2.2 Quantile regression with joint slope and grouping on intercepts

In this section, we consider the setting in Example 2.3. The spectral clustering approach is based on the estimators for the slopes and variances described in Section 3.2.3. More precisely, recall the definition of $\tilde{\alpha}_1, \tilde{\boldsymbol{\beta}}$ in (21) and $\hat{\Sigma}_{i,j}$ defined in (24).

The variation matrix $\hat{V}$ which we use as input to the spectral clustering algorithm is given by $\hat{V}_{ij} := \hat{\Sigma}_{i,j}^{-1/2} |\tilde{\alpha}_i - \tilde{\alpha}_j|$. For comparison, we also consider spectral clustering setting all variance estimators set to be equal, the naive $k$-means approach on estimated $\tilde{\alpha}_i$ from (21), and the convex clustering procedure of Gu and Volgushev (2019). Tuning parameters for Gu and Volgushev (2019) were set as described in the latter paper. The following model corresponds to DGP1 location scale shift model in Gu and Volgushev

| n | T | Perfect Match | | | | | | | Average Match | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 |
| $N(0,1), \tau = 0.5$ | | | | | | | | | | | | | | | |
| 30 | 60 | 0.31 | 0.32 | 0.26 | 0.31 | 0 | 0 | 0.06 | 0.95 | 0.96 | 0.94 | 0.96 | 0.70 | 0.69 | 0.84 |
| 30 | 90 | 0.67 | 0.69 | 0.60 | 0.67 | 0 | 0 | 0.23 | 0.98 | 0.99 | 0.98 | 0.98 | 0.74 | 0.73 | 0.88 |
| 30 | 120 | 0.87 | 0.88 | 0.83 | 0.87 | 0 | 0 | 0.32 | 0.99 | 1.00 | 0.99 | 1.00 | 0.79 | 0.78 | 0.88 |
| 60 | 60 | 0.13 | 0.13 | 0.07 | 0.13 | 0 | 0 | 0.01 | 0.96 | 0.96 | 0.95 | 0.96 | 0.68 | 0.67 | 0.84 |
| 60 | 90 | 0.49 | 0.50 | 0.43 | 0.50 | 0 | 0 | 0.09 | 0.99 | 0.99 | 0.98 | 0.99 | 0.75 | 0.73 | 0.86 |
| 60 | 120 | 0.77 | 0.78 | 0.74 | 0.78 | 0 | 0 | 0.19 | 1.00 | 1.00 | 1.00 | 1.00 | 0.81 | 0.78 | 0.86 |
| 90 | 60 | 0.05 | 0.05 | 0.03 | 0.05 | 0 | 0 | 0.00 | 0.96 | 0.97 | 0.96 | 0.97 | 0.69 | 0.68 | 0.83 |
| 90 | 90 | 0.38 | 0.38 | 0.30 | 0.36 | 0 | 0 | 0.05 | 0.99 | 0.99 | 0.99 | 0.99 | 0.75 | 0.74 | 0.84 |
| 90 | 120 | 0.71 | 0.71 | 0.61 | 0.70 | 0 | 0 | 0.08 | 1.00 | 1.00 | 1.00 | 1.00 | 0.81 | 0.79 | 0.83 |
| $t(3), \tau = 0.5$ | | | | | | | | | | | | | | | |
| 30 | 60 | 0.20 | 0.20 | 0.14 | 0.19 | 0 | 0 | 0.02 | 0.93 | 0.94 | 0.91 | 0.94 | 0.67 | 0.66 | 0.80 |
| 30 | 90 | 0.51 | 0.53 | 0.44 | 0.52 | 0 | 0 | 0.11 | 0.97 | 0.98 | 0.97 | 0.98 | 0.72 | 0.71 | 0.84 |
| 30 | 120 | 0.74 | 0.76 | 0.70 | 0.75 | 0 | 0 | 0.19 | 0.99 | 0.99 | 0.99 | 0.99 | 0.76 | 0.76 | 0.86 |
| 60 | 60 | 0.03 | 0.04 | 0.02 | 0.04 | 0 | 0 | 0.00 | 0.94 | 0.95 | 0.93 | 0.95 | 0.66 | 0.66 | 0.79 |
| 60 | 90 | 0.30 | 0.31 | 0.26 | 0.30 | 0 | 0 | 0.02 | 0.98 | 0.98 | 0.98 | 0.98 | 0.72 | 0.71 | 0.82 |
| 60 | 120 | 0.61 | 0.62 | 0.56 | 0.60 | 0 | 0 | 0.08 | 0.99 | 0.99 | 0.99 | 0.99 | 0.77 | 0.75 | 0.83 |
| 90 | 60 | 0.01 | 0.01 | 0.00 | 0.01 | 0 | 0 | 0.00 | 0.95 | 0.95 | 0.93 | 0.95 | 0.67 | 0.66 | 0.78 |
| 90 | 90 | 0.18 | 0.18 | 0.14 | 0.18 | 0 | 0 | 0.01 | 0.98 | 0.98 | 0.98 | 0.98 | 0.72 | 0.71 | 0.80 |
| 90 | 120 | 0.49 | 0.50 | 0.41 | 0.49 | 0 | 0 | 0.03 | 0.99 | 0.99 | 0.99 | 0.99 | 0.78 | 0.76 | 0.79 |

Table 5: Membership estimation based on Spectral (the proposed method), ZWZ19 Zhang et al. (2019a), and the vanilla $k-$means method without variance information for Model 2 with $\tau = 0.5$ and two error distributions.

(2019).

**Model 4:**

$$y_{it} = \alpha_i + x_{it}\beta + (1 + x_{it}\gamma)e_{it}\,.$$

where $e_{it} \overset{iid}{\sim} N(0,1)$ or $e_{it} \overset{iid}{\sim} t(3)$, $\alpha_i \in \{1,2,3\}$ with the same proportions, and $\beta = 1, \gamma = 0.1, x_{it} = \gamma_i + v_{it}$, where $\gamma_i$ and $v_{it}$ are independent and identically distributed from standard normal distribution over $i, t$, respectively.

Tables 7 summarizes the proportion of perfect classification and the average of the percentage of correct classification based on the proposed method with both exponential and the Gaussian kernel (denoted as S and $S^g$ respectively). Spectral clustering ignoring variance information is denoted as S-Iden, and $k$-means clustering on $\tilde{\alpha}_i$ is denoted as $k$-means$^\circ$ along with the PAM method for clustering. The procedure from Gu and Volgushev (2019) is denoted as GV.

In this model, including variance information is not helpful (S versus S-Iden). A possible explanation for variance information not being useful in this model is that the $\alpha_i$ are one-dimensional and there are no directions of larger or smaller variation in their estimates. The PAM and vanilla k-means performs identical in this model. The key difference is that PAM picks a representative point as the center of a group while k-means will take a cluster based average, this does not materialize any differences for grouping estimation in this Model. The method proposed in Gu and Volgushev (2019), which uses convex clustering method

| n | Perfect Match | | | | | | | Average Match | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 | $S^g$ | S | PAM | S-Diag | S-Iden | $k$-means$^\circ$ | ZWZ19 |
| $N(0,1), \tau = 0.5$ | | | | | | | | | | | | | | |
| 30 | 0.09 | 0.10 | 0.05 | 0.08 | 0 | 0 | 0.04 | 0.92 | 0.93 | 0.86 | 0.92 | 0.61 | 0.61 | 0.88 |
| 60 | 0.01 | 0.02 | 0.00 | 0.01 | 0 | 0 | 0.00 | 0.94 | 0.94 | 0.87 | 0.93 | 0.59 | 0.59 | 0.90 |
| 90 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.93 | 0.93 | 0.88 | 0.93 | 0.59 | 0.58 | 0.90 |
| $t(3), \tau = 0.5$ | | | | | | | | | | | | | | |
| 30 | 0.04 | 0.04 | 0.02 | 0.04 | 0 | 0 | 0.01 | 0.89 | 0.90 | 0.82 | 0.89 | 0.58 | 0.58 | 0.84 |
| 60 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.91 | 0.91 | 0.83 | 0.91 | 0.57 | 0.56 | 0.86 |
| 90 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.91 | 0.91 | 0.83 | 0.91 | 0.56 | 0.55 | 0.87 |

Table 6: Membership estimation based on Spectral (the proposed method), ZWZ19 Zhang et al. (2019a), and the vanilla $k-$means method without variance information for Model 3 with $\tau = 0.5$ and two error distributions.

to group the intercept shows slightly inferior performance for smaller $T$, but is otherwise comparable for larger $T$.

| n | T | Perfect Match | | | | | | Average Match | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | S | $S^g$ | S-Iden | PAM | $k$-means$^\circ$ | GV | S | $S^g$ | S-Iden | PAM | $k$-means$^\circ$ | GV |
| $N(0,1)$, $\tau = 0.5$ | | | | | | | | | | | | | |
| 30 | 15 | 0.07 | 0.08 | 0.06 | 0.05 | 0.05 | 0.03 | 0.90 | 0.90 | 0.89 | 0.89 | 0.91 | 0.67 |
| 30 | 30 | 0.52 | 0.53 | 0.52 | 0.45 | 0.49 | 0.39 | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.88 |
| 30 | 60 | 0.94 | 0.94 | 0.94 | 0.91 | 0.92 | 0.91 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| 60 | 15 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 | 0.92 | 0.92 | 0.91 | 0.90 | 0.92 | 0.66 |
| 60 | 30 | 0.35 | 0.36 | 0.36 | 0.27 | 0.34 | 0.22 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.90 |
| 60 | 60 | 0.91 | 0.90 | 0.90 | 0.87 | 0.90 | 0.86 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| 90 | 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.91 | 0.91 | 0.91 | 0.90 | 0.92 | 0.66 |
| 90 | 30 | 0.19 | 0.20 | 0.20 | 0.16 | 0.17 | 0.12 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.91 |
| 90 | 60 | 0.89 | 0.88 | 0.88 | 0.84 | 0.87 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| $t(3)$, $\tau = 0.5$ | | | | | | | | | | | | | |
| 30 | 15 | 0.01 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.88 | 0.88 | 0.85 | 0.86 | 0.88 | 0.64 |
| 30 | 30 | 0.34 | 0.34 | 0.34 | 0.29 | 0.32 | 0.23 | 0.96 | 0.96 | 0.96 | 0.95 | 0.96 | 0.84 |
| 30 | 60 | 0.88 | 0.88 | 0.88 | 0.83 | 0.85 | 0.80 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.98 |
| 60 | 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.88 | 0.88 | 0.88 | 0.86 | 0.88 | 0.60 |
| 60 | 30 | 0.16 | 0.16 | 0.17 | 0.13 | 0.15 | 0.08 | 0.97 | 0.97 | 0.97 | 0.96 | 0.97 | 0.85 |
| 60 | 60 | 0.78 | 0.79 | 0.78 | 0.73 | 0.78 | 0.71 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.98 |
| 90 | 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.89 | 0.89 | 0.89 | 0.86 | 0.89 | 0.55 |
| 90 | 30 | 0.05 | 0.06 | 0.04 | 0.04 | 0.05 | 0.03 | 0.97 | 0.97 | 0.97 | 0.96 | 0.97 | 0.85 |
| 90 | 60 | 0.69 | 0.68 | 0.69 | 0.63 | 0.66 | 0.63 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 |

Table 7: Membership estimation based on Spectral, $k$-means and the method proposed in Gu and Volgushev (2019) (GV) for Model 4 with $\tau = 0.5$.

### 4.3 Determining the Number of Groups

In this section, we compare the proposed heuristic in (5) for selecting the number of groups with other proposals from the literature. A general principle for determining the number of clusters using cross-validation (CV) in combination with the stability of cluster assignments was proposed by Wang (2010) and adapted to quantile regression with grouping on the

slopes in Zhang et al. (2019a). The underlying idea is directly applicable to any clustering algorithm, and hence we consider two versions: CV-kmeans corresponding to the proposal of Zhang et al. (2019a), and CV-Spectral which uses spectral clustering as proposed in the present paper as the underlying clustering algorithm. The maximum numbers of clusters to consider, denoted by Gmax, is set to 10 throughout. Results for Model 1 are presented in Table 8 and those for Model 2 are summarized in Table 9. All results reported in this section are based on 500 simulation repetitions.

For Model 1, the proposed heuristic has the best performance for all settings except for $t(3)$ errors with $n = 30, 60, T = 60$ where the CV–Spectral outperforms slightly. CV–Spectral shows better performance than CV–kmeans consistently. We note that Model 1 is perfectly symmetric with an odd number of groups, this corresponds to a setting that is favorable for stability–based methods. Model 2 demonstrates a situation where stability based method performs badly.

Model 2 corresponds to an even number of groups, and both CV methods fail in this setting because they always pick 2 groups. In light of the findings in Ben-David et al. (2006), this is not surprising; see also von Luxburg (2010). The issue is that a wrong grouping with two groups corresponding to coefficients $(0.1, 0.1), (0.2, 0.2)$ in one group and $(3, 3), (3.1, 3.1)$ in the other is very stable under variations of the data which leads to confusion of the stability–based methods. In the online appendix, we plot the paths of cross-validated stability scores for different $n, T$ combinations and different realizations of the data (Section 8 Figure 6). For larger $T$ there is a local minimum at the true number of groups $G = 4$, but the global minima are always at $G = 2$. The proposed heuristic works reasonably well and is able to pick up the correct number of groups as $T$ increases.

Model 4 corresponds to common slopes and group structure on the intercept (see also Example 2.3). Since this setting was also considered in Gu and Volgushev (2019), we consider the information criterion proposed in there. Results are presented in Table 10. We also include cross–validation with spectral clustering, denoted by CV-spectral, for comparison. Note that CV–kmeans is not applicable in this setting.

For $\tau = 0.5, n = 30, T = 15$, the best performing method is Gu and Volgushev (2019) with about a $10\% - 15\%$ advantage over the other two methods which show comparable performance. In all other settings, CV-Spectral is the best or close to best (within 5%) performer. The heuristic method performs better or is similar to Gu and Volgushev (2019) for most cases with $n = 90, T \geq 60$ while the results between those two are mixed in other settings.

In conclusion, there is no clear winner that performs best across all models and settings. This is not surprising because selecting the number of clusters is a very difficult problem in general. This also explains why there exists no unifying approach for selecting the number of groups. Our proposed eigenvalue heuristic is competitive in most cases considered, and clearly the best on some. Stability–based methods have two major limitations: they cannot

select one group by construction, and they can fail for models with stable clusters for the wrong number of groups. The information criterion in Gu and Volgushev (2019) can select one group and performs well when $n, T$ are smaller but falls behind when $n$ is large. No information criterion is known for quantile regression models with unrestricted intercepts and grouping on the slopes. Such a criterion could potentially be derived, but it would only be valid in this specific setting and we refrained from taking this route since we aimed to propose a method that is applicable in more generality.

| n | T | $N(0,1)$ | | | | | $t(3)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **≥5** | **1** | **2** | **3** | **4** | **≥5** |
| CV-Spectral, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 60 | – | 0.05 | **0.84** | 0.09 | 0.02 | – | 0.08 | **0.72** | 0.15 | 0.05 |
| 30 | 90 | – | 0.01 | **0.98** | 0.01 | 0.00 | – | 0.03 | **0.91** | 0.06 | 0.00 |
| 30 | 120 | – | 0.01 | **0.99** | 0.00 | 0.00 | – | 0.01 | **0.98** | 0.01 | 0.00 |
| 60 | 60 | – | 0.01 | **0.98** | 0.01 | 0.00 | – | 0.01 | **0.95** | 0.02 | 0.02 |
| 60 | 90 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.00 | **0.99** | 0.01 | 0.00 |
| 60 | 120 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.00 | **1.00** | 0.00 | 0.00 |
| CV-kmeans, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 60 | – | 0.29 | **0.40** | 0.17 | 0.14 | – | 0.34 | **0.32** | 0.17 | 0.17 |
| 30 | 90 | – | 0.13 | **0.69** | 0.12 | 0.06 | – | 0.19 | **0.58** | 0.16 | 0.07 |
| 30 | 120 | – | 0.13 | **0.80** | 0.06 | 0.01 | – | 0.13 | **0.72** | 0.11 | 0.04 |
| 60 | 60 | – | 0.09 | **0.39** | 0.25 | 0.27 | – | 0.13 | **0.27** | 0.16 | 0.44 |
| 60 | 90 | – | 0.06 | **0.74** | 0.15 | 0.05 | – | 0.07 | **0.63** | 0.18 | 0.12 |
| 60 | 120 | – | 0.05 | **0.85** | 0.08 | 0.02 | – | 0.03 | **0.78** | 0.15 | 0.04 |
| Heuristic, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 60 | 0.00 | 0.07 | **0.91** | 0.02 | 0.00 | 0.05 | 0.25 | **0.69** | 0.01 | 0.00 |
| 30 | 90 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **0.98** | 0.02 | 0.00 |
| 30 | 120 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **0.99** | 0.01 | 0.00 |
| 60 | 60 | 0.01 | 0.00 | **0.98** | 0.01 | 0.00 | 0.09 | 0.07 | **0.84** | 0.00 | 0.00 |
| 60 | 90 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 |
| 60 | 120 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 |

Table 8: Percentage of estimated number of groups based on CV-Spectral, CV-kmeans, and Heuristic for Model 1 with $\tau = 0.5$. The true $G$ is 3 (highlighted column).

# 5 Empirical Applications

## 5.1 Heterogeneity in environmental Kuznet curves

We first apply our methodology to a panel data quantile regression analysis on the environmental Kuznet curves (EKC). The concept first emerged in the influential study of Grossman and Krueger (1991). Various empirical studies have since then provided evidence in different countries that there exists an inverse-U relationship between economic development and the pollution level. As income per capita increases, we expect to see first deterioration of the environment, and then an improvement as income continues to rise. Understanding the relationship between pollution and per capita income is important for the design of the optimal environmental policy. Here we focus our analysis on using

| n | T | $N(0,1)$ | | | | | $t(3)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **≥5** | **1** | **2** | **3** | **4** | **≥5** |
| CV-Spectral, $\tau = 0.5$ | | | | | | | | | | | |
| 40 | 40 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 40 | 80 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 40 | 160 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 40 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 80 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 160 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| CV-kmeans, $\tau = 0.5$ | | | | | | | | | | | |
| 40 | 40 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 40 | 80 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 40 | 160 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 40 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 80 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| 60 | 160 | – | 1.00 | 0.00 | **0.00** | 0.00 | – | 1.00 | 0.00 | **0.00** | 0.00 |
| Heuristic, $\tau = 0.5$ | | | | | | | | | | | |
| 40 | 40 | 0.00 | 0.70 | 0.00 | **0.30** | 0.00 | 0.00 | 0.92 | 0.00 | **0.08** | 0.00 |
| 40 | 80 | 0.00 | 0.00 | 0.00 | **0.99** | 0.01 | 0.00 | 0.02 | 0.00 | **0.97** | 0.01 |
| 40 | 160 | 0.00 | 0.00 | 0.00 | **0.99** | 0.01 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 |
| 60 | 40 | 0.00 | 0.49 | 0.00 | **0.51** | 0.00 | 0.00 | 0.91 | 0.00 | **0.09** | 0.00 |
| 60 | 80 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.01 | 0.00 | **0.99** | 0.00 |
| 60 | 160 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 |

Table 9: Percentage of estimated number of groups with CV-Spectral, CV-kmeans and Heuristic methods for Model 2 with $\tau = 0.5$.. The true $G$ is 4 (highlighted column).

state-level panel data in the United States during the period of 1929 - 1994 and for brevity, we focus on the emission of $SO_2$. The dataset is available from the National Air Pollutant Emission Trends, 1900 - 1994, published by the US Environmental Protection Agency. Most early empirical work on EKC uses least squares methods pooling all the states together and utilizes either a quadratic or cubic specification to estimate the relationship between the emission level and per capita income. Millimet et al. (2003) discusses in detail some of the model specification issues and explores semi–parametric methods that provide a set of more flexible modelling tools. Given concerns that different states may take a different environmental transition path as income level arises, List and Gallet (1999) estimates the EKC with both the quadratic and cubic specification state by state to account for potential state heterogeneity. They then group these states into three groups depending on whether the estimated peak of the state-specific EKC falls below, inside or above the 95% confidence interval implied by a pooled model. This provides an interesting piece of evidence for some form of group heterogeneity, yet how group membership is constructed is ad hoc and does not account for the statistical uncertainty of the state-specific least square estimates. On the other hand, Flores et al. (2014) has criticized the least square approach and advocates the use of quantile regression methods. They document that quantile regression offers a more complete picture of the relationship between pollution and income. However, for a given quantile, they estimate the panel data quantile regression with state fixed effect with-

| n | T | $N(0,1)$ | | | | | $t(3)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **≥5** | **1** | **2** | **3** | **4** | **≥5** |
| CV-Spectral, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 15 | – | 0.35 | **0.43** | 0.09 | 0.13 | – | 0.45 | **0.31** | 0.09 | 0.15 |
| 30 | 30 | – | 0.04 | **0.92** | 0.03 | 0.01 | – | 0.11 | **0.79** | 0.07 | 0.03 |
| 30 | 60 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.01 | **0.99** | 0.00 | 0.00 |
| 60 | 15 | – | 0.13 | **0.63** | 0.02 | 0.22 | – | 0.20 | **0.44** | 0.01 | 0.35 |
| 60 | 30 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.01 | **0.97** | 0.01 | 0.01 |
| 60 | 60 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.00 | **1.00** | 0.00 | 0.00 |
| 90 | 15 | – | 0.09 | **0.79** | 0.00 | 0.12 | – | 0.18 | **0.61** | 0.01 | 0.20 |
| 90 | 30 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.00 | **1.00** | 0.00 | 0.00 |
| 90 | 60 | – | 0.00 | **1.00** | 0.00 | 0.00 | – | 0.00 | **1.00** | 0.00 | 0.00 |
| Heuristic, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 15 | 0.05 | 0.40 | **0.45** | 0.06 | 0.04 | 0.10 | 0.49 | **0.31** | 0.06 | 0.04 |
| 30 | 30 | 0.01 | 0.04 | **0.92** | 0.02 | 0.01 | 0.00 | 0.15 | **0.79** | 0.03 | 0.03 |
| 30 | 60 | 0.00 | 0.00 | **0.99** | 0.00 | 0.01 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 |
| 60 | 15 | 0.18 | 0.37 | **0.44** | 0.00 | 0.01 | 0.50 | 0.28 | **0.22** | 0.00 | 0.00 |
| 60 | 30 | 0.00 | 0.01 | **0.99** | 0.00 | 0.00 | 0.00 | 0.04 | **0.95** | 0.00 | 0.01 |
| 60 | 60 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 |
| 90 | 15 | 0.02 | 0.31 | **0.64** | 0.01 | 0.02 | 0.12 | 0.40 | **0.46** | 0.00 | 0.02 |
| 90 | 30 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.02 | **0.98** | 0.00 | 0.00 |
| 90 | 60 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 |
| GV, $\tau = 0.5$ | | | | | | | | | | | |
| 30 | 15 | 0.00 | 0.06 | **0.51** | 0.34 | 0.09 | 0.00 | 0.16 | **0.50** | 0.27 | 0.08 |
| 30 | 30 | 0.00 | 0.00 | **0.81** | 0.16 | 0.03 | 0.00 | 0.01 | **0.76** | 0.20 | 0.04 |
| 30 | 60 | 0.00 | 0.00 | **0.98** | 0.02 | 0.00 | 0.00 | 0.00 | **0.96** | 0.03 | 0.00 |
| 60 | 15 | 0.00 | 0.04 | **0.52** | 0.32 | 0.12 | 0.00 | 0.11 | **0.44** | 0.33 | 0.12 |
| 60 | 30 | 0.00 | 0.00 | **0.86** | 0.13 | 0.02 | 0.00 | 0.00 | **0.78** | 0.18 | 0.04 |
| 60 | 60 | 0.00 | 0.00 | **0.99** | 0.01 | 0.00 | 0.00 | 0.00 | **0.98** | 0.02 | 0.00 |
| 90 | 15 | 0.00 | 0.03 | **0.51** | 0.32 | 0.14 | 0.00 | 0.11 | **0.37** | 0.33 | 0.19 |
| 90 | 30 | 0.00 | 0.00 | **0.88** | 0.11 | 0.01 | 0.00 | 0.00 | **0.79** | 0.17 | 0.03 |
| 90 | 60 | 0.00 | 0.00 | **0.99** | 0.01 | 0.00 | 0.00 | 0.00 | **0.98** | 0.02 | 0.00 |

Table 10: Percentage of estimated number of groups based on CV-Spectral, Heuristic, and Gu and Volgushev (2019) (GV) methods for Model 3 with $\tau = 0.5$. . The true $G$ is 3 (highlighted column).

out allowing the EKC coefficients to be state-dependent. Combining the insights of List and Gallet (1999) and Flores et al. (2014), we apply our methodology in a panel data quantile regression model which allows individual fixed effects while estimating the group structure of the slope coefficients that determine the shape of the EKC curves across different states.

For a given quantile level $\tau$, our model specification is:

$$q_{i,\tau}(Z_{it}) = \alpha_i(\tau) + \lambda_t(\tau) + Z_{it}\beta_{1,g_i}(\tau) + Z_{it}^2\beta_{2,g_i}(\tau)\,,$$

where $i$ corresponds to states, $t$ it the time index and $g_i$ records the group membership. We denote by $q_{i,\tau}(Z_{it})$ the conditional quantile function of $Y_{it}$ given $Z_{it}$ where the response $Y_{it}$ is the state-year per capita emission level of $SO_2$ and $Z_{it}$ is the per capita real income using 1987 dollar. We focus on the quadratic specification for better visualization of the estimation results. Cubic specification leads to similar grouping results. Other control

variables can be added, for example, population density and the number of days with extreme temperature as considered in Flores et al. (2014). However, Flores et al. (2014) report that these additional control variables do not change the estimates for the quadratics of the EKC. We first obtain state-specific estimates $\hat{\beta}_{1,i}(\tau), \hat{\beta}_{2,i}(\tau)$ as well as their associated covariance matrix.

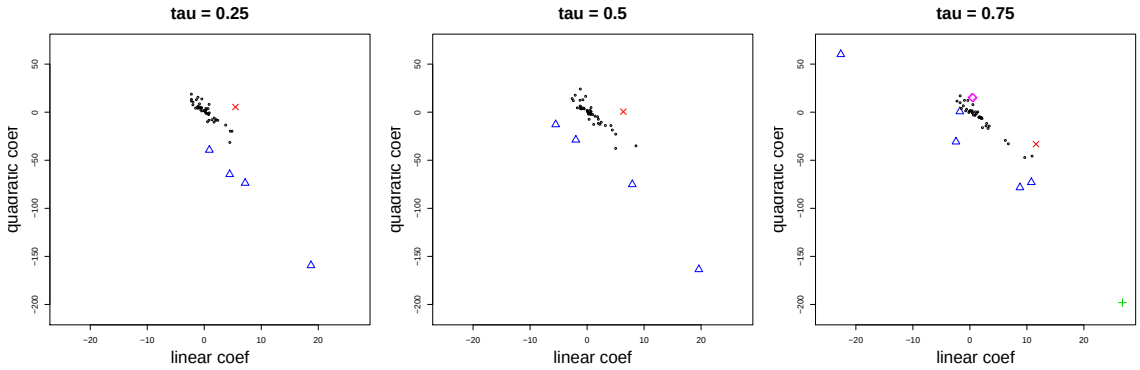



Figure 2: Raw state specific estimates for $\hat{\beta}_{1,i}(\tau), \hat{\beta}_{2,i}(\tau))$ and grouping of states for quantile levels $\tau = \{0.25, 0.5, 0.75\}$. Each symbol represents a different group.

To estimate the number of groups for different quantile levels $\tau = \{0.25, 0.5, 0.75\}$, we apply the heuristic in (5); see Figure 7 in the online supplement for corresponding plots. For both 25 and 50th quantile, we find three groups and for 75th quantile, we find 5 groups. Given these estimates, we then apply the spectral clustering method on these raw estimates, accounting for the statistical uncertainty. Figure 2 shows the estimated group membership for $(\hat{\beta}_{1,i}(\tau), \hat{\beta}_{2,i}(\tau))$. Noticeably, for both 25th and 50th quantile, the grouping of the states are the same. The red cross in Figure 2 corresponds to West Virginia, while the blue triangles correspond to Arizona, Montana, Nevada, and Utah. A close inspection of the data suggests that the EKC for West Virginia looks to be closer to a linear trend within the range of years under consideration, while Arizona, Montana, Nevada, and Utah are states that have relatively higher emission level and a much more positive linear coefficient and a much negative quadratic coefficient when compared to all other states. Interestingly, these four states are also noted as the "outlier" states in Flores et al. (2014) which documents that the residuals of these states are alarmingly high. Since their specification requires the EKC coefficients to be the same for all states, this provides some evidence that these states might have a different EKC. This is clearly confirmed by our analysis. For the 75th quantile, the State of Arizona has a more extreme estimate and now becomes a group by herself, as well as West Virginia. Two smaller groups consist of North Dakota and Wyoming as one group and Illinois, Montana, Nevada, New Mexico, and Utah as the other group.

We note that some groups resulting from this empirical analysis are very small. The results should thus be interpreted with caution since this violates our theoretical assumptions which require proportional group sizes to be bounded from below.

## 5.2 Heterogeneity in intergenerational income mobility

The study on intergenerational income mobility across the United States by Chetty et al. (2014), Chetty et al. (2018) and Chetty and Hendren (2018) has been influential. Using tax records on the entire U.S. population, they document how children's expected incomes conditional on their parents' incomes differ across different geographical regions in the United States. Although the raw data used to obtain these estimates are not publicly available, they publish the region specific estimates at the commuting zone, country or census tract level, together with their associated standard errors. These estimates are used for policy purposes to encourage welfare improvements for children resides in the areas that have low mobility rates as for instance considered in Bergman et al. (2019). The categorization of a region having low mobility is often solely based on the point estimates without accounting for the associated statistical uncertainty. Our analysis focuses on the plausible hypothesis that although different geographical locations are likely to have heterogeneous mobility ratings, they may be divided into a few distinct groups and we let the data determine the number of groups utilizing both point estimates and their levels of precision. It is worth noting that, in contrast to most proposals in the existing literature, our method remains applicable even when raw individual-level data are not available due to privacy or other concerns and only estimated coefficients and their uncertainty estimates are given.

We focus on the 100 most populous commuting zones. Let the point estimates of the income mobility to be $\hat{\beta}_i$ and the associated standard error to be $\hat{\Sigma}_i$. To apply the heuristic for the estimation of the number of groups, we also know $T_i$ which is the amount of data that leads to the estimates $(\hat{\beta}_i, \hat{\Sigma}_i)$.[5]

We first use our method to select the number of groups. The left plot in Figure 5.2 shows that the number of groups is estimated to be nine and the right plot illustrates the gap of the adjacent eigen values. We then apply our algorithm to estimate the group membership, which is illustrated in Figure 5.2. Further details on the grouping of the hundred most populous commuting zones is provided in Table 11. Fayetteville and Memphis have the lowest point estimates for their income mobility among all the hundred commutting zones considered and they are grouped together. There are ten commuting zones grouped together as the top tier. The grouping provides a parsimonious description of the mobility heterogeneity. It also suggests that citizens in the commuting zones that belong to the same group, although having different point estimates, are likely to have similar true mobility ratings.

[5] All these information are publicly available from `https://opportunityinsights.org/data`. Note in this application, $T_i$ varies across individuals, we take the minimum $T_i$ when constructing the scaled dissimilarity measure for the estimation of $G$. Using the average value of $T_i$ leads to similar result.

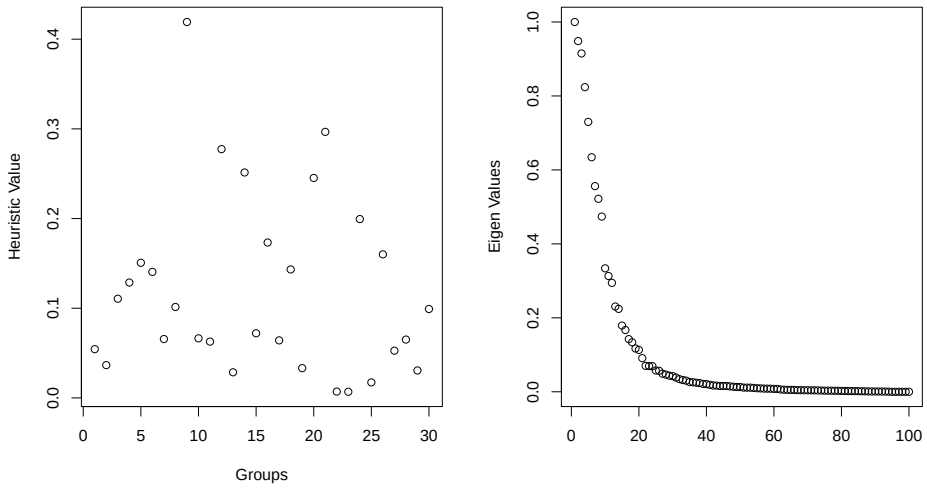



Figure 3: The heuristic value for group selection and the associated eigen values for the 100 most populous commuting zones using publicly data in Chetty and Hendren (2018).

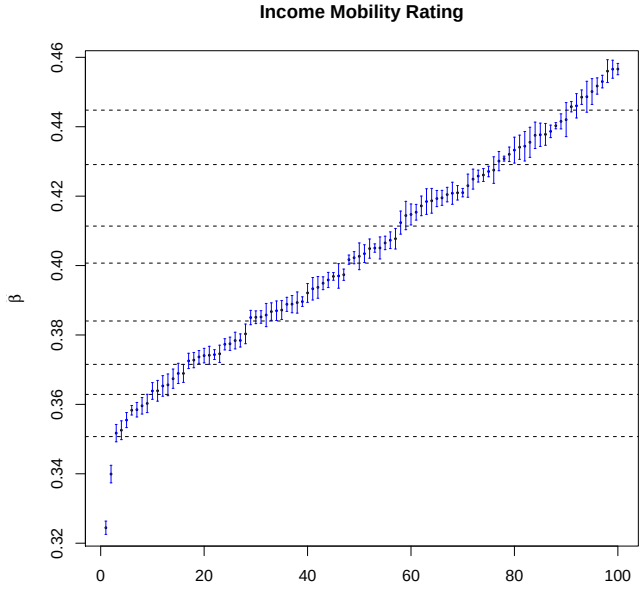



Figure 4: Points in the figure are the sorted point estimates $\hat{\beta}_i$ for the 100 most populous commuting zones in the United States. The blue bars indicates the confidence set of each point estimates with $\pm$ 2 s.e. The dotted line are the division lines for the 9 groups based on the estimated group membership.

| | Grouped Commuting Zones |
|---|---|
| 1 | Boston, Des Moines, Honolulu, Minneapolis, Newark, Toms River, Salt Lake City San Francisco, San Jose, Scranton |
| 2 | Albany, Allentown, Brownsville, Los Angeles, Madison, Manchester, New York Pittsburgh, Providence, Reading, Santa Barbara, Santa Rosa, Seattle, Spokane |
| 3 | Bakersfield, Buffalo, Bridgeport, Canton, Denver, El Paso, Erie Harrisburg, Houston,Modesto ,Omaha, Portland, Poughkeepsie, Sacramento San Diego, Springfield, Syracuse, Washington DC |
| 4 | Austin, Eugene, Fort Worth, Miami, Oklahoma City, Philadelphia Rockford, San Antonio, Tulsa, Youngstown |
| 5 | Albuquerque, Baton Rouge, Cape Coral, Chicago, Cleveland, Dallas Fresno, Gary, Grand Rapids , Kansas City ,Las Vegas, Milwaukee, Orlando Port St. Lucie, Phoenix, Sarasota, South Bend, Toledo, Tucson |
| 6 | Baltimore, Cincinnati, Columbus, Dayton, Detroit, Louisville Nashville, New Orleans, Pensacola, St. Louis, Tampa, Virginia Beach |
| 7 | Knoxville, Indianapolis, Lakeland, Little Rock, Mobile, Raleigh, Richmond |
| 8 | Atlanta, Birmingham, Charlotte, Columbia, Greensboro, Greenville, Jacksonville |
| 9 | Fayetteville, Memphis |

Table 11: The 100 most populous commuting zones grouped using our method. First group are for those with the highest income mobility rating and the last group the lowest.

# 6 Conclusion

In this paper, we propose a general methodology for studying group heterogeneity of effects in panel data models. We provide high-level conditions for the proposed method to achieve correct group identification and verify these conditions for several leading non-linear models often applied in empirical studies. We demonstrate that incorporating uncertainty information in individual-level estimates is useful for estimating group patterns. Although we focus on non-linear models, our methodology is naturally applicable to linear models, as well to situations where micro-level data is not available and only summary statistics are accessible to the researcher. We have proposed a method for selecting the number of groups, but left the theoretical validation of this method open to future research.

There are several important questions that merit further research. Our implementation of the dissimilarity measure and its theoretical analysis requires independence across individuals. In many settings, dependence across individuals is present. Properly modeling such dependence and accounting for it in our approach is an important question.

Another substantial practical and theoretical challenge is dealing with short and highly unbalanced panels where individual level estimators are of very poor quality. In addition, as pointed out by the Associate Editor and a referee, there are cases when only some coefficients contain group structure. Using a two step procedure where coefficients containing group information are determined in a first step and grouping is only performed on those coefficients in a second step could lead to improvements in grouping accuracy. Implementing

such an approach poses substantial theoretical challenges and is worthwhile investigating. The Associate Editor and referee also suggested In some cases, re-scaling the covariates first to bring all coefficients on the same scale before clustering could also be advantageous. This merits further exploration which we leave for future research. Finally, extending our approach to more complex models with time-varying group heterogeneity is another natural next step which we plan to address in future research.

## References


Ando, T., J. Bai, and K. Li (2022). Bayesian and maximum likelihood analysis of large-scale panel choice models with unobserved heterogeneity. *Journal of Econometrics 230*(1), 20–38.

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

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4–29.

Ben-David, S., U. von Luxburg, and D. Pál (2006). A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pp. 5–19. Springer.

Bergman, P., R. Chetty, S. DeLuca, N. Hendren, L. F. Katz, and C. Palmer (2019). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Technical report, National Bureau of Economic Research.

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

Chao, S.-K., S. Volgushev, and G. Cheng (2017). Quantile processes for semi and nonparametric regression. *Electronic Journal of Statistics 11*(2), 3272–3331.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018). The opportunity atlas: Mapping the childhood roots of social mobility. Technical report, National Bureau of Economic Research.

Chetty, R. and N. Hendren (2018). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics 133*(3), 1107–1162.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics 129*(4), 1553–1623.

Chetverikov, D. and E. Manresa (2022). Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*.

Chung, F. and M. Radcliffe (2011). On the spectra of general random graphs. *The Electronic Journal of Combinatorics 18*(1), P215.

Chung, F. R. and F. C. Graham (1997). *Spectral Graph Theory*. Number 92. American Mathematical Soc.

Flores, C. A., A. Flores-Lagunes, and D. Kapetanakis (2014). Lessons from quantile panel estimation of the environmental kuznets curve. *Econometric Reviews 33*(8), 815–853.

Galvao, A. F., J. Gu, and S. Volgushev (2020). On the unbiased asymptotic normality of quantile regression with fixed effects. *Journal of Econometrics 218*(1), 178–215.

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

Grossman, G. M. and A. B. Krueger (1991). Environmental impacts of a north american free trade agreement. Working Paper 3914, National Bureau of Economic Research.

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

Harding, M. and C. Lamarche (2017). Penalized quantile regression with semiparametric correlated effects: An application with heterogeneous preferences. *Journal of Applied Econometrics 32*(2), 342–358.

Hendricks, W. and R. Koenker (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association 87*(417), 58–68.

Hocking, T. D., A. Joulin, F. Bach, and J.-P. Vert (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, pp. 1–7.

John, C. R., D. Watson, M. R. Barnes, C. Pitzalis, and M. J. Lewis (2020). Spectrum: Fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics 36*(4), 1159–1166.

Kato, K., A. F. Galvao Jr, and G. V. Montes-Rojas (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics 170*(1), 76–91.

Kaufman, L. and P. J. Rousseeuw (2005). *Finding groups in data: an introduction to cluster analysis*, Volume 344. John Wiley & Sons.

Ke, Z. T., J. Fan, and Y. Wu (2015). Homogeneity pursuit. *Journal of the American Statistical Association 110*(509), 175–194.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics 157*(2), 396–408.

Leng, X., W. Wang, and H. Chen (2023). Multi-dimensional latent group structures with heterogeneous distributions. *Journal of Econometrics 233*(1), 1–21.

Lin, C.-C. and S. Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods 1*(1), 42–55.

List, J. A. and C. A. Gallet (1999). The environmental kuznets curve: does one size fit all? *Ecological economics 31*(3), 409–423.

Little, A., M. Maggioni, and J. M. Murphy (2020). Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *Journal of Machine Learning Research 21*, 1–66.

Lumsdaine, R. L., R. Okui, and W. Wang (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics 233*(1), 45–65.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Miao, K., L. Su, and W. Wang (2020). Panel threshold regressions with latent group structures. *Journal of Econometrics 214*(2), 451–481.

Millimet, D. L., J. A. List, and T. Stengos (2003). The environmental kuznets curve: real progress or misspecified models? *Review of Economics and Statistics 85*(4), 1038–1047.

Ng, A. Y., M. I. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849–856.

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

Reynolds, A. P., G. Richards, B. de la Iglesia, and V. J. Rayward-Smith (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms 5*(4), 475–504.

Schubert, E. and P. J. Rousseeuw (2019). Faster k-medoids clustering: improving the pam, clara, and clarans algorithms. In *International conference on similarity search and applications*, pp. 171–187. Springer.

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

van Delft, A. and H. Dette (2021). A similarity measure for second order properties of non-stationary functional time series with applications to clustering and testing. *Bernoulli 27*(1), 469–501.

van der Vaart, A. and J. Wellner (1996). Weak convergence and empirical processes. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

von Luxburg, U. (2010). *Clustering stability: an overview*. Now Publishers Inc.

von Luxburg, U., M. Belkin, and O. Bousquet (2008). Consistency of spectral clustering. *The Annals of Statistics 36*(2), 555–586.

von Luxburg, U., O. Bousquet, and M. Belkin (2004). Limits of spectral clustering. *Advances in neural information processing systems 17*.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika 97*(4), 893–904.

Wang, W., P. C. Phillips, and L. Su (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics 33*(6), 797–815.

Wang, W. and L. Su (2021). Identifying latent group structures in nonlinear panels. *Journal of Econometrics 220*(2), 272–295.

Yu, Y., T. Wang, and R. J. Samworth (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika 102*(2), 315–323.

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

Zhang, Y., H. J. Wang, and Z. Zhu (2019b). Robust subgroup identification. *Statistica Sinica 29*(1), 1873–1889.

**ONLINE SUPPLEMENTARY MATERIAL**

In this supplementary material, we provide more simulation details, additional plots as well as proofs of the main results in Section 3.1 and Section 3.2.

# 7 Simulation Studies

We provide more details about the simulation studies in Section 4.

## 7.1 Simulations in Section 4.1

We implement the CLASSO estimator using CVX in Matlab with the mosek solver with version Mosek 8. The algorithm is initiated with $\boldsymbol{\beta}_i$ being the individual logistic regression estimates and $\eta_k$ being the origin for all $K_0$ groups. The algorithm is terminated when the objective function differs by a quantity less than 0.001 and when the $\ell_2$ norm of the estimated group centre $\eta_k$ changes by less than 0.1%.

## 7.2 Simulations in Section 4.2.1

Simulations are done in the **quantreg** package in R. Covariance estimates are computed using the function `summary.rq()` with option `se="nid"` and default bandwidth choice `hs=true`.

## 7.3 Simulations in Section 4.2.2

The bandwidth $d_T$ used in (24) is based on the method implemented in the **quantreg** package in R (function `summary.rq()` with `se="nid"` and default choice `hs=true`).

## 7.4 Simulations in Section 4.3

The maximum numbers of clusters to consider Gmax is set to Gmax = 5 for $n <= 30$ and Gmax = 10 for $n > 30$ cases for the CV methods, and we set Gmax = 10 across all settings for the heuristic method. For the cross-validation method, we use 100 (4:4:2) random splits of the dataset into training and validation data (see Wang (2010) for details on the meaning of this splitting).

# 8 Plots

## 8.1 Plots in Section 4.2

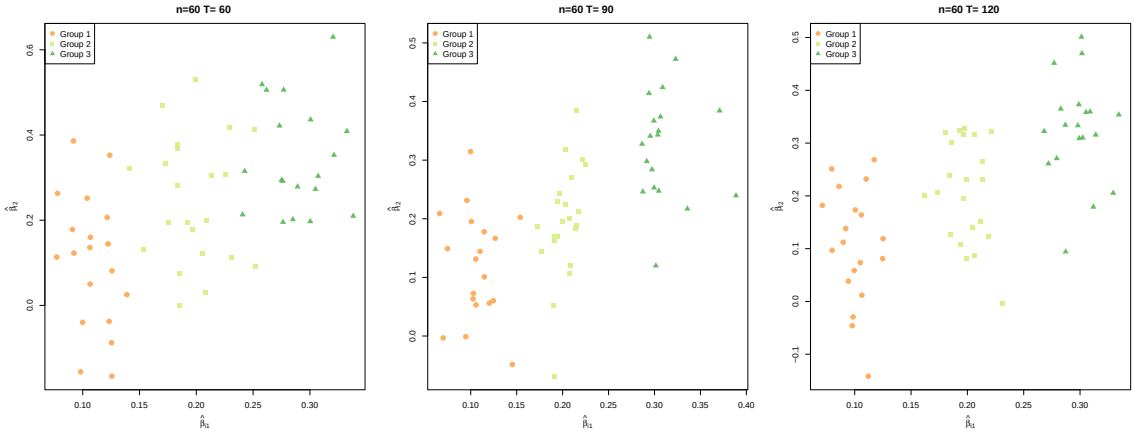
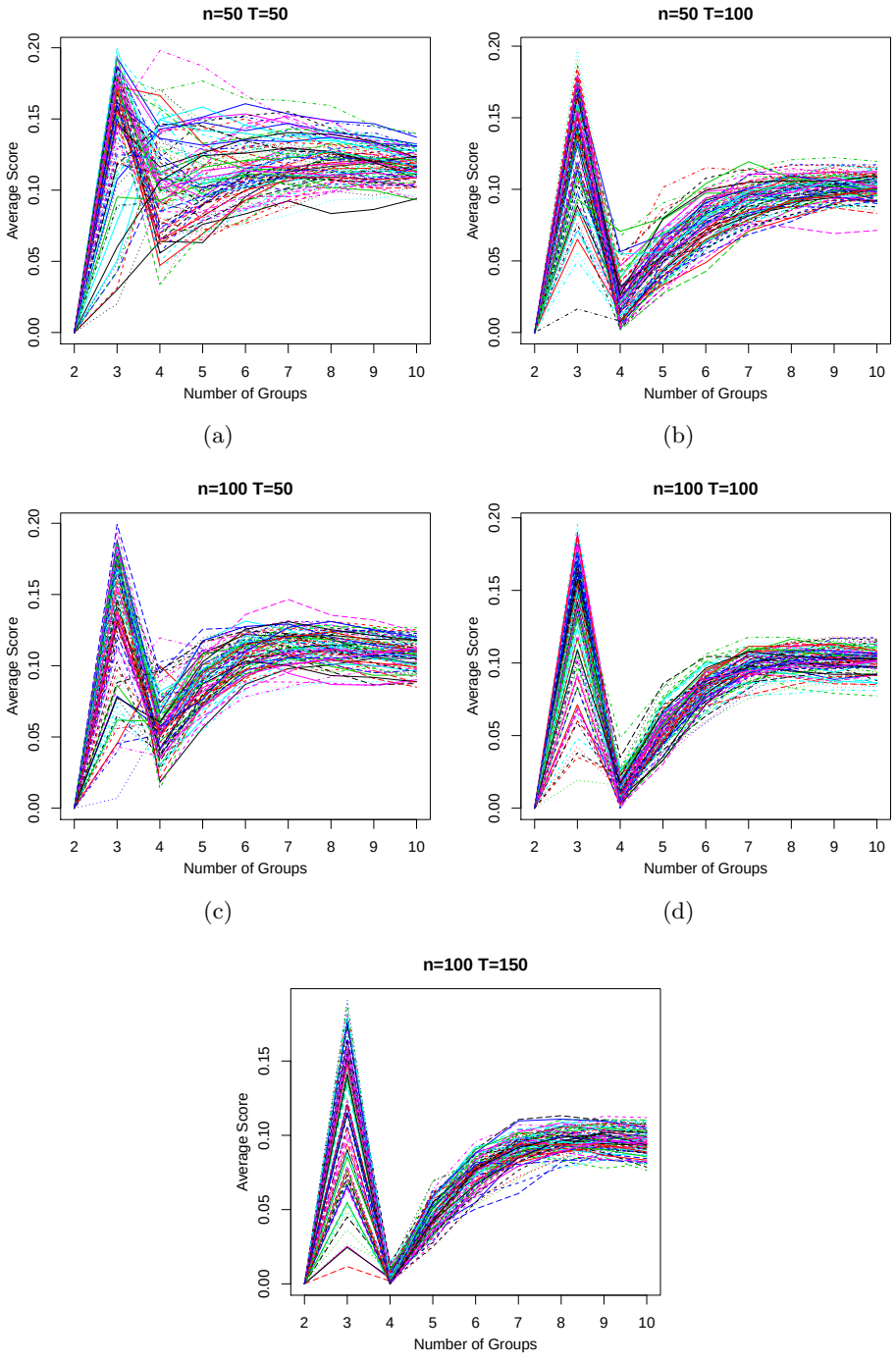



Figure 5: Scatter plots of $\{\hat{\boldsymbol{\beta}}_i\}_{i=1}^n$ for Model 1 with $t(3)$ error and $\tau = 0.5$.

## 8.2 Plots in Section 4.3




(a) (b)

(c) (d)

(e)

Figure 6: Stability score for Model 3 with $t(3)$ error and $\tau = 0.5$.

### 8.3 Plots in Section 5

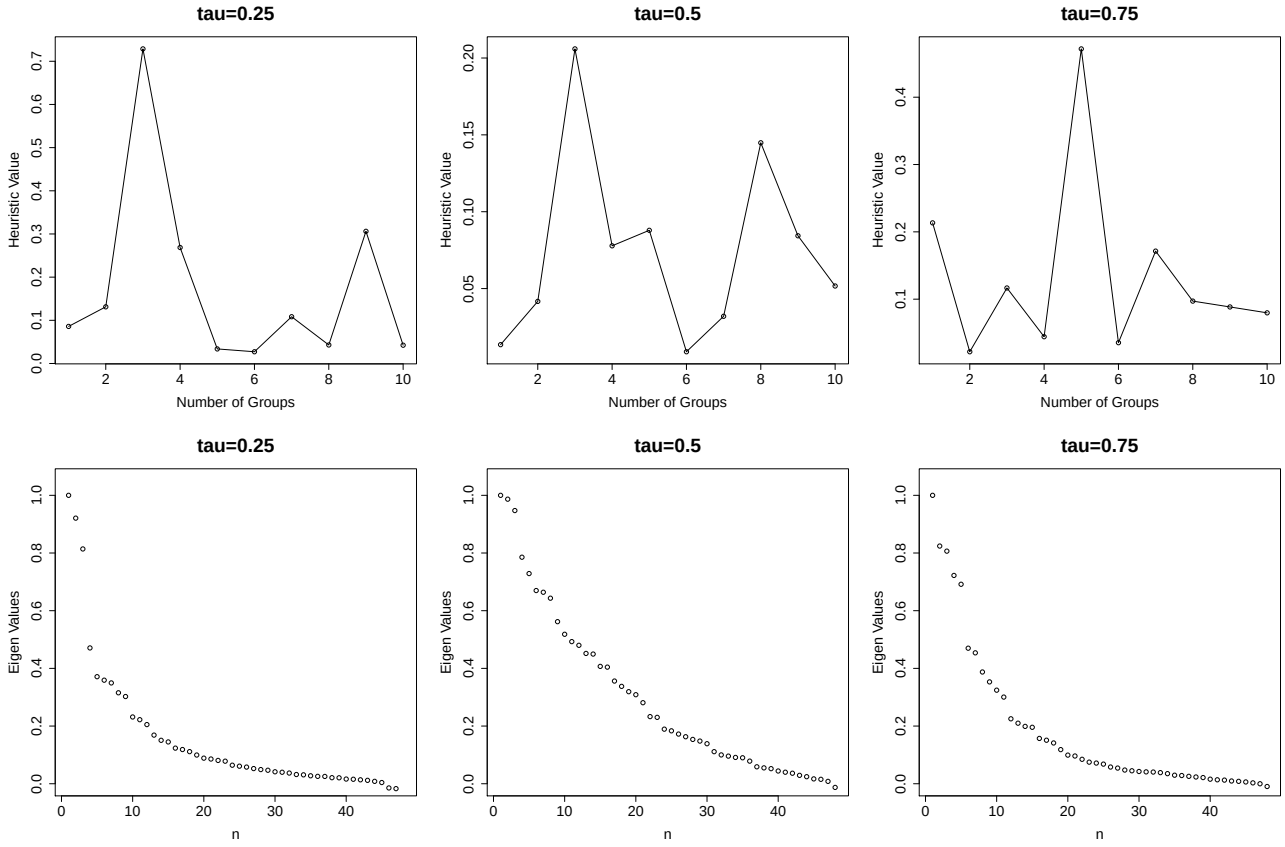



Figure 7: The heuristic values and the eigen-value plot for three different quantile levels $\tau = \{0.25, 0.5, 0.75\}$.

## 9 Proofs

### 9.1 Notation

Let $a_n \lesssim_p b_n$ denotes that there exists a non-random constant $C \in (0, \infty)$ that is independent of $n, T$, such that $\mathbb{P}(a_n \leq Cb_n) \to 1$. For a matrix $A \in \mathbb{R}^{n \times p}$, we define the operator norm of $A$ as the maximum absolute column sum of the matrix $|||A|||_\infty := \max_{1 \leq i \leq n} \sum_{j=1}^{p} |A_{ij}|$, define the Frobenius norm of $A$ as the square root of the sum of the absolute squares of all elements $|||A|||_{\mathrm{F}} := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} |A_{ij}^2|}$, and define the spectral norm of $A$ as its largest singular value $|||A|||_2 := \sigma_{\max}(A)$.

To lighten notation we abbreviate the true number of groups as $G$ instead of $G^*$ whenever there is no risk of confusion.

### 9.1.1 Proof of Theorem 2.1

The proof consists of two parts. First, we verify that under the assumptions made the following expansions and convergences hold. Second, we prove that (25)-(28) imply the statement of the Theorem.

For any (possibly random) sequence $\boldsymbol{\Delta}_{\boldsymbol{\gamma},T} := (\Delta_\alpha, \boldsymbol{\Delta}_{\boldsymbol{\beta}}) = \mathcal{O}_{\mathbb{P}}(1)$ we have an expansion of the form

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\gamma},T}) - \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*)$$
$$= \frac{1}{T^{1/2}} \sum_{t=1}^{T} \boldsymbol{\Delta}_{\boldsymbol{\gamma},T}^\top \nabla_{\boldsymbol{\gamma}} \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*) + \frac{1}{2}\boldsymbol{\Delta}_{\boldsymbol{\gamma},T}^\top A \boldsymbol{\Delta}_{\boldsymbol{\gamma},T} + o_{\mathbb{P}}(1). \quad (25)$$

Moreover,

$$\sqrt{T}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = \frac{1}{T^{1/2}} \sum_{t=1}^{T} A^{-1} \nabla_{\boldsymbol{\gamma}} \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*) + o_{\mathbb{P}}(1), \quad (26)$$

$$\frac{1}{T^{1/2}} \sum_{t=1}^{T} \nabla_{\boldsymbol{\gamma}} \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*) \xrightarrow{d} N(0, B) \quad (27)$$

for a non-degenerate covariance matrix $B$. Finally, letting

$$\tilde{\alpha} := \arg\min_{\alpha} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha, \boldsymbol{\beta}^* + T^{-1/2}\Delta_{\boldsymbol{\beta}})$$

we have

$$\tilde{\alpha} - \alpha^* = \mathcal{O}_{\mathbb{P}}(T^{-1/2}). \quad (28)$$

We now prove that (25)–(28) hold under Assumption 2.1. Consider the class of functions

$$\mathcal{F} := \left\{ (\mathbf{x}, Y) \mapsto \mathcal{L}(\mathbf{x}, Y; \boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \tilde{\Gamma} \right\}$$

where $\tilde{\Gamma}$ is the original parameter space if $\Gamma$ is bounded and a ball of Euclidean radius 1 around $\boldsymbol{\gamma}^*$ if $\Gamma$ is unbounded but the objective is convex. In both cases the bracketing numbers $N_{[]}(\varepsilon, \mathcal{F}, L_2(\mathbb{P}))$ are at most polynomial in $1/\varepsilon$ (see Example 19.7 in van der Vaart (2000)). Thus Corollary 19.35 in van der Vaart (2000) implies

$$\sup_{\boldsymbol{\gamma} \in \tilde{\Gamma}} \left| \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}) - m(\boldsymbol{\gamma}) \right| = \mathcal{O}_{\mathbb{P}}(T^{-1/2}).$$

Since the minimum is by assumption well-separated this implies $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* = o_{\mathbb{P}}(1)$ in the case

where $\tilde{\Gamma} = \Gamma$. In the case of convexity we find that

$$\inf_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|=1} \Big|\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*)\Big| \\ \geq \inf_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|=1} |m(\boldsymbol{\gamma}) - m(\boldsymbol{\gamma}^*)| + \mathcal{O}_{\mathbb{P}}(T^{-1/2})$$

and hence by convexity of $\boldsymbol{\gamma} \mapsto \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma})$ the minimizer of the latter must lie in $\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\| = 1$ with probability tending to one. This reduces the problem to the case of bounded parameter spaces $\Gamma$. In either case we have proved $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* = o_{\mathbb{P}}(1)$. Now (26) follows from Theorem 5.23 in van der Vaart (2000) while the expansion in (25) is established in the first line of the proof of the latter Theorem. Convergence in (27) follows from the CLT after observing that under Assumption 2.1(ii) $\mathcal{L}(\mathbf{x}_t, Y_t; \boldsymbol{\gamma}^*)$ has a finite second moment.

To establish (28), note that the proof of Theorem 5.52 in van der Vaart (2000) yields the following more general result: assume that we have a sequence of functions $m_T : \Theta \to \mathbb{R}$ and estimated functions $\hat{m}_T$ such that for any sufficiently small $\delta > 0$

(a) $\sup_{\|\theta-\theta_T\|\leq\delta} m_T(\theta) - m_T(\theta_T) \leq C\delta^2$,

(b) $\mathbb{E}\Big[\sup_{\|\theta-\theta_T\|\leq\delta} \sqrt{T}|\hat{m}_T(\theta) - \hat{m}_T(\theta_T) - m_T(\theta) + m_T(\theta_T)|\Big] \leq C\delta$,

(c) $\hat{m}_T(\hat{\theta}) = \inf_\theta \hat{m}_T(\theta) + \mathcal{O}_{\mathbb{P}}(T^{-1})$.

(d) $\hat{\theta} = \theta_T + o_{\mathbb{P}}(1)$.

Then $\hat{\theta} - \theta_T = \mathcal{O}_{\mathbb{P}}(T^{-1/2})$. We will apply this with $m_T(\theta) := \mathbb{E}[\mathcal{L}(\mathbf{x}, Y; \theta, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}})]$, $\theta_T$ the well-separated global minimizer of $m_T(\theta)$ which exists by assumption for $T$ sufficiently large, and

$$\hat{m}_T(\theta) = T^{-1}\sum_{t=1}^{T}\mathcal{L}(\mathbf{x}_t, Y_t; \theta, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}})$$

Of those conditions, (a) follows by a Taylor expansion noting that the gradient of $m_T$ vanishes at $\theta_T$ and (c) follows by assuming the computed minimizer is sufficiently close to the global minimizer. Next observe that

$$\sup_{\|\theta-\theta_T\|\leq\delta} \sqrt{T}|\hat{m}_T(\theta) - \hat{m}_T(\theta_T) - m_T(\theta) + m_T(\theta_T)| \\ \leq \sup_{\|\theta-\theta_T\|\leq\delta} \mathbb{G}_T(\mathcal{L}(\cdot; \theta, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}}) - \mathcal{L}(\cdot; \theta_T, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}}))$$

where $\mathbb{G}_T$ denotes the empirical process corresponding to the observations $(\mathbf{x}_t, Y_t)_{t=1,\ldots,T}$. The class of functions

$$\mathcal{F}_T := \Big\{(\mathbf{x}, Y) \mapsto \mathcal{L}(\mathbf{x}, Y; \theta, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}}) - \mathcal{L}(\mathbf{x}, Y; \theta_T, \boldsymbol{\beta}^* + T^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\beta}}) : |\theta - \theta_T| \leq \delta\Big\}$$

has envelope $\dot{m}(\cdot)\delta$ and bracketing numbers satisfying $N_{[]}(\varepsilon, \mathcal{F}_T, L_2(\mathbb{P})) \lesssim \frac{\delta}{\varepsilon}$ (compare Example 19.7 in van der Vaart (2000)) and thus by Corollary 19.35 in van der Vaart (2000) we have

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}_T} |\mathbb{G}_T(f)|\Big] \lesssim \int_0^{\|\dot{m}\|_{P,2}\delta} \sqrt{1+\log(\delta/\varepsilon)}d\varepsilon \lesssim \delta$$

where the last equality follows after a change of variables. This implies (b). The statement in (d) follows by similar arguments as the proof of consistency of $\hat{\boldsymbol{\gamma}}$ given earlier since we assumed that each $m_T$ has a unique and well-separated global minimizer. This completes the proof of (a)–(d) and hence (28).

From now on assume that (25)–(28) hold. We first analyze $\hat{k}^{PAM}$. Let

$$\hat{G}_\alpha = T^{-1/2}\sum_{t=1}^{T} \nabla_\alpha \mathcal{L}(\mathbf{x}_t, Y_t; \alpha, \boldsymbol{\beta}^*)\Big|_{\alpha=\alpha^*},$$
$$\hat{G}_{\boldsymbol{\beta}} = T^{-1/2}\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha^*, \boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

By (27) we have

$$(\hat{G}_\alpha, \hat{G}_{\boldsymbol{\beta}}^\top)^\top \xrightarrow{d} (G_\alpha, G_{\boldsymbol{\beta}}^\top)^\top \sim N(0, B)$$

and by (26) and (27)

$$T^{1/2}(\hat{\alpha}-\alpha^*, (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^\top)^\top = -A^{-1}(\hat{G}_\alpha, \hat{G}_{\boldsymbol{\beta}}^\top)^\top + o_{\mathbb{P}}(1) \xrightarrow{d} -A^{-1}(G_\alpha, G_{\boldsymbol{\beta}}^\top)^\top. \tag{29}$$

In what follows, for squared matrices $M$ of dimension $p+1$ consider the following block structures

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

with $M_{11} \in \mathbb{R}$. With this notation we find that

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) &\xrightarrow{d} \Big[[A^{-1}]_{21} \;\; [A^{-1}]_{22}\Big](G_\alpha, G_{\boldsymbol{\beta}}^\top)^\top = [A^{-1}]_{21}G_\alpha + [A^{-1}]_{22}G_{\boldsymbol{\beta}} \\ &= [A^{-1}]_{22}G_{\boldsymbol{\beta}} - \frac{1}{A_{11}}[A^{-1}]_{22}A_{21}G_\alpha = [A^{-1}]_{22}\Big(G_{\boldsymbol{\beta}} - \frac{A_{21}}{A_{11}}G_\alpha\Big)\end{aligned}$$

where we used block matrix inversion combined with the fact that $A_{11}$ is a scalar. Denoting by $C$ the covariance matrix of $G_{\boldsymbol{\beta}} - \frac{A_{21}}{A_{11}}G_\alpha$, we have

$$\sqrt{T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \xrightarrow{d} N(0, [A^{-1}]_{22}C[A^{-1}]_{22}).$$

In this notation, $\Sigma_{\boldsymbol{\beta}} = [A^{-1}]_{22}C[A^{-1}]_{22}$. Since $\hat{\Sigma}_{\boldsymbol{\beta}}$ is a consistent estimator for $\Sigma_{\boldsymbol{\beta}}$ by assumption and since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 = \boldsymbol{\beta}^* + T^{-1/2}\Delta$ by assumption, we obtain by the definition

of $\hat{k}^{PAM}$

$$P(\hat{k}^{PAM}=1) \to P\Big(\|Z+\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2 > \|Z\|_2\Big).$$

where $Z \sim N(0, I_{p\times p})$ with $p$ denoting he dimension of $\boldsymbol{\beta}$. This can be further simplified as follows

$$\begin{aligned}
&\|Z+\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2 > \|Z\|_2 \\
\iff &\|Z+\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2^2 > \|Z\|_2^2 \\
\iff &\|Z\|_2^2 + \|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2^2 + 2Z^\top\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta > \|Z\|_2^2 \\
\iff &\|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2^2 > -2Z^\top\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta \\
\iff &\frac{1}{2}\|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2^2 > \|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2 N(0,1).
\end{aligned}$$

Thus

$$P(\hat{k}^{PAM}=1) \to \Phi(\|\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta\|_2/2). \tag{30}$$

Next we derive the corresponding limit for $\hat{k}^{BM}$. Let

$$\begin{aligned}
\tilde{\Delta}_\alpha &:= \sqrt{T}(\tilde{\alpha}-\alpha^*) \\
\Delta_\alpha^* &:= -\frac{\hat{G}_\alpha + A_{12}\Delta}{A_{11}}.
\end{aligned}$$

Apply the expansion in (25) with $\boldsymbol{\Delta}_{\boldsymbol{\gamma}} = (\tilde{\Delta}_\alpha, \Delta)$ and with $\boldsymbol{\Delta}_{\boldsymbol{\gamma}} = (\Delta_\alpha^*, \Delta)$ and subtract those expansions to obtain

$$\begin{aligned}
0 \geq &\sum_{t=1}^T \mathcal{L}(\mathbf{x}_t, Y_t; \tilde{\alpha}, \boldsymbol{\beta}^* + T^{-1/2}\Delta) - \sum_{t=1}^T \mathcal{L}(\mathbf{x}_t, Y_t; \alpha^* + T^{-1/2}\Delta_\alpha^*, \boldsymbol{\beta}^* + T^{-1/2}\Delta) \\
=&\tilde{\Delta}_\alpha\hat{G}_\alpha + \Delta^\top\hat{G}_{\boldsymbol{\beta}} + \frac{1}{2}A_{11}\tilde{\Delta}_\alpha^2 + \frac{1}{2}\Delta^\top A_{22}\Delta + \tilde{\Delta}_\alpha A_{12}\Delta \\
&- \Big\{\Delta_\alpha^*\hat{G}_\alpha + \Delta^\top\hat{G}_{\boldsymbol{\beta}} + \frac{1}{2}A_{11}(\Delta_\alpha^*)^2 + \frac{1}{2}\Delta^\top A_{22}\Delta + \Delta_\alpha^* A_{12}\Delta\Big\} + o_{\mathbb{P}}(1) \\
=&\frac{A_{11}}{2}(\Delta_\alpha^* - \tilde{\Delta}_\alpha)^2 + o_{\mathbb{P}}(1).
\end{aligned}$$

where the inequality in the first line follows because $\tilde{\alpha}$ is defined as minimizer and the last line is obtained by plugging in the definition of $\Delta_\alpha^*$. This implies $\Delta_\alpha^* = \tilde{\Delta}_\alpha + o_{\mathbb{P}}(1)$. Next

observe that

$$\begin{aligned}
&\inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha, \boldsymbol{\beta}^* + T^{-1/2}\Delta) - \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha^*, \boldsymbol{\beta}^*) \\
&= \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \tilde{\alpha}, \boldsymbol{\beta}^* + T^{-1/2}\Delta) - \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha^*, \boldsymbol{\beta}^*) \\
&= \tilde{\Delta}_\alpha \hat{G}_\alpha + \Delta^\top \hat{G}_{\boldsymbol{\beta}} + \frac{1}{2}(\tilde{\Delta}_\alpha, \Delta^\top) A (\tilde{\Delta}_\alpha, \Delta^\top)^\top + o_{\mathbb{P}}(1) \\
&= \Delta^\top \hat{G}_{\boldsymbol{\beta}} + \frac{1}{2}\Delta^\top A_{22}\Delta - \frac{1}{2A_{11}}(\hat{G}_\alpha + A_{12}\Delta)^2 + o_{\mathbb{P}}(1)
\end{aligned}$$

where we used the definition of $\tilde{\alpha}$ and the expansion $\Delta^*_\alpha = \tilde{\Delta}_\alpha + o_{\mathbb{P}}(1)$ in the last line and (25) in the second to last line. Similarly, setting $\Delta = 0$ in the above expansion we find

$$\inf_{\alpha} \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha, \boldsymbol{\beta}^*) - \sum_{t=1}^{T} \mathcal{L}(\mathbf{x}_t, Y_t; \alpha^*, \boldsymbol{\beta}^*) = -\frac{1}{2A_{11}}\hat{G}_\alpha^2 + o_{\mathbb{P}}(1).$$

Subtracting those two expansions and expanding the square in $(\hat{G}_\alpha + A_{12}\Delta)^2$ we find

$$\begin{aligned}
\hat{k}^{BM} = 1 &\iff \Delta^\top \hat{G}_{\boldsymbol{\beta}} + \frac{1}{2}\Delta^\top A_{22}\Delta - \frac{\hat{G}_\alpha A_{12}\Delta}{A_{11}} - \frac{(A_{12}\Delta)^2}{2A_{11}} + o_{\mathbb{P}}(1) > 0 \\
&\iff \Delta^\top \Big(\hat{G}_{\boldsymbol{\beta}} - \frac{\hat{G}_\alpha A_{21}}{A_{11}}\Big) + \frac{1}{2}\Delta^\top \Big(A_{22} - \frac{A_{21}A_{12}}{A_{11}}\Big)\Delta + o_{\mathbb{P}}(1) > 0 \\
&\iff \Delta^\top \Big(\hat{G}_{\boldsymbol{\beta}} - \frac{\hat{G}_\alpha A_{21}}{A_{11}}\Big) + \frac{1}{2}\Delta^\top \Big[[A^{-1}]_{22}\Big]^{-1}\Delta + o_{\mathbb{P}}(1) > 0
\end{aligned}$$

where the last line follows by block inversion for matrices since $A_{11}$ is a scalar. Thus

$$\begin{aligned}
P(\hat{k}^{BM} = 1) &\to P\Big(\frac{1}{2}\Delta\Big[[A^{-1}]_{22}\Big]^{-1}\Delta > N(0, \Delta^\top C \Delta)\Big) \\
&= \Phi\Big(\frac{\Delta\Big[[A^{-1}]_{22}\Big]^{-1}\Delta}{2(\Delta^\top C\Delta)^{1/2}}\Big). \qquad (31)
\end{aligned}$$

To lighten notation, let $D := \Big[[A^{-1}]_{22}\Big]^{-1}$. Note that by the Cauchy-Schwarz inequality and the definition of $\Sigma_{\boldsymbol{\beta}}$

$$\frac{\Delta^\top D\Delta}{(\Delta^\top C\Delta)^{1/2}} = \frac{\Delta^\top C^{1/2}C^{-1/2}D\Delta}{(\Delta^\top C\Delta)^{1/2}} \leq \frac{\|\Delta^\top C^{1/2}\|_2 \|C^{-1/2}D\Delta\|_2}{(\Delta^\top C\Delta)^{1/2}} = (\Delta^\top \Sigma_{\boldsymbol{\beta}}^{-1}\Delta)^{1/2}.$$

This inequality is strict unless $C^{1/2}\Delta$ is a scalar multiple of $C^{-1/2}D\Delta$. Thus

$$\lim_{T\to\infty} P(\hat{k}^{PAM} = 1) \geq \lim_{T\to\infty} P(\hat{k}^{BM} = 1)$$

with strict inequality unless $C^{1/2}\Delta$ is a scalar multiple of $C^{-1/2}D\Delta$.

For a proof that

$$\lim_{T\to\infty}\mathbb{P}\Big(\hat{k}^{PAM}=1\Big)\geq\lim_{T\to\infty}\mathbb{P}\Big(\hat{k}^{PAM,K_T}=1\Big),$$

we obtain by similar computations as above that

$$\lim_{T\to\infty}\mathbb{P}\Big(\hat{k}^{PAM,K_T}=1\Big)=\Phi\Big(\frac{\Delta^\top K^\top K\Delta}{2(\Delta^\top K^\top K\Sigma_{\boldsymbol{\beta}}K^\top K\Delta)^{1/2}}\Big)=\Phi\Big(\frac{\Delta K^\top K\Sigma_{\boldsymbol{\beta}}^{1/2}\Sigma_{\boldsymbol{\beta}}^{-1/2}\Delta}{2(\Delta^\top K^\top K\Sigma_{\boldsymbol{\beta}}K^\top K\Delta)^{1/2}}\Big)$$
$$\leq\Phi\Big(\frac{(\Delta K^\top K\Sigma_{\boldsymbol{\beta}}K^\top K\Delta)^{1/2}(\Delta\Sigma_{\boldsymbol{\beta}}^{-1}\Delta)^{1/2}}{2(\Delta^\top K^\top K\Sigma_{\boldsymbol{\beta}}K^\top K\Delta)^{1/2}}\Big)=\Phi\Big(\frac{(\Delta\Sigma_{\boldsymbol{\beta}}^{-1}\Delta)^{1/2}}{2}\Big)=\lim_{T\to\infty}\mathbb{P}\Big(\hat{k}^{PAM}=1\Big).$$

□

## 9.2 Proof of the generic spectral clustering results (Theorems 3.1, 3.2)

Since the result is trivial when $G^*=1$, we will without loss of generality assume that $G^*\geq 2$. We will further write $G$ instead of $G^*$ since there is no risk of confusion in this subsection.

To simplify notation, we will without loss of generality assume that the units are ordered according to their true grouping, i.e. unit $1,...,|I_1^*|$ belong to group 1, unit $|I_1^*|+1,...,|I_2^*|$ belong to group 2, etc. This is to shorten notation only, all arguments will work with more complex notation if this assumption is dropped.

To proceed to the proof, we first consider the decomposition $\hat{A}=\hat{A}_{\text{diag}}+\hat{A}_{\text{off-diag}}$, where

$$\hat{A}_{\text{diag}}:=\begin{pmatrix}\hat{A}^{(11)}&\mathbf{0}&\cdots&\mathbf{0}\\\mathbf{0}&\hat{A}^{(22)}&\cdots&\mathbf{0}\\\cdots&\cdots&\cdots&\cdots\\\mathbf{0}&\mathbf{0}&\cdots&\hat{A}^{(GG)}\end{pmatrix}$$

and

$$\hat{A}_{\text{off-diag}}:=\begin{pmatrix}\mathbf{0}&\hat{A}^{(12)}&\cdots&\hat{A}^{(1G)}\\\hat{A}^{(21)}&\mathbf{0}&\cdots&\hat{A}^{(2G)}\\\cdots&\cdots&\cdots&\cdots\\\hat{A}^{(G1)}&\hat{A}^{(G2)}&\cdots&\mathbf{0}\end{pmatrix},$$

with $\hat{A}^{(ij)}\in\mathbb{R}^{|I_i^*|\times|I_j^*|}, i,j=1,\ldots,G$. Define the degree matrix $\hat{D}_{\text{diag}}$ corresponding to $\hat{A}_{\text{diag}}$ as

$$\hat{D}_{\text{diag}}:=\text{diag}\big((\hat{D}_{\text{diag}})_1,\ldots,(\hat{D}_{\text{diag}})_n\big),$$

with the elements $(\hat{D}_{\text{diag}})_i:=\sum_{j=1}^n(\hat{A}_{\text{diag}})_{ij}$. Define the corresponding graph Laplacian $\hat{L}_{\text{diag}}$ as

$$\hat{L}_{\text{diag}}:=I-\hat{D}_{\text{diag}}^{-1/2}\hat{A}_{\text{diag}}D_{\text{diag}}^{-1/2}.$$

The remaining proof proceeds as follows: in step 1, we show that $\hat{L}_{\text{diag}}$ has non-negative eigenvalues and that the eigenvalue zero has multiplicity $G$. Moreover, the eigenspace corresponding to that eigenvalue is spanned by the vectors $\hat{D}_{\text{diag}}\mathbb{1}_{I_j^*} \in \mathbb{R}^n$ with entries

$$(\mathbb{1}_{I_j^*})_k = \begin{cases} 1, & k \in I_j^* \\ 0, & k \notin I_j^* \end{cases}, \tag{32}$$

see Lemma 9.1. In step 2, we bound the distance in operator norm between $\hat{L}$ and $\hat{L}_{\text{diag}}$ (Lemma 9.2). In step 3, we quantify the gap between the $G$-th and $(G+1)$-th smallest eigenvalues of $\hat{L}_{\text{diag}}$ (Lemma 9.3). In step 4, we use the results from step 2 and step 3 to show that the matrix $\hat{U}$ defined in step 4 of the spectral clustering algorithm is close to a rotation of the matrix $U \in \mathbb{R}^{n \times G}$ defined via

$$U := \left(\mathbb{1}_{I_1^*}, ..., \mathbb{1}_{I_G^*}\right)$$

in Frobenius norm (Lemma 9.4), i.e. the Frobenius norm of the difference between those matrices converges to zero. This convergence together with a simple analysis of the $k$-means algorithm yields our main result in step 5.

**Step 1:** *Eigenstructure of* $\hat{L}_{\text{diag}}$.
The following result is essentially a reformulation of Proposition 4 from von Luxburg (2007) in our setting. The proof follows by exactly the same type of arguments as in the latter paper, for the sake of completeness and for the reader's convenience we provide a short proof in our specific setting.

**Lemma 9.1.** *The multiplicity of the eigenvalue 0 of* $\hat{L}_{\text{diag}}$ *equals* $G$. *The eigenspace of the eigenvalue 0 of* $\hat{L}_{\text{diag}}$ *is spanned by the vectors* $\hat{D}_{\text{diag}}^{1/2}\mathbb{1}_{I_j^*}$ *where* $\mathbb{1}_{I_j^*}$ *are defined in* (32).

*Proof of Lemma 9.1.* Begin by observing that $\hat{L}_{\text{diag}}$ is block-diagonal with $G$ blocks, say $\hat{L}^{(11)}, ..., \hat{L}^{(GG)}$, of size $|I_1^*| \times |I_1^*|, ..., |I_G^*| \times |I_G^*|$. It thus suffices to show that the eigenvalues of each block are non-negative and that the multiplicity of the eigenvalue 0 for each block equals 1. Since all blocks share a similar structure we will focus on the first block. Assume that $v = (v_1, \ldots, v_{|I_1^*|})^\top$ is an eigenvector of $\hat{L}_{\text{diag}}^{(11)}$ with norm 1 corresponding to eigenvalue

$\lambda$. Then, we have

$$\begin{aligned}
\lambda =& v^\top \hat{L}^{(11)}_{\text{diag}} v \\
=& \sum_{i \in I_1} v_i^2 - \sum_{i,j \in I_1} v_i \frac{\hat{A}_{ij}}{\sqrt{(\hat{D}_{\text{diag}})_i}\sqrt{(\hat{D}_{\text{diag}})_j}} v_j \\
=& \frac{1}{2}\left( \sum_{i \in I_1} v_i^2 - 2 \sum_{i,j \in I_1} \hat{A}_{ij} \frac{v_i}{\sqrt{(\hat{D}_{\text{diag}})_i}} \frac{v_j}{\sqrt{(\hat{D}_{\text{diag}})_j}} + \sum_{j \in I_1} v_j^2 \right) \\
=& \frac{1}{2} \sum_{i,j \in I_1} \hat{A}_{ij} \left( \frac{v_i}{\sqrt{(\hat{D}_{\text{diag}})_i}} - \frac{v_j}{\sqrt{(\hat{D}_{\text{diag}})_j}} \right)^2 \\
\geq& 0 \,,
\end{aligned}$$

where $(\hat{D}_{\text{diag}})_i$ denotes the $i$-th diagonal elements of $\hat{D}_{\text{diag}}$, and the last line follows since by construction $\hat{A}_{ij} > 0$. The latter also implies that $\lambda = 0$ if and only if $v_i / \sqrt{(\hat{D}_{\text{diag}})_i} = v_j / \sqrt{(\hat{D}_{\text{diag}})_j}$ for all $i, j$, which is only possible if $v_i = C\sqrt{(\hat{D}_{\text{diag}})_i}$ for a constant $C$ independent of $i$. This completes the proof.

□

**Step 2:** *Bound on operator norm distance between $\hat{L}$ and $\hat{L}_{\text{diag}}$.*

Now we consider the distance between $\hat{L}$ and $\hat{L}_{\text{diag}}$ in operator norm.

**Lemma 9.2.** *On the event $\frac{nA_{1,max}}{A_{0,min} \min_k |I_k^*|} \leq 1$ it holds that*

$$\left| \left| \left| \hat{L} - \hat{L}_{\text{diag}} \right| \right|_\infty \leq \frac{4\sqrt{2} n A_{1,max}}{A_{0,min} \min_k |I_k^*|} \sqrt{\frac{A_{0,max} \max_k |I_k^*|}{A_{0,min} \min_k |I_k^*|}} \,.$$

*Proof of Lemma 9.2.* The proof follows a similar strategy as in Chung and Radcliffe (2011), and Lemma 3.1 of van Delft and Dette (2021) but modified to account for the fact that $n$ can diverge while it is fixed in the latter paper. Decompose the difference $\hat{L} - \hat{L}_{\text{diag}}$ as follows

$$\begin{aligned}
& \hat{L} - \hat{L}_{\text{diag}} \\
=& (\hat{D}^{-1/2} - \hat{D}_{\text{diag}}^{-1/2}) \hat{A} \hat{D}^{-1/2} + \hat{D}_{\text{diag}}^{-1/2} \hat{A} (\hat{D}^{-1/2} - \hat{D}_{\text{diag}}^{-1/2}) + \hat{D}_{\text{diag}}^{-1/2} (\hat{A} - \hat{A}_{\text{diag}}) \hat{D}_{\text{diag}}^{-1/2} \\
=& (I - \hat{D}_{\text{diag}}^{-1/2} \hat{D}^{1/2}) \hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2} + (\hat{D}_{\text{diag}}^{-1/2} \hat{D}^{1/2}) \hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2} (I - \hat{D}^{1/2} \hat{D}_{\text{diag}}^{-1/2}) \\
& + \hat{D}_{\text{diag}}^{-1/2} (\hat{A} - \hat{A}_{\text{diag}}) \hat{D}_{\text{diag}}^{-1/2} \,.
\end{aligned}$$

Now, we bound the terms on the right hand side separately. Define the $i$-th diagonal

elements of the diagonal matrix $D$ by $(D)_i$. By definition of the norm $|||\cdot|||_\infty$, we have

$$\begin{aligned}\left|\left|\left|I-\hat{D}_{\text{diag}}^{-1/2}\hat{D}^{1/2}\right|\right|\right|_\infty &= \max_i\left|1-\sqrt{\frac{\hat{D}_i}{(\hat{D}_{\text{diag}})_i}}\right| \\ &\le \max_i\left|1-\frac{\hat{D}_i}{(\hat{D}_{\text{diag}})_i}\right| \\ &\le \frac{\max_i\left|(\hat{D}_{\text{diag}})_i-\hat{D}_i\right|}{\min_i(\hat{D}_{\text{diag}})_i},\end{aligned}$$

where we used the fact that $|1-x|=|1-\sqrt{x}||1+\sqrt{x}|\ge|1-\sqrt{x}|, \forall x>0$. We also have

$$\left|\left|\left|\hat{D}_{\text{diag}}^{-1/2}\hat{D}^{-1/2}\right|\right|\right|_\infty = \left|\left|\left|I-(I-\hat{D}_{\text{diag}}^{-1/2}\hat{D}^{-1/2})\right|\right|\right|_\infty \le 1+\frac{\max_i\left|(\hat{D}_{\text{diag}})_i-\hat{D}_i\right|}{\min_i(\hat{D}_{\text{diag}})_i},$$

and

$$\begin{aligned}\left|\left|\left|\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}\right|\right|\right|_\infty &= \max_i\left\{\sum_{j=1}^n\frac{\hat{A}_{ij}}{\sqrt{\hat{D}_i}\sqrt{\hat{D}_j}}\right\} \\ &\le \max_i\left\{\frac{1}{\sqrt{\hat{D}_i}}\frac{1}{\min_k\sqrt{\hat{D}_k}}\sum_{j=1}^n\hat{A}_{ij}\right\} \\ &= \frac{\max_i\sqrt{\hat{D}_i}}{\min_j\sqrt{\hat{D}_j}}.\end{aligned}$$

Moreover, by the sub-multiplicativity of the norm $|||\cdot|||_\infty$, it holds that

$$\left|\left|\left|\hat{D}_{\text{diag}}^{-1/2}(\hat{A}-\hat{A}_{\text{diag}})\hat{D}_{\text{diag}}^{-1/2}\right|\right|\right|_\infty \le \frac{1}{\min_i(\hat{D}_{\text{diag}})_i}\left|\left|\left|\hat{A}-\hat{A}_{\text{diag}}\right|\right|\right|_\infty.$$

Collecting pieces gives

$$\begin{aligned}&\left|\left|\left|\hat{L}-\hat{L}_{\text{diag}}\right|\right|\right|_\infty \\ \le&\frac{\max_i\left|(\hat{D}_{\text{diag}})_i-\hat{D}_i\right|}{\min_i(\hat{D}_{\text{diag}})_i}\frac{\max_i\sqrt{\hat{D}_i}}{\min_j\sqrt{\hat{D}_j}}\left(2+\frac{\max_i\left|(\hat{D}_{\text{diag}})_i-\hat{D}_i\right|}{\min_i(\hat{D}_{\text{diag}})_i}\right)+\frac{1}{\min_i(\hat{D}_{\text{diag}})_i}\left|\left|\left|\hat{A}_{\text{diag}}-\hat{A}\right|\right|\right|_\infty.\end{aligned}$$

Define

$$\begin{aligned}S_{0,i} &:= \sum_{j:i,j \text{ in same group}}\hat{A}_{ij}, \\ S_{1,i} &:= \sum_{j:i,j \text{ in different groups}}\hat{A}_{ij}.\end{aligned}$$

With this notation we have

$$\left|(\hat{D}_{\text{diag}})_i - \hat{D}_i\right| = S_{0,i}$$

and

$$(\hat{D}_{\text{diag}})_i = S_{1,i}$$

as well as $\hat{D}_i = S_{0,i} + S_{1,i}$. Moreover, by definition,

$$\left|\!\left|\!\left|\hat{A}_{\text{diag}} - \hat{A}\right|\!\right|\!\right|_\infty = \max_i S_{1,i}.$$

Collecting pieces yields

$$\left|\!\left|\!\left|\hat{L} - \hat{L}_{\text{diag}}\right|\!\right|\!\right|_\infty \leq \frac{\max_i S_{1,i}}{\min_i S_{0,i}} \left(1 + \sqrt{\frac{\max_i S_{0,i} + S_{1,i}}{\min_i S_{0,i} + S_{1,i}}} \Big(2 + \frac{\max_i S_{1,i}}{\min_i S_{0,i}}\Big)\right)$$

Recall the definition of $A_{1,max}, A_{0,\min}$. We have

$$S_{1,i} \leq n A_{1,max}\,,$$

and

$$A_{0,max} \max_k |I_k^*| \geq S_{0,i} \geq \min_k |I_k^*| A_{0,min}\,. \tag{33}$$

This further yields

$$\left|\!\left|\!\left|\hat{L} - \hat{L}_{\text{diag}}\right|\!\right|\!\right|_\infty \leq \frac{nA_{1,max}}{A_{0,min} \min_k |I_k^*|} \left(1 + \sqrt{\frac{\max_k |I_k^*| + nA_{1,max}}{A_{0,min} \min_k |I_k^*|}} \Big(2 + \frac{nA_{1,max}}{A_{0,min} \min_k |I_k^*|}\Big)\right).$$

Assuming $\frac{nA_{1,max}}{A_{0,min} \min_k |I_k^*|} \leq 1$ and noting $\frac{\max_k |I_k^*|}{A_{0,min} \min_k |I_k^*|} \geq 1$ this can be further bounded by

$$\left|\!\left|\!\left|\hat{L} - \hat{L}_{\text{diag}}\right|\!\right|\!\right|_\infty \leq \frac{4\sqrt{2} n A_{1,max}}{A_{0,min} \min_k |I_k^*|} \sqrt{\frac{\max_k |I_k^*|}{A_{0,min} \min_k |I_k^*|}}.$$

This completes the proof. □

**Step 3:** *Bounding the $G+1$'st smallest eigenvalue of $\hat{L}_{\text{diag}}$.*
Denote the $i$-th smallest eigenvalue of $\hat{L}_{\text{diag}}$ by $\lambda_i$. By Lemma 9.1, we know that $\lambda_1 = \cdots = \lambda_G = 0$. Thus, we need to find a lower bound on the $G+1$'st smallest eigenvalue $\lambda_{G+1}$. This is done in the following Lemma.

**Lemma 9.3.** *We have*

$$\lambda_{G+1} \geq \frac{A_{0,min}}{8 A_{0,max}}\,.$$

*Proof of Lemma 9.3.* Recall the *Cheeger constant* (see for instance equation (2.2) in Chung and Graham (1997)) of a undirected graph $\mathcal{G} = (V, E)$ ($V$ denotes the set of vertices and

$E$ denotes the set of edges) with weights $w_{i,j}$ on the vertices $(i,j) \in E$:

$$\mathfrak{H} := \min_{\mathcal{I} \subset V} \frac{\sum_{j \in \mathcal{I}, k \notin \mathcal{I}} w_{j,k}}{\min\left\{ \sum_{j \in \mathcal{I}} d_j, \sum_{k \in V \setminus \mathcal{I}} d_k \right\}},$$

where

$$d_k := \sum_{(i,j) \in E : i \in \mathcal{I}} w_{i,j}\,.$$

Then, Theorem 2.2 in Chung and Graham (1997) implies that the eigengap of the normalized graph Laplacian is bounded below by $\mathfrak{H}^2/2$. To translate this result to our setting consider the fully connected graph with vertices given by $V = I_k^*$ and edge weights $w_{i,j} := \hat{A}_{ij}, i, j \in V$. Hence, the Cheeger constant corresponding to block $\hat{L}^{(mm)}$ on the diagonal of $\hat{L}_{\text{diag}}$ is defined as

$$\mathfrak{H}_m := \min_{\mathcal{I} \subset I_m^*} \frac{\sum_{j \in \mathcal{I}, i \in I_m^* \setminus \mathcal{I}} \hat{A}_{ij}}{\min\{\sum_{j \in \mathcal{I}} \hat{d}_j(m), \sum_{k \in I_m^* \setminus \mathcal{I}} \hat{d}_k(m)\}},$$

where $\hat{d}_j(m) := \sum_{i \in I_m} \hat{A}_{ij}$. Since the non-zero eigenvalues of $\hat{L}_{\text{diag}}$ are exactly the eigenvalues of the corresponding block diagonal pieces, it follows that

$$\lambda_{G+1} \geq \frac{\min_{m=1,...,G} \mathfrak{H}_m^2}{2}\,.$$

Hence, it suffices to prove that

$$\min_{m=1,...,G} \mathfrak{H}_m \geq A_{0,min}/2A_{0,max}\,.$$

Observe that

$$\sum_{j \in \mathcal{I}} \hat{d}_j(m) \leq |\mathcal{I}||I_m^*|A_{0,max}$$

and

$$\sum_{j \in \mathcal{I}, i \in I_m^* \setminus \mathcal{I}} \hat{A}_{ij} \geq |\mathcal{I}||I_m^* \setminus \mathcal{I}|A_{0,min}\,.$$

Let $\bar{\mathcal{I}} := I_m^* \setminus \mathcal{I}$. It then holds that

$$\mathfrak{H}_m \geq \frac{A_{0,min}}{A_{0,max}} \min_{\mathcal{I} \subset I_m^*} \frac{|\mathcal{I}||\bar{\mathcal{I}}|}{|I_m^*| \min\{|\mathcal{I}|, |\bar{\mathcal{I}}|\}} = \frac{A_{0,min}}{A_{0,max}} \min_{\mathcal{I} \subset I_m^*} \frac{|\mathcal{I}| \vee |\bar{\mathcal{I}}|}{|I_m^*|} \geq \frac{A_{0,min}}{2A_{0,max}}, \quad m = 1, ..., G\,.$$

This completes the proof. □

**Step 4:** *Frobenius norm convergence of $\hat{U}$ to a transformation of $U$.*

**Lemma 9.4.** *There exists a orthogonal matrix* $O_{n,T} \in \mathbb{R}^{G\times G}$ *such that on the event* $\frac{nA_{1,max}}{A_{0,min}\min_k |I_k^*|} \leq 1$ *we have*

$$\left|\left|\left|\hat{U} - UO_{n,T}\right|\right|\right|_{\mathrm{F}}^2 \leq \frac{2^{16} n^2 G \max_k |I_k^*|^2 A_{1,max}^2 A_{0,max}^3}{A_{0,min}^5 \min_k |I_k^*|^3}\,.$$

*Proof of Lemma 9.4.* In the first step we apply Theorem 2 from Yu et al. (2015). In the notation of the latter paper let $d = G, s = n, r = n - G + 1$, $\hat{\Sigma} = \hat{L}, \Sigma = \hat{L}_{\mathrm{diag}}$. Let $\hat{Z}, Z$ denote the matrices which contain the eigenvectors corresponding to the $G$ smallest eigenvalues of $\hat{L}_{\mathrm{diag}}$ and $\hat{L}$, respectively (in the notation of Yu et al. (2015) we have $\hat{V} = \hat{Z}, V = Z$). Note that by Lemma 9.1 we can choose $Z$ to have columns $\hat{D}_{\mathrm{diag}} \mathbb{1}_{I_j^*}, j = 1, ..., G$. By equation (3) in Theorem 2 from Yu et al. (2015) there exists an orthonormal matrix $\hat{O} \in \mathbb{R}^{G\times G}$ with

$$\left|\left|\left|\hat{Z}\hat{O} - Z\right|\right|\right|_{\mathrm{F}} \leq \frac{2^{3/2}\sqrt{G}\left|\left|\left|\hat{L} - \hat{L}_{\mathrm{diag}}\right|\right|\right|_{\infty}}{\lambda_{G+1}}\,. \tag{34}$$

Here we note that for symmetric matrices the operator norm $\|\cdot\|_{\mathrm{op}}$ used in Yu et al. (2015) coincides with our $|||\cdot|||_2$ and the latter satisfies $|||A|||_2 \leq |||A|||_\infty$ for symmetric matrices $A$. Let $O_{n,T} := \hat{O}^\top$ and note that by orthogonality of $\hat{O}$ we have $\left|\left|\left|\hat{Z}\hat{O} - Z\right|\right|\right|_{\mathrm{F}} = \left|\left|\left|\hat{Z} - ZO_{n,T}\right|\right|\right|_{\mathrm{F}}$. In what follows write $O$ for $O_{n,T}$ to simplify notation. Note that $\hat{U}_{i,\cdot} = \frac{\hat{Z}_{i,\cdot}}{\left\|\hat{Z}_{i,\cdot}\right\|_2}$, and $(UO)_{i,\cdot} = \frac{(ZO)_{i,\cdot}}{\|Z_{i,\cdot}\|_2}$. Similarly to Lemma 3.2 in van Delft and Dette (2021), it follows that

$$\begin{aligned}
\left|\left|\left|\hat{U} - UO\right|\right|\right|_{\mathrm{F}}^2 &= \sum_{i=1}^n \left\| \frac{\hat{Z}_{i,\cdot}}{\left\|\hat{Z}_{i,\cdot}\right\|_2} - \frac{(ZO)_{i,\cdot}}{\|Z_{i,\cdot}\|_2} \right\|_2^2 \\
&= \sum_{i=1}^n \left\| \frac{\hat{Z}_{i,\cdot}\,\|Z_{i,\cdot}\|_2 - \hat{Z}_{i,\cdot}\left\|\hat{Z}_{i,\cdot}\right\|_2 + \hat{Z}_{i,\cdot}\left\|\hat{Z}_{i,\cdot}\right\|_2 - (ZO)_{i,\cdot}\left\|\hat{Z}_{i,\cdot}\right\|_2}{\left\|\hat{Z}_{i,\cdot}\right\|_2 \|Z_{i,\cdot}\|_2} \right\|_2^2 \\
&\leq 2\sum_{i=1}^n \left\| \frac{\hat{Z}_{i,\cdot}(\,\|Z_{i,\cdot}\|_2 - \left\|\hat{Z}_{i,\cdot}\right\|_2)}{\left\|\hat{Z}_{i,\cdot}\right\|_2 \|Z_{i,\cdot}\|_2} \right\|_2^2 + \left\| \frac{\hat{Z}_{i,\cdot} - (ZO)_{i,\cdot}}{\|Z_{i,\cdot}\|_2} \right\|_2^2 \\
&= 2\sum_{i=1}^n \frac{(\,\|Z_{i,\cdot}\|_2 - \left\|\hat{Z}_{i,\cdot}\right\|_2)^2}{\|Z_{i,\cdot}\|_2^2} + \frac{\left\|\hat{Z}_{i,\cdot} - (ZO)_{i,\cdot}\right\|_2^2}{\|Z_{i,\cdot}\|_2^2} \\
&\leq 4\sum_{i=1}^n \frac{\left\|\hat{Z}_{i,\cdot} - (ZO)_{i,\cdot}\right\|_2^2}{\|Z_{i,\cdot}\|_2^2} \\
&\leq \frac{4}{\min_i \|Z_{i,\cdot}\|_2^2} \left|\left|\left|\hat{Z} - (ZO)\right|\right|\right|_{\mathrm{F}}^2.
\end{aligned}$$

Combining this with (34) yields

$$\left|\left|\left|\hat{U} - UO\right|\right|\right|_{\mathrm{F}}^2 \leq \frac{32G}{(\lambda_{G+1})^2 \min_i \|Z_{i,\cdot}\|_2^2} \left|\left|\left|\hat{L} - \hat{L}_{\mathrm{diag}}\right|\right|\right|_{\infty}^2 . \tag{35}$$

Recalling the definition of $Z$, we obtain

$$\|Z_{i,\cdot}\|_2^2 = \frac{(\hat{D}_{\mathrm{diag}})_i}{\sum_{j \in I_k^*} (\hat{D}_{\mathrm{diag}})_j} = 1/|I_k^*|, \quad \forall i \in I_k^* ,$$

where the last line follows since $(\hat{D}_{\mathrm{diag}})_i$ is the same for all $i$ from the same group. This yields

$$1/\min_i \|Z_{i,\cdot}\|_2^2 = \max_k |I_k^*|$$

and thus

$$\left|\left|\left|\hat{U} - UO\right|\right|\right|_{\mathrm{F}}^2 \leq \frac{32G \max_k |I_k^*|}{(\lambda_{G+1})^2} \left|\left|\left|\hat{L} - \hat{L}_{\mathrm{diag}}\right|\right|\right|_{\infty}^2 .$$

Combining this with the bounds in Lemma 9.2 we find

$$\left|\left|\left|\hat{U} - UO\right|\right|\right|_{\mathrm{F}}^2 \leq \frac{2^{16} n^2 G \max_k |I_k^*|^2 A_{1,max}^2 A_{0,max}^3}{A_{0,min}^5 \min_k |I_k^*|^3}$$

□

**Step 5:** *Completing the argument*

Recall that the last step of the algorithm consists of applying $k$-means clustering to the $n$ embedded points $\hat{U}_{1,\cdot}, \ldots, \hat{U}_{n,\cdot}$. In other words, this step determines group centers $\hat{c}_1, ..., \hat{c}_G$ through

$$\{\hat{c}_1, \ldots, \hat{c}_G\} \in \operatorname*{arg\,min}_{c_1,\ldots,c_G \in \mathbb{R}^G} \left\{ \sum_{i=1}^{n} \min_{j \in \{1,\ldots,G\}} \left\| \hat{U}_{i,\cdot} - c_j \right\|_2^2 \right\} .$$

The data points $\hat{U}_{i,\cdot}$ and $\hat{U}_{j,\cdot}$ are grouped together if and only if

$$\arg\min_k \|\hat{c}_k - \hat{U}_{i,\cdot}\|_2 = \arg\min_k \|\hat{c}_k - \hat{U}_{j,\cdot}\|_2 .$$

We will prove that as soon as $\left|\left|\left|\hat{U} - UO_{n,T}\right|\right|\right|_{\mathrm{F}} < 1/2$ all individuals are clustered correctly. Combined with Lemma 9.4 and noting that under the assumptions of the theorem we have $\frac{nA_{1,max}}{A_{0,min} \min_k |I_k^*|} \leq 1$, this will complete the proof. By orthogonality of $O_{n,T}$ and the definition of $U$ we have for $i, j$ in different groups

$$\|(UO_{n,T})_{i,\cdot} - (UO_{n,T})_{j,\cdot}\|_2 = \|U_{i,\cdot} - U_{j,\cdot}\|_2 = \sqrt{2} .$$

Note that by definition of the Frobenius norm

$$\max_{i\neq j}\Big\{\|\hat{U}_{i,\cdot}-(UO_{n,T})_{i,\cdot}\|_2+\|\hat{U}_{j,\cdot}-(UO_{n,T})_{j,\cdot}\|_2\Big\}\leq\sqrt{2}\left|\!\left|\!\left|\hat{U}-UO_{n,T}\right|\!\right|\!\right|_{\mathrm{F}}<1/\sqrt{2}\,.$$

Combining the above inequality with the reverse triangle inequality we have for $i,j$ in different groups

$$\min_{i,j\text{ in different groups}}\|\hat{U}_{i,\cdot}-\hat{U}_{j,\cdot}\|_2\geq\|U_{i,\cdot}-U_{j,\cdot}\|_2-\sqrt{2}\left|\!\left|\!\left|\hat{U}-UO_{n,T}\right|\!\right|\!\right|_{\mathrm{F}}>1/\sqrt{2}\,.$$

Similarly, we have

$$\max_{i,j\text{ in the same group}}\|\hat{U}_{i,\cdot}-\hat{U}_{j,\cdot}\|_2\leq\sqrt{2}\left|\!\left|\!\left|\hat{U}-UO_{n,T}\right|\!\right|\!\right|_{\mathrm{F}}<1/\sqrt{2}\,.$$

Hence any two points in the same group are closer to each other than to any point outside of that group. This implies that $\hat{c}_j$ are just the group means of group $I_j^*$ (modulo permutation of group labels) and that individuals $i,j$ are grouped together if and only if $i,j\in I_k^*$ for some $k$. This completes step 5 and thus the proof of Theorem 3.1. □

### 9.2.1 Proof of Theorem 3.2

Note that $G^*\leq n, \min_k|I_k^*|\geq 1, \max_k|I_k^*|\leq n$. Thus the bound in (6) holds with probability approaching one if $A_{1,max}^2A_{0,max}^3/A_{0,min}^5=o_{\mathbb{P}}(n^{-5})$. Now letting $\hat{\eta}_{max}:=\lambda_{max}(b_T^{-1/2}\hat{\Sigma}_{i,j}^{-1/2})$ and $\hat{\eta}_{min}:=\lambda_{min}(b_T^{-1/2}\hat{\Sigma}_{i,j}^{-1/2})$ we find that under Assumption 3.2 $\hat{\eta}_{min}$ is bounded away from zero and $\hat{\eta}_{max}$ is bounded from above by a fixed constant, both with probability tending to one. Moreover, by definition of $\hat{A}_{ij}$, that $A_{0,max}\leq 1$ and

$$\begin{aligned}A_{1,max}&\leq\exp(-b_T^{1/2}\hat{\eta}_{min}(\min_{k\neq\ell}\|\boldsymbol{\beta}_k^*-\boldsymbol{\beta}_\ell^*\|_2-2\max_i\|\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\|_2))\\A_{0,min}&\geq\exp(-2b_T^{1/2}\hat{\eta}_{max}\max_i\|\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\|_2).\end{aligned}$$

Thus

$$A_{1,max}^2A_{0,max}^3/A_{0,min}^5\leq\exp\Big(-2b_T^{1/2}\{\hat{\eta}_{min}\Delta_{min}-(2\hat{\eta}_{min}+5\hat{\eta}_{max})a_{n,T}\}\Big)\,.$$

The assumption $a_{n,T}=o_{\mathbb{P}}(\Delta_{min})$ ensures that the exponent is bounded from below by a (positive) constant multiple of $\Delta_{min}b_T^{1/2}$ which grows faster than $\log n$ by assumption. This completes the proof. □

## 9.3 Proofs for examples section

Throughout this section, we will use the following empirical process notation: let $\mathbb{P}_{T,i}$ denote the empirical measure of the sample $(\mathbf{z}_{it},Y_{it})_{t=1,\ldots,T}$ and let $\mathbb{P}_i$ denote the measure corresponding to the distribution of $(\mathbf{z}_{i1},Y_{i1})$ and let $\mathbb{G}_{T,i}:=\sqrt{T}(\mathbb{P}_{T,i}-\mathbb{P}_i)$ denote the

corresponding empirical process. For a function $f : (\mathbf{z}, y) \mapsto f(\mathbf{z}, y)$ and a signed measure $\mathbb{P}$ let $\mathbb{P}f$ stand for $\int f d\mathbb{P}$. For a class of functions $\mathcal{G}$ define $\|\mathbb{G}_{T,i}\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |\mathbb{G}_{T,i} f|$. Given the function $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, define $\sigma_{q,i}(f) := \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{q}} \sum_{i=1}^{q} f(\mathbf{z}_{it}, Y_{it})\right)$.

### 9.3.1 Proofs for logistic regression in the independent case (Theorem 3.3)

Throughout this section we will use the following additional notation. Let

$$f(y; \boldsymbol{\gamma}, \mathbf{z}) = \exp\Big\{y\mathbf{z}^\top\boldsymbol{\gamma} - g(\mathbf{z}^\top\boldsymbol{\gamma})\Big\}$$

denote the pmf of $y \in \mathbb{R}$ conditional on $\mathbf{z} \in \mathbb{R}^{p+1}$; here the function $g$ is defined via

$$\begin{aligned} g : \mathbb{R} &\to \mathbb{R} \\ z &\mapsto \log(1 + e^z)\,. \end{aligned}$$

We abbreviate the corresponding log-likelihood as $\ell(\mathbf{z}, y; \boldsymbol{\gamma}) := y\mathbf{z}^\top\boldsymbol{\gamma} - g(\mathbf{z}^\top\boldsymbol{\gamma})$. Define

$$\mathbb{M}_{i,T}(\boldsymbol{\gamma}) := \frac{1}{T}\sum_t [Y_{it}\mathbf{z}_{it}^\top\boldsymbol{\gamma} - g(\mathbf{z}_{it}^\top\boldsymbol{\gamma})]$$

and

$$\mathbb{M}_i(\boldsymbol{\gamma}) := \mathbb{E}[\mathbb{M}_{i,T}(\boldsymbol{\gamma})], i = 1, \ldots, n\,.$$

**Lemma 9.5.** *Given $p \in \mathbb{Z}^+$, we have $\int_0^1 \sqrt{1 + \log(\epsilon^{-p})}d\epsilon \le 1 + \sqrt{2\pi p}e^{1/p}$.*

*Proof of Lemma 9.5.* Set $t = \sqrt{1 + \log(\epsilon^{-p})}$, we then have

$$\begin{aligned} &\int_0^1 \sqrt{1 + \log(\epsilon^{-p})}d\epsilon \\ \le &\frac{2}{p}e^{\frac{1}{p}} \int_1^\infty t^2 e^{\frac{-t^2}{p}} dt \\ = &-e^{\frac{1}{p}} \int_1^\infty t d(e^{-\frac{t^2}{p}}) \\ = &-e^{\frac{1}{p}} \Big(te^{-\frac{t^2}{p}}\Big|_1^\infty - \int_1^\infty e^{-\frac{t^2}{p}} dt\Big) \\ = &1 + e^{\frac{1}{p}} \int_1^\infty e^{-\frac{t^2}{p}} dt \\ \le &1 + \sqrt{2\pi p}e^{1/p}\,. \end{aligned}$$

□

*Proof of Theorem 3.3 (i).* Define set $\Gamma_i(\delta) := \{\boldsymbol{\gamma} \in \mathbb{R}^{p+1} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 \le \delta\}$. By the concavity of function $\mathbb{M}_{i,T}$ and definition of $\hat{\boldsymbol{\gamma}}_i$, when all the directional derivatives on the

boundary of the set $\Gamma_i(\delta)$ is negative, that is,

$$\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\delta} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) < 0\,,$$

it follows that $\hat{\boldsymbol{\gamma}}_i \in \Gamma_i(\delta)$.

This implies

$$\mathbb{P}\bigg(\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) < 0\bigg) \le \mathbb{P}\bigg(\sup_i \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 \le \tilde{C}\sqrt{\frac{\log n}{T}}\bigg)\,.$$

Hence it suffices to show that under the stated assumptions it holds that

$$\mathbb{P}\bigg(\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) < 0\bigg) \to 1 \tag{36}$$

provided that $\tilde{C}$ is picked sufficiently large. Note that

$$(\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) = (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \Big(\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\Big) + (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma})\,. \tag{37}$$

We now handle the last two terms on the right hand side of the last equality separately. More precisely, we will show that for any $\tilde{C} > 0$ there exists $\delta > 0$ such that for $\log n/T < \delta$ we have

$$\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma}) \le -C_1\frac{\log n}{T}, \quad i=1,\dots,n\,, \tag{38}$$

where $C_1 = \tilde{C}^2\kappa_2 \inf_i\{\lambda_{\min}(\mathbb{E}[\mathbf{z}_{it}\mathbf{z}_{it}^\top])\}$ with $\kappa_1 := \max_i\{\|\boldsymbol{\gamma}_i^*\|_2\}$ and $\kappa_2 := \frac{e^{\kappa(1+\kappa_1)}}{(1+e^{\kappa(1+\kappa_1)})^2}$.

Additionally, we will prove that for $\tilde{C}$ sufficiently large (where "sufficiently large" does not depend on $n, T$), it holds that

$$\mathbb{P}\bigg(\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \Big(\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\Big) > \frac{C_1}{2}\frac{\log n}{T}\bigg) \to 0\,. \tag{39}$$

Combining the above statements with the decomposition in (37) yields (36).

**Proof of display** (38). In what follows assume that the vector $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ satisfies $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 = \tilde{C}\sqrt{\frac{\log n}{T}}$. Using Taylor expansion, we have

$$\nabla \mathbb{M}_i(\boldsymbol{\gamma}) = \nabla \mathbb{M}_i(\boldsymbol{\gamma}_i^*) + \nabla^2 \mathbb{M}_i(\tilde{\boldsymbol{\gamma}})(\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)\,,$$

where $\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{p+1}$ is on the line connecting $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_i^*$. Note that $\nabla \mathbb{M}_i(\boldsymbol{\gamma}_i^*) = \mathbf{0}$. Multiplying

both sides of the equation by $(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)$ gives

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma}) = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \nabla^2 \mathbb{M}_i(\tilde{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)\,.$$

Note that $\nabla^2 \mathbb{M}_i(\tilde{\boldsymbol{\gamma}}) = -\mathbb{E}\big[g^{''}(\mathbf{z}_{it}^\top \tilde{\boldsymbol{\gamma}})\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]$. It then follows that

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma}) = -\,(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \mathbb{E}\big[g^{''}(\mathbf{z}_{it}^\top \tilde{\boldsymbol{\gamma}})\mathbf{z}_{it}\mathbf{z}_{it}^\top\big](\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)\,.$$

This implies

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma}) \leq -\,\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2^2\, \lambda_{\min}\Big(\mathbb{E}\big[g^{''}(\mathbf{z}_{it}^\top \tilde{\boldsymbol{\gamma}})\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\Big)\,. \tag{40}$$

For any $\tilde{C}$ we have $\tilde{C}\sqrt{\frac{\log n}{T}} < 1$ provided that $\log n/T$ is small enough. Note that $\tilde{\boldsymbol{\gamma}}$ is in between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_i^*$ and $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 = \tilde{C}\sqrt{\frac{\log n}{T}}$. We then find $\|\tilde{\boldsymbol{\gamma}}\|_2 \leq 1 + \kappa_1$. Combining this with Assumption 3.4 yields

$$\Big\|\mathbf{z}_{it}^\top \tilde{\boldsymbol{\gamma}}\Big\|_2 \leq \sup_{i,t} \|\mathbf{z}_{it}\|_2\, \|\tilde{\boldsymbol{\gamma}}\|_2 \leq \kappa(1 + \kappa_1)\,.$$

Note that the function $z \mapsto g^{''}(z) = \frac{e^z}{(1+e^z)^2}$ is positive and decreasing on $\mathbb{R}_+$. It then follows that

$$g^{''}(\mathbf{z}_{it}^\top \tilde{\boldsymbol{\gamma}}) \geq \frac{e^{\kappa(1+\kappa_1)}}{(1 + e^{\kappa(1+\kappa_1)})^2}\,.$$

Define $\kappa_2 := \frac{e^{\kappa(1+\kappa_1)}}{(1+e^{\kappa(1+\kappa_1)})^2}$. Plugging the last display into the inequality (40) and using $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 = \tilde{C}\sqrt{\frac{\log n}{T}}$ yields

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \nabla \mathbb{M}_i(\boldsymbol{\gamma}) \leq -\kappa_2 \tilde{C}^2 \frac{\log n}{T} \lambda_{\min}\Big(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\Big)\,.$$

By Assumption 3.3 that $\inf_i\{\lambda_{\min}\big(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\big)\}$ is bounded from zero, we obtain (38) by setting $C_1 := \kappa_2 \tilde{C}^2 \inf_i\{\lambda_{\min}\big(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\big)\}$.

**Proof of display** (39) We will show

$$\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 = \tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \Big(\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\Big) > \frac{C_1}{2}\frac{\log n}{T}\right) \to 0\,.$$

By Cauchy-Schwaz inequality, it holds that

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*)^\top \Big(\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\Big) \leq \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2\, \|\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\|_2\,,$$

which implies

$$\begin{aligned}
&\mathbb{P}\Bigg( \sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \Big( \nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma}) \Big) > \frac{C_1}{2}\frac{\log n}{T} \Bigg) \\
&\leq \mathbb{P}\Bigg( \sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 \, \|\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\|_2 > \frac{C_1}{2}\frac{\log n}{T} \Bigg) \\
&= \mathbb{P}\Bigg( \sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} \|\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\|_2 > \frac{C_1}{2\tilde{C}}\sqrt{\frac{\log n}{T}} \Bigg) \\
&\leq \mathbb{P}\Bigg( \sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2\leq 1} \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 \, \|\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\|_2 > C_2\sqrt{\frac{\log n}{T}} \Bigg),
\end{aligned}$$

where $C_2 := \frac{C_1}{2\tilde{C}} = \tilde{C}\kappa_2 \inf_i\{\lambda_{\min}(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big])\}/2$. We will show that the last line in the display above converges to zero provided that $\tilde{C}$ is large enough. Define the vector

$$M_{it}(\boldsymbol{\gamma}) := Y_{it}\mathbf{z}_{it} - \frac{\mathbf{z}_{it}e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}} - \mathbb{E}\left[ Y_{it}\mathbf{z}_{it} - \frac{\mathbf{z}_{it}e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}} \right].$$

Denote the $j$-th entry of the vector $M_{it}(\boldsymbol{\gamma})$ by $M_{it,j}(\boldsymbol{\gamma})$. We now show that

$$\mathbb{P}\Bigg( \sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2\leq 1} \frac{1}{T}\sum_{t=1}^T \Big| M_{it,j}(\boldsymbol{\gamma}) \Big| > C_2\sqrt{\frac{\log n}{T}} \Bigg) \to 0\,,$$

Define the function $h_{\boldsymbol{\gamma}}^j(\mathbf{z},y)$ via

$$\begin{aligned}
&h_{\boldsymbol{\gamma}}^j : \mathbb{R}^{p+1}\times\{0,1\} \to \mathbb{R} \\
&\qquad\qquad (\mathbf{z},y) \mapsto h_{\boldsymbol{\gamma}}^j(\mathbf{z},y) := z_j\Big( y - \frac{e^{\mathbf{z}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}^\top\boldsymbol{\gamma}}} \Big) \mathbb{1}\{\|\mathbf{z}\|_2 \leq \kappa\}\,,
\end{aligned}$$

where $z_j$ denotes the $j$-th element of the vector $\mathbf{z}$. Consider the function class $\mathcal{H}_{i,j}(\delta) := \big\{ h_{\boldsymbol{\gamma}}^j(\mathbf{z},y) : \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 \leq \delta \big\}$. Set $\mathcal{H}_{i,j} := \mathcal{H}_{i,j}(1)$. It follows that

$$\mathbb{P}\Bigg( \sup_{\boldsymbol{\gamma}:\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2\leq 1} \Big| \frac{1}{T}\sum_{t=1}^T M_{it,j}(\boldsymbol{\gamma}) \Big| > C_2\sqrt{\frac{\log n}{T}} \Bigg) = \mathbb{P}\Big( \|\mathbb{G}_{T,i}\|_{\mathcal{H}_{i,j}} > C_2\sqrt{\log n} \Big)\,.$$

Now, we study the probability

$$\mathbb{P}\Big( \|\mathbb{G}_{T,i}\|_{\mathcal{H}_{i,j}} > C_2\sqrt{\log n} \Big)\,.$$

Note that for any $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ satisfying $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 < 1$, it holds that

$$\left| yz_j - \frac{e^{\mathbf{z}^\top \boldsymbol{\gamma}} z_j}{1 + e^{\mathbf{z}^\top \boldsymbol{\gamma}}} \right| = |z_j| \left| y - \frac{e^{\mathbf{z}^\top \boldsymbol{\gamma}}}{1 + e^{\mathbf{z}^\top \boldsymbol{\gamma}}} \right| \leq \|\mathbf{z}\|_2 \,, \tag{41}$$

and thus an envelope for the class $\mathcal{H}_{i,j}$ is given by $\kappa$. Moreover, for any functions $h_{\boldsymbol{\gamma}}^j(\mathbf{z}, y), h_{\tilde{\boldsymbol{\gamma}}}^j(\mathbf{z}, y) \in \mathcal{H}_{i,j}$, it holds that

$$\left| yz_j - \frac{e^{\mathbf{z}^\top \boldsymbol{\gamma}} z_j}{1 + e^{\mathbf{z}^\top \boldsymbol{\gamma}}} - yz_j + \frac{e^{\mathbf{z}^\top \tilde{\boldsymbol{\gamma}}} z_j}{1 + e^{\mathbf{z}^\top \tilde{\boldsymbol{\gamma}}}} \right| \tag{42}$$

$$\leq \|\mathbf{z}\|_2 \left| \frac{e^{\mathbf{z}^\top \boldsymbol{\gamma}} z_j}{1 + e^{\mathbf{z}^\top \boldsymbol{\gamma}}} - \frac{e^{\mathbf{z}^\top \tilde{\boldsymbol{\gamma}}} z_j}{1 + e^{\mathbf{z}^\top \tilde{\boldsymbol{\gamma}}}} \right| \tag{43}$$

$$\leq \|\mathbf{z}\|_2^2 \, \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_2 \,, \tag{44}$$

where the last inequality follows from the mean value theorem and the bound $e^z/(1+e^z)^2 \leq 1$. Thus, the $\epsilon$-bracketing number of the function class $\mathcal{H}_{i,j}$ satisfies

$$\sup_{i,j} N_{[]}(\epsilon, \mathcal{H}_{i,j}, \|\cdot\|_2) \leq C_4 \epsilon^{-p-1} \tag{45}$$

for a constant $C_4$ independent of $n$. By Theorem 2.14.2 in van der Vaart and Wellner (1996) and Assumption 3.4, we have

$$\mathbb{E}\left[ \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 < 1} \left( \sqrt{T} \left| \frac{1}{T} \sum_{t=1}^{T} M_{it,j}(\boldsymbol{\gamma}) \right| \right) \right] \lesssim 2\kappa J_{[]}(1, \mathcal{H}_{i,j}) \,,$$

where

$$J_{[]}(1, \mathcal{H}_{i,j}) := \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{H}_{i,j}, \|\cdot\|_2)} d\epsilon \leq \int_0^1 \sqrt{1 + \log(C_4 \epsilon^{-p-1})} d\epsilon < \infty$$

by Lemma 9.5. This implies

$$\mu := \sup_{i,j} \mathbb{E}\left[ \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 < 1} \left( \sqrt{T} \left| \frac{1}{T} \sum_{t=1}^{T} M_{it,j}(\boldsymbol{\gamma}) \right| \right) \right] \leq C_5 \kappa \tag{46}$$

for a constant $C_5$ independent of $n$. For a function class $\mathcal{H}$ define $\mu_i(\mathcal{H}) := \mathbb{E}\big[\|\mathbb{G}_{T,i}\|_{\mathcal{H}}\big]$, and $\sigma_i^2(\mathcal{H}) := \|\mathbb{P}_i[(h - \mathbb{P}_i h)^2]\|_{\mathcal{H}}$. By (46) we have $\mu_i\big(\mathcal{H}_{i,j}\big) \leq \mu$. Since the envelope for the class $\mathcal{H}_{i,j}$ is $\kappa$, it holds by Assumption 3.4 that

$$\sigma^2 := \sup_{i,j} \sigma_i^2\big(\mathcal{H}_{i,j}\big) \leq \kappa^2 \,.$$

Define $\tilde{C}_2 := C_2 C^*$ where $C^*$ denotes the universal constant $C$ from Theorem 2.14.25 in

van der Vaart and Wellner (1996). Set $t = C_2\sqrt{\log n} - \mu$ to obtain for $\log n > \mu^2/\tilde{C}_2^2$, it holds that

$$\sup_{i,j} \mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_{i,j}} > \tilde{C}_2\sqrt{\log n}\Big) \leq \sup_{i,j} \mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_{i,j}} > C^*\{\mu_i(\mathcal{H}_{i,j}) + t\}\Big)\,.$$

Invoking Theorem 2.14.25 in van der Vaart and Wellner (1996) yields

$$\begin{aligned}
&\sup_{i,j} \mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_{i,j}} > C^*\{\mu_i(\mathcal{H}_{i,j}) + t\}\Big) \\
&\leq \exp\left(-D\Big(\frac{(C_2\sqrt{\log n} - \mu)^2}{\sigma^2} \bigwedge \frac{(C_2\sqrt{\log n} - \mu)\sqrt{T}}{\kappa}\Big)\right),
\end{aligned}$$

where $D$ is a universal constant independent of $n, T, C_2$. Collecting pieces gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \leq 1} \Big|\frac{1}{T}\sum_{t=1}^{T} M_{it,j}(\boldsymbol{\gamma})\Big| > C_2\sqrt{\frac{\log n}{T}}\right) \\
&\leq \sum_{i=1}^{n} \mathbb{P}\left(\sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \leq 1} \Big|\frac{1}{T}\sum_{t=1}^{T} M_{it,j}(\boldsymbol{\gamma})\Big| > C_2\sqrt{\frac{\log n}{T}}\right) \\
&\leq \exp\left(\log n - D\Big(\frac{(C_2\sqrt{\log n} - \mu)^2}{\sigma^2} \bigwedge \frac{(C_2\sqrt{\log n} - \mu)\sqrt{T}}{\kappa}\Big)\right).
\end{aligned}$$

By assumption that $\log n/T \to 0$ and $n, T \to \infty$, we can pick a $\tilde{C}$ sufficiently large such that $C_2$ is large enough to obtain

$$\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \leq 1} \Big|\frac{1}{T}\sum_{t=1}^{T} M_{it,j}(\boldsymbol{\gamma})\Big| > C_2\sqrt{\frac{\log n}{T}}\right) \to 0\,.$$

This completes the proof. □

*Proof of Theorem 3.3 (ii).* Define the functions $h_i : \mathbb{R}^{p+1} \to \mathbb{R}^{(p+1)\times(p+1)}$ through

$$h_i(\boldsymbol{\gamma}) := \mathbb{E}\left[\frac{e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}{(1+e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}})^2}\mathbf{z}_{it}\mathbf{z}_{it}^\top\right].$$

Note that

$$\sup_i \Big|\Big|\Big|\widetilde{\Sigma}_i^{-1} - h_i(\hat{\boldsymbol{\gamma}}_i)\Big|\Big|\Big|_2 \leq \kappa \sup_i \lambda_{\max}(\mathbb{E}[\mathbf{z}_{it}\mathbf{z}_{it}^\top]) \sup_i \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2\,.$$

By Theorem 3.3 and Assumption 3.3, we obtain

$$\sup_i \Big|\Big|\Big|\widetilde{\Sigma}_i^{-1} - h_i(\hat{\boldsymbol{\gamma}}_i)\Big|\Big|\Big|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

Since

$$\sup_i \left|\left|\left|\hat{\widetilde{\Sigma}}_i^{-1} - \widetilde{\Sigma}_i^{-1}\right|\right|\right|_2 \leq \sup_i \left|\left|\left|\hat{\widetilde{\Sigma}}_i^{-1} - h_i(\hat{\boldsymbol{\gamma}}_i)\right|\right|\right|_2 + \sup_i \left|\left|\left|h_i(\hat{\boldsymbol{\gamma}}_i) - \widetilde{\Sigma}_i^{-1}\right|\right|\right|_2,$$

it remains to show that

$$\sup_i \left|\left|\left|\hat{\widetilde{\Sigma}}_i^{-1} - h_i(\hat{\boldsymbol{\gamma}}_i)\right|\right|\right|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{47}$$

For ease of notation, we define the matrix $N_{it}(\boldsymbol{\gamma}) \in \mathbb{R}^{(p+1)\times(p+1)}$ via

$$N_{it}(\boldsymbol{\gamma}) := \frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}})^2}\mathbf{z}_{it}\mathbf{z}_{it}^\top - \mathbb{E}\Big[\frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}})^2}\mathbf{z}_{it}\mathbf{z}_{it}^\top\Big].$$

Define the $(j,\ell)$-th entry of matrix $N_{it}(\boldsymbol{\gamma})$ by $N_{it,j,\ell}(\boldsymbol{\gamma})$. Given $\delta > 0$, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\sup_i \Big|\frac{1}{T}\sum_{t=1}^T N_{it,j,\ell}(\hat{\boldsymbol{\gamma}}_i)\Big| > C\sqrt{\frac{\log n}{T}}\right) \\
\leq &\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 \leq \delta} \Big|\frac{1}{T}\sum_{t=1}^T N_{it,j,\ell}(\boldsymbol{\gamma})\Big| > C\sqrt{\frac{\log n}{T}}\right) + \mathbb{P}\Big(\sup_i \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 > \delta\Big).
\end{aligned}$$

By equation (13) of Theorem 3.3, it holds that

$$\mathbb{P}\Big(\sup_i \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 > \delta\Big) \to 0.$$

It remains to bound the probability

$$\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2 \leq \delta} \Big|\frac{1}{T}\sum_{t=1}^T N_{it,j,\ell}(\boldsymbol{\gamma})\Big| > C\sqrt{\frac{\log n}{T}}\right).$$

Define the function $h_{\boldsymbol{\gamma}}^{j,\ell}(\mathbf{z})$ via

$$\begin{aligned}
&h_{\boldsymbol{\gamma}}^{j,\ell} : \mathbb{R}^{p+1} \to \mathbb{R} \\
&\qquad \mathbf{z} \mapsto h_{\boldsymbol{\gamma}}^{j,\ell}(\mathbf{z}) := \frac{e^{\mathbf{z}^\top \boldsymbol{\gamma}}}{(1+e^{\mathbf{z}^\top \boldsymbol{\gamma}})^2} z_j z_l \mathbb{1}\{\|\mathbf{z}\|_2 \leq \kappa\},
\end{aligned}$$

where $z_j$ denotes the $j$-th element of the vector $\mathbf{z}$. Consider the function class $\mathcal{H}_i^{j,\ell}(\delta) := \big\{h_{\boldsymbol{\gamma}}^{j,\ell}(\mathbf{z}) : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 \leq \delta\big\}$. It follows that

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \leq \delta} \Big|\frac{1}{T}\sum_{t=1}^T N_{it,j,\ell}(\boldsymbol{\gamma})\Big| > C\sqrt{\frac{\log n}{T}}\right) = \mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_i^{j,\ell}(\delta)} > C\sqrt{\log n}\Big).$$

Moreover, by Assumption 3.4, it holds for any function $h_{\boldsymbol{\gamma}}^{j,\ell}(\mathbf{z}) \in \mathcal{H}_i^{j,\ell}(\delta)$ that $|h_{\boldsymbol{\gamma}}^{j,\ell}(\mathbf{z})| \leq$

$\kappa^2/4$. Employing the similar entropy method as in the proof of equation (13) of Theorem 3.3, we obtain

$$\mathbb{E} \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le \delta} \left( \sqrt{T} \Big| \frac{1}{T} \sum_{t=1}^{T} N_{it,j,\ell}(\boldsymbol{\gamma}) \Big| \right) \le C_6 \kappa^2$$

for a constant $C_6$ independent of $n, T$. Defining $\mu_i(\mathcal{H}) := \mathbb{E}\big[\|\mathbb{G}_{T,i}\|_{\mathcal{H}}\big]$, and $\sigma_i^2(\mathcal{H}) := \|\mathbb{P}_i[(h - \mathbb{P}_i h)^2]\|_{\mathcal{H}}$ we have

$$\mu := \sup_i \mu_i\Big(\mathcal{H}_i^{j,\ell}(\delta)\Big) \le C_6 \kappa^2 \,,$$
$$\sigma^2 := \sup_i \sigma_i^2\Big(\mathcal{H}_i^{j,\ell}(\delta)\Big) \le \kappa^4 \,.$$

Denote the universal constant $C$ from Theorem 2.14.25 in van der Vaart and Wellner (1996) by $C^*$ and set $t = \tilde{C}\sqrt{\log n} - \mu$ with $\tilde{C} := C/C^*$ for $\log n > \mu_1^2/\tilde{C}^2$ to obtain

$$\mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_i^{j,\ell}(\delta)} > C\sqrt{\log n}\Big) \le \mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_i^{j,\ell}(\delta)} > C^*\{\mu_i\big(\mathcal{H}_i^{j,\ell}(\delta)\big) + t\}\Big) \,.$$

Invoking Theorem 2.14.25 in van der Vaart and Wellner (1996) yields

$$\begin{aligned}
&\mathbb{P}\Big(\|\mathbb{G}_{T,i}\|_{\mathcal{H}_i^{j,\ell}(\delta)} > C^*\{\mu_i\big(\mathcal{H}_i^{j,\ell}(\delta)\big) + t\}\Big) \\
&\le \exp\left( -D\Big( \frac{(\tilde{C}\sqrt{\log n} - \mu)^2}{\sigma^2} \bigwedge \frac{(\tilde{C}\sqrt{\log n} - \mu)\sqrt{T}}{\kappa^2} \Big) \right),
\end{aligned}$$

where $D$ is an universal constant independent of $n, T, C$. Collecting pieces gives

$$\begin{aligned}
&\mathbb{P}\left( \sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le \delta} \Big| \frac{1}{T} \sum_{t=1}^{T} N_{it,j,\ell}(\boldsymbol{\gamma}) \Big| > C\sqrt{\frac{\log n}{T}} \right) \\
&\le \sum_{i=1}^{n} \mathbb{P}\left( \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le \delta} \Big| \frac{1}{T} \sum_{t=1}^{T} N_{it,j,\ell}(\boldsymbol{\gamma}) \Big| > C\sqrt{\frac{\log n}{T}} \right) \\
&\le \exp\left( \log n - D\Big( \frac{(\tilde{C}\sqrt{\log n} - \mu)^2}{\sigma^2} \bigwedge \frac{(\tilde{C}\sqrt{\log n} - \mu)\sqrt{T}}{\kappa^2} \Big) \right).
\end{aligned}$$

By assumption $\log n/T \to 0$, and hence we can pick $C$ sufficiently large to obtain as $\min(n, T) \to \infty$

$$\mathbb{P}\left( \sup_i \sup_{\boldsymbol{\gamma}:\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2 \le \delta} \Big| \frac{1}{T} \sum_{t=1}^{T} N_{it,j,\ell}(\boldsymbol{\gamma}) \Big| > C\sqrt{\frac{\log n}{T}} \right) \to 0 \,.$$

This establishes (47). Note that the eigenvalues of $\mathbb{E}\Big[\frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*})^2} \mathbf{z}_{it}\mathbf{z}_{it}^\top\Big]$ are bounded uniformly away from zero and from above – indeed, boundedness from above follows since $\|\mathbf{z}_{it}\|_2 \le \kappa$ by assumption, for boundedness from below recall that $z \mapsto e^z/(1+e^z)^2$ is

decreasing and non-negative on $\mathbb{R}_+$ so that for $\kappa_1 := \max_i\{\|\boldsymbol{\gamma}_i^*\|_2\} < \infty$, it holds that

$$\frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*})^2} \geq \frac{e^{\kappa\kappa_1}}{(1+e^{\kappa\kappa_1})^2} > 0\,.$$

A Taylor expansion of the map $A \mapsto A^{-1}$ together completes the proof. □

### 9.3.2 Proofs for logistic regression under dependence: Theorem 3.4

*Proof of Theorem 3.4 (i).* The proof of Theorem 3.4 (i) is similar to the proof of Theorem 3.3 (i), the only difference is that we employ Proposition C.2 in Kato et al. (2012) instead of Talagrand's inequality for i.i.d. random variables used in the previous proof. We use the same notation as in the proof of Theorem 3.3 (i). To establish the desired result, we need to to derive the bounds (38) and (39). Note that the proof of display (38) remains unchanged under the dependent setting, so we omit the proof for the sake of brevity. We aim to show the bound (39), i.e.

$$\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2=\tilde{C}\sqrt{\frac{\log n}{T}}} (\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*)^\top \Big(\nabla \mathbb{M}_{i,T}(\boldsymbol{\gamma}) - \nabla \mathbb{M}_i(\boldsymbol{\gamma})\Big) = \mathcal{O}_{\mathbb{P}}\Big(\frac{\log n}{T}\Big)\,. \tag{48}$$

By the proof of Theorem 3.3 (i), it suffices to show that

$$\mathbb{P}\left(\sup_i \sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2\le 1} \frac{1}{T}\sum_{t=1}^T \Big|M_{it,j}(\boldsymbol{\gamma})\Big| > C_2\sqrt{\frac{\log n}{T}}\right) \to 0\,,$$

where $C_2$ is the constant defined in the proof of Theorem 3.3 (i),

$$M_{it}(\boldsymbol{\gamma}) = Y_{it}\mathbf{z}_{it} - \frac{\mathbf{z}_{it}e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}} - \mathbb{E}\left[Y_{it}\mathbf{z}_{it} - \frac{\mathbf{z}_{it}e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}_{it}^\top\boldsymbol{\gamma}}}\right].$$

and $M_{it,j}(\boldsymbol{\gamma})$ denotes the $j$-th entry of the vector $M_{it}(\boldsymbol{\gamma})$. Define the function class

$$\mathcal{H}_{i,j} := \Bigg\{(\mathbf{z},y)\mapsto\left[\Big(y\mathbf{z} - \frac{\mathbf{z}e^{\mathbf{z}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}^\top\boldsymbol{\gamma}}}\Big)_j - \mathbb{E}\Big[\Big(y\mathbf{z} - \frac{\mathbf{z}e^{\mathbf{z}^\top\boldsymbol{\gamma}}}{1+e^{\mathbf{z}^\top\boldsymbol{\gamma}}}\Big)_j\Big]\right]\mathbb{1}\{\|\mathbf{z}\|_2\le\kappa\} : \\ y\in\{0,1\}, \mathbf{z}\in\mathbb{R}^{p+1}, \boldsymbol{\gamma}\in\mathbb{R}^{p+1}, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_i^*\|_2\le 1\Bigg\}.$$

It then follows that

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma}:\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2\le 1}\Big|\frac{1}{T}\sum_{t=1}^T M_{it,j}(\boldsymbol{\gamma})\Big| > C_2\sqrt{\frac{\log n}{T}}\right) = \mathbb{P}\left(\|\mathbb{P}_{T,i}-\mathbb{P}_i\|_{\mathcal{H}_{i,j}} > C_2\sqrt{\frac{\log n}{T}}\right).$$

We will show that

$$\|\mathbb{P}_{T,i}-\mathbb{P}_i\|_{\mathcal{H}_{i,j}} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

By displays (41) and Assumption 3.4, it holds for any $i,j$ and $h\in\mathcal{H}_{i,j}$ that

$$\|h\|_\infty \le U_1\,,\ \text{and } \mathbb{V}\text{ar}\,(h)\le U_2\,,$$

with some universal constants $U_1, U_2 > 0$. Applying Lemma 4 in Galvao et al. (2020) to the function $h/U_2$ gives

$$\sup_{i,j} \sup_{h \in \mathcal{H}_{i,j}} \sup_{1 \le q \le T} \mathbb{V}\mathrm{ar}\Big( \frac{1}{q^{1/2}} \sum_{t=1}^{q} h(\mathbf{z}_{it}, Y_{it}) \Big) \le U_3 \,,$$

with some positive universal constant $U_3 < \infty$.

Note that the envelope for function class $\mathcal{H}_{i,j}$ is $2\kappa$ and the upper bound for the $\epsilon$-bracketing number in (45) holds for any $L_p$-norm and any probability measure $Q$. Then, we obtain the following bounds of the $\epsilon$-covering number for any probability measure $Q$ and any $0 < \epsilon < 1$ that

$$N(\mathcal{H}_{i,j}, L_1(Q), \epsilon) \le N_{[\,]}(\mathcal{H}_{i,j}, L_1(Q), \epsilon/2) \le (2A/\epsilon)^{\nu} \,,$$

with some constants $A, \nu < \infty$.

By Proposition C.2 of Kato et al. (2012), it holds for any $q_{n,T} \ge 1$ satisfying $q_{n,T}^2 \log(q_{n,T}) = o(T)$, any $i$, and any $s_{n,T} > 0$ that

$$\mathbb{P}\bigg( \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{H}_{i,j}} \ge C\Big( \sqrt{\frac{\log(q_{n,T})}{T}} + \sqrt{\frac{s_{n,T}}{T}} + \frac{s_{n,T} q_{n,T}}{T} \Big) \bigg) \le 2e^{-s_{n,T}} + 2T\beta(q_{n,T}) \,, \quad (49)$$

where $C > 0$ is a constant independent of $T, n, i, j$. Let $q_{n,T} := C_1(\log n + \log T)$ with the constant $C_1 > 1$ satisfying $b_\beta^{C_1} \le e^{-2}$. With the assumption that $T$ grows at most polynomially in $n$ and $(\log n)^3 = o(T)$, one can verify that $q_{n,T}^2 \log(q_{n,T}) = o(T)$. Let $s_{n,T} := 2 \log n$, it then holds for large $n, T$ that

$$\sqrt{\frac{\log(q_{n,T})}{T}} + \sqrt{\frac{s_{n,T}}{T}} + \frac{s_{n,T} q_{n,T}}{T} \lesssim \sqrt{\frac{\log n}{T}}$$

and

$$2e^{-s_{n,T}} + 2T\beta(q_{n,T}) \lesssim \frac{1}{n^2} + \frac{1}{n^2 T} \,.$$

Taking the union bound for (49) over $i = 1, \ldots, n$ gives the desired result.

□

*Proof of Theorem 3.4 (ii).* The proof of $\sup_i \Big\| \widehat{B}_{iT}^{-1} - B_i^{-1} \Big\|_2 = o_{\mathbb{P}}(1)$ is similar to the proof of Theorem 3.3 (ii), which boils down to show that

$$\mathbb{P}\bigg( \sup_i \sup_{\boldsymbol{\gamma}: \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_i^*\|_2 \le \delta} \Big| \frac{1}{T} \sum_{t=1}^{T} N_{it,j,\ell}(\boldsymbol{\gamma}) \Big| > C \sqrt{\frac{\log n}{T}} \bigg) \to 0 \,,$$

where

$$N_{it}(\boldsymbol{\gamma}) := \frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}})^2}\mathbf{z}_{it}\mathbf{z}_{it}^\top - \mathbb{E}\left[\frac{e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}}}{(1+e^{\mathbf{z}_{it}^\top \boldsymbol{\gamma}})^2}\mathbf{z}_{it}\mathbf{z}_{it}^\top\right],$$

and $N_{it,j,\ell}(\boldsymbol{\gamma})$ denotes the $(j,\ell)$-th entry of the matrix $N_{it}(\boldsymbol{\gamma})$. The desired result follows by an application of the union bound and Proposition C.2 of Kato et al. (2012) with similar arguments as in the proof of Theorem 3.4 (i) after noting that by Lemma 4 in Galvao et al. (2020), we have

$$\sup_{j,\ell}\sup_{i}\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{q}}\sum_{t=1}^{q}N_{it,j,\ell}(\boldsymbol{\gamma})\right) = \mathcal{O}(1).$$

Thus, it remains to show that $\sup_i \left|\!\left|\!\left|\widehat{H}_{iT} - H_i\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1)$. Similar to the proof of the convergence of $\widehat{B}_{iT}^{-1}$, one can verify that $\sup_i \left|\!\left|\!\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{it}^\top - \mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i1}^\top\big]\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1)$. We now aim to show that

$$\sup_i \left|\!\left|\!\left|\sum_{1\le j\le m_T}\Big(1-\frac{j}{T}\Big)\left(\frac{1}{T}\sum_{t=1}^{T-j}\big(\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{i,t+j}^\top + \widehat{\mathbf{w}}_{i,t+j}\widehat{\mathbf{w}}_{it}^\top\big)\right) - \sum_{j=1}^{\infty}\mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i,1+j}^\top + \mathbf{w}_{i,1+j}\mathbf{w}_{i1}^\top\big]\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1).$$

To this end, we introduce an intermediate term

$$\widetilde{A}_{iT} := \sum_{1\le j\le m_T}\Big(1-\frac{j}{T}\Big)\left(\frac{1}{T}\sum_{t=1}^{T-j}\big(\mathbf{w}_{it}\mathbf{w}_{i,t+j}^\top + \mathbf{w}_{i,t+j}\mathbf{w}_{it}^\top\big)\right),$$

and we define

$$\widehat{A}_{iT} := \sum_{1\le j\le m_T}\Big(1-\frac{j}{T}\Big)\left(\frac{1}{T}\sum_{t=1}^{T-j}\big(\widehat{\mathbf{w}}_{it}\widehat{\mathbf{w}}_{i,t+j}^\top + \widehat{\mathbf{w}}_{i,t+j}\widehat{\mathbf{w}}_{it}^\top\big)\right), \quad A_i := \sum_{j=1}^{\infty}\mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i,1+j}^\top + \mathbf{w}_{i,1+j}\mathbf{w}_{i1}^\top\big].$$

So, we aim to show that $\sup_i \left|\!\left|\!\left|\widehat{A}_{iT} - A_i\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1)$. Consider the decomposition

$$\sup_i \left|\!\left|\!\left|\widehat{A}_{iT} - A_i\right|\!\right|\!\right|_2 \le \sup_i \left|\!\left|\!\left|\mathbb{E}[\widetilde{A}_{iT}] - A_i\right|\!\right|\!\right|_2 + \sup_i \left|\!\left|\!\left|\mathbb{E}[\widetilde{A}_{iT}] - \widehat{A}_{iT}\right|\!\right|\!\right|_2.$$

We note that $\sup_i \left|\!\left|\!\left|\mathbb{E}[\widetilde{A}_{iT}] - A_i\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1)$ follows by similar arguments as the proof of the last display in the proof of Lemma 12 in Galvao et al. (2020).

It remains to show that $\sup_i \left|\!\left|\!\left|\mathbb{E}[\widetilde{A}_{iT}] - \widehat{A}_{iT}\right|\!\right|\!\right|_2 = o_{\mathbb{P}}(1)$.

Invoking the triangle inequality again, we have

$$\left|\!\left|\!\left|\mathbb{E}[\widetilde{A}_{iT}] - \widehat{A}_{iT}\right|\!\right|\!\right|_2 \le \left|\!\left|\!\left|\widetilde{A}_{iT} - \mathbb{E}[\widetilde{A}_{iT}]\right|\!\right|\!\right|_2 + \left|\!\left|\!\left|\widetilde{A}_{iT} - \widehat{A}_{iT}\right|\!\right|\!\right|_2.$$

To bound $\sup_i \left|\!\left|\!\left| \widetilde{A}_{iT} - \mathbb{E}[\widetilde{A}_{iT}] \right|\!\right|\!\right|_2$, observe that

$$\left|\!\left|\!\left| \widetilde{A}_{iT} - \mathbb{E}[\widetilde{A}_{iT}] \right|\!\right|\!\right|_2 \leq m_T \max_{j=1,\dots,m_T} \left|\!\left|\!\left| \frac{1}{T} \sum_{t=1}^{T-j} \mathbf{w}_{it}\mathbf{w}_{i,t+j}^\top + \mathbf{w}_{i,t+j}\mathbf{w}_{it}^\top - \mathbb{E}\big[\mathbf{w}_{i1}\mathbf{w}_{i,1+j}^\top + \mathbf{w}_{i,1+j}\mathbf{w}_{i1}^\top\big] \right|\!\right|\!\right|_2 .$$

By similar computations as in the proof of of display (53) in Galvao et al. (2020) one can show that

$$\sup_{j=1,\dots m_T} \sup_{i=1,\dots,n} \sup_{q\geq 1} \sup_{k,\ell} \mathbb{V}\mathrm{ar}\Big( \frac{1}{\sqrt{q}} \sum_{t=1}^{q} (\mathbf{w}_{it}\mathbf{w}_{i,t+j}^\top + \mathbf{w}_{i,t+j}\mathbf{w}_{it}^\top)_{k,\ell} \Big) = O(m_T)$$

By applying Corollary C.1 in Kato et al. (2012) with $q = C\log(nm_T), s = C\log(nm_T)$ for a suitable constant $C$ it follows that

$$\sup_i \left|\!\left|\!\left| \widetilde{A}_{iT} - \mathbb{E}[\widetilde{A}_{iT}] \right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}\Big( m_T \sqrt{\frac{m_T \log(nm_T)}{T}} \Big)$$

Finally, note that for all $t = 1, \dots, T$, by a Taylor expansion and Assumption 3.4

$$\|\hat{w}_{it} - w_{it}\|_2 \leq \kappa^2 \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2.$$

Thus by elementary computations

$$\max_i \left|\!\left|\!\left| \widetilde{A}_{iT} - \widehat{A}_{iT} \right|\!\right|\!\right|_2 \lesssim m_T \max_i \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2 = \mathcal{O}_{\mathbb{P}}\Big( m_T \sqrt{\frac{\log n}{T}} \Big) = o_{\mathbb{P}}(1).$$

Combining all bounds obtained so far we have

$$\sup_i \left|\!\left|\!\left| \widehat{A}_{iT} - A_i \right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}\Big( m_T \sqrt{\frac{\log n}{T}} + m_T \sqrt{\frac{m_T \log(nm_T)}{T}} \Big) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$$

by the assumptions on $m_T$. □

## 9.4 Proofs for quantile regression in the independent case (Theorem 3.5 and Theorem 3.7)

*Proof of Theorem 3.5(i).* Define $\boldsymbol{\gamma}_{n,T,i} := \hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*$. The Theorem 5.1 in Chao et al. (2017) can be used in our framework by setting $n = T, m = p+1, \xi_m = \kappa, g_n = 0$, and $c_n = 0$. We then find

$$\boldsymbol{\gamma}_{n,T,i} = -\frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \psi_{i,\tau}(\mathbf{z}_{it}, Y_{it}) + \boldsymbol{\gamma}_{n,T,i,1} + \boldsymbol{\gamma}_{n,T,i,2} + \boldsymbol{\gamma}_{n,T,i,3}\,, \tag{50}$$

where $B_i := \mathbb{E}[f_{Y|\mathbf{z}}(q_{i,\tau}(\mathbf{z}_{i1}) \mid \mathbf{z}_{i1})\mathbf{z}_{i1}\mathbf{z}_{i1}^\top]$ and $\psi_{i,\tau}(\mathbf{z}, Y) := \mathbf{z}(\mathbb{1}(Y \le q_{i,\tau}(\mathbf{z})) - \tau)$. Define $\tilde{\boldsymbol{\gamma}}_{n,T,i} := \boldsymbol{\gamma}_{n,T,i,1} + \boldsymbol{\gamma}_{n,T,i,2} + \boldsymbol{\gamma}_{n,T,i,3}$. We now prove that

$$\sup_i \|\tilde{\boldsymbol{\gamma}}_{n,T,i}\|_2 = o_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{51}$$

For this, we show that

$$\sup_i \|\boldsymbol{\gamma}_{n,T,i,k}\|_2 = o_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right), \quad k = 1,2,3\,.$$

Now, we handle the three remainder terms $\boldsymbol{\gamma}_{n,T,i,1}, \boldsymbol{\gamma}_{n,T,i,2}, \boldsymbol{\gamma}_{n,T,i,3}$ separately. By equation (5.1) in Theorem 5.1 of Chao et al. (2017), we have almost surely

$$\sup_i \|\boldsymbol{\gamma}_{n,T,i,1}\|_2 \le C/T$$

for a constant $C$ independent of $n, T, i$. Since $1/T = o(\sqrt{\log n/T})$ it follows that

$$\sup_i \|\boldsymbol{\gamma}_{n,T,i,1}\|_2 = o_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{52}$$

By equation (5.2) in Theorem 5.1 of Chao et al. (2017) applied with $\kappa_n = 2\log n \ll T$, there exists a constant $C_1$ independent of $n, T$ (and bounded uniformly in $i$ as seen by a close inspection of the corresponding proof in Chao et al. (2017)) such that for all sufficiently large $T$

$$\mathbb{P}\left( \|\boldsymbol{\gamma}_{n,T,i,2}\|_2 > C_1\Big(\sqrt{\frac{\log T}{T}} + \sqrt{\frac{2\log n}{T}}\Big)^2 \right) \le 2\exp(-\kappa_n) = 2/n^2\,. \tag{53}$$

Since

$$\left(\sqrt{\frac{\log T}{T}} + \sqrt{\frac{2\log n}{T}}\right)^2 \le 2\frac{2\log n + \log T}{T} = o\left(\sqrt{\frac{\log n}{T}}\right),$$

an application of the union bound shows that

$$\sup_i \|\boldsymbol{\gamma}_{n,T,i,2}\|_2 = o_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

Next apply (5.2) in Theorem 5.1 from Chao et al. (2017) with $\kappa_n = 2\log n \ll T$ to obtain the existence of a constant $C_2$ independent of $T$ (and bounded uniformly in $i$ as seen by a close inspection of the corresponding proof in Chao et al. (2017)) such that for all sufficiently large $T$

$$\mathbb{P}\left( \|\boldsymbol{\gamma}_{n,T,i,3}\|_2 > C_2\Big(\sqrt{\frac{\log T}{T}} + \sqrt{\frac{2\log n}{T}}\Big)^{3/2}\right) < 2/n^2\,. \tag{54}$$

Note that

$$\left(\sqrt{\frac{\log T}{T}} + \sqrt{\frac{2\log n}{T}}\right)^3 \leq 8\frac{(2\log n)^{3/2} + (\log T)^{3/2}}{T^{3/2}} = o\left(\frac{\log n}{T}\right)$$

by the assumption that $\log n = o(T)$. Combining this with the union bound and (54) shows that

$$\sup_i \|\boldsymbol{\gamma}_{n,T,i,3}\|_2 = o_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{T}}\Big),$$

and collecting pieces yields (51).

To complete the proof, define the classes of functions

$$\mathcal{G}_i := \Big\{(\mathbf{z}, y) \mapsto \mathbf{a}^\top \mathbf{z}(\mathbb{1}\{y \leq \mathbf{z}^\top \mathbf{b}\} - \tau)\mathbb{1}\{\|\mathbf{z}\|_2 \leq \kappa\} : \mathbf{b} \in \mathbb{R}^{p+1}, \mathbf{a} \in \mathbb{R}^{p+1}, \|\mathbf{a}\|_2 = 1\Big\}$$

and note that

$$\sup_i \left\| \frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \psi_{i,\tau}(\mathbf{z}_{it}, Y_{it}) \right\|_2 \leq \sup_i \left|\left| B_i^{-1} \right|\right|_2 \sup_i \|\mathbb{P}_{T,i} - \mathbb{P}_i\|_{\mathcal{G}_i} \tag{55}$$

for $\mathbb{P}_i$ denoting the measure of and $\mathbb{P}_{T,i}$ corresponding to the empirical measure of $\{(\mathbf{z}_{it}, Y_{it}), t = 1, \ldots, T\}$. Under the assumptions made we have $\sup_i \left|\left| B_i^{-1} \right|\right|_2 = O(1)$, and Lemma C.3 from Chao et al. (2017) applied with $\kappa_n = 2\log n \ll T$ shows that there exists a constant $C_3$, independent of $n, T$ (and bounded uniformly in $i$ as revealed by a close look at the corresponding proof) such that

$$\mathbb{P}\left(\|\mathbb{P}_{T,i} - \mathbb{P}_i\|_{\mathcal{G}_i} > C_3\sqrt{\frac{\log n}{T}}\right) \leq n^{-2}\,.$$

Applying the union bound shows that

$$\sup_i \|\mathbb{P}_{T,i} - \mathbb{P}_i\|_{\mathcal{G}_i} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).$$

Combining this with (51) completes the proof. □

Next we proceed to the proof of Theorem 3.5(ii). The proof will make use of the following additional notation

$$\begin{aligned}\psi_{i,\tau}(\mathbf{z},Y) &= \mathbf{z}(\mathbb{1}(Y \le q_{i,\tau}(\mathbf{z})) - \tau)\\ f_{it} &:= \frac{2d_T}{q_{i,\tau+d_T}(\mathbf{z}_{it}) - q_{i,\tau-d_T}(\mathbf{z}_{it})}\\ e_{it} &:= 1/f_{it}\\ B_{iT} &= \frac{1}{T}\sum_{t=1}^{T} f_{it}\mathbf{z}_{it}\mathbf{z}_{it}^\top\\ \widetilde{\Sigma}_{iT}^{-1} &= \mathbb{E}[B_{iT}]H_i^{-1}\mathbb{E}[B_{iT}]\,.\end{aligned}$$

We begin by stating and proving an intermediate technical result.

**Lemma 9.6.** *Let Assumptions 3.6-3.9 hold and assume* $\log n = o(T), \frac{\log n}{Td_T} = o(1)$. *Let* $\widehat{e}_{it} := \widehat{f}_{it}^{-1}$, *then* $\sup_{i,t}|\widehat{e}_{it} - e_{it}| = \mathcal{O}_{\mathbb{P}}(b_{n,T})$ *with* $b_{n,T} = \sqrt{\frac{\log n}{Td_T^2}}$.

*Proof of Lemma 9.6.* The proof essentially follows from the arguments in the proof of Lemma 9 of Galvao et al. (2020), but modifications are needed to take into account that $n(\log T)^2/T = o(1)$ made in that paper is replaced by $\log n = o(T)$ and that the rate changes accordingly. By definitions of $\widehat{e}_{is}$ and $e_{is}$, it holds that

$$\widehat{e}_{is} - e_{is} = \mathbf{z}_{is}^\top\Big(\big(\hat{\boldsymbol{\gamma}}_i(\tau+d_T) - \boldsymbol{\gamma}_i^*(\tau+d_T)\big) - \big(\hat{\boldsymbol{\gamma}}_i(\tau-d_T) - \boldsymbol{\gamma}_i^*(\tau-d_T)\big)\Big)/2d_T\,.$$

We know from the display (50) and Theorem 3.5 that

$$\begin{aligned}&\mathbf{z}_{is}^\top(\hat{\boldsymbol{\gamma}}_i(\tau\pm d_T) - \boldsymbol{\gamma}_i^*(\tau\pm d_T))\\ &\qquad = -\frac{1}{T}\mathbf{z}_{is}^\top B_i^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}\Big(\mathbb{1}\{Y_{it} \le q_{i,\tau\pm d_T}(\mathbf{z}_{it})\} - (\tau\pm d_T)\Big) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right).\end{aligned}$$

Hence with $U_{it} := F_{Y|\mathbf{z}}(Y_{it}|\mathbf{z}_{it}) \sim U[0,1]$ independent of $\mathbf{z}_{it}$, it holds that

$$\widehat{e}_{is} - e_{is} = -\frac{1}{2Td_T}\mathbf{z}_{is}^\top B_i^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}\Big(\mathbb{1}\{U_{it} \le \tau+d_T\} - \mathbb{1}\{U_{it} \le \tau-d_T\} - 2d_T\Big) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{d_T}\sqrt{\frac{\log n}{T}}\right). \tag{56}$$

Define the vectors $M_{it} \in \mathbb{R}^{p+1}$ via

$$M_{it} := \mathbf{z}_{it}\Big(\mathbb{1}\{U_{it} \le \tau+d_T\} - \mathbb{1}\{U_{it} \le \tau-d_T\} - 2d_T\Big)\Big/2d_T\,.$$

Fix an arbitrary $k \in \{1, ..., p+1\}$ and let $M_{it,k}$ denote the $k$-th entry of the vector $M_{it}$. It then follows that $\mathbb{E}[M_{it,k}] = 0$ and $\sup_i \mathbb{V}\mathrm{ar}\,[M_{it,k}] \leq \frac{C_1}{d_T}$ for some constant $C_1$ under Assumption 3.6. Under Assumption 3.6, we also have $\sup_{i,t,k} |M_{it,k}| \leq C_2/d_T$ for some constant $C_2 > 0$. Invoking the Bernstein inequality yields

$$\begin{aligned}\mathbb{P}\bigg(\Big|\sum_{t=1}^{T} M_{it,k}\Big| > T\epsilon\bigg) \leq & 2\exp\bigg(-\frac{\frac{1}{2}T^2\epsilon^2}{\sum_{t=1}^{T}\mathbb{E}[M_{it,k}^2] + \frac{1}{3}C_2 d_T^{-1} T\epsilon}\bigg) \\ = & 2\exp\bigg(-\frac{\frac{1}{2}T^2\epsilon^2}{C_1 T d_T^{-1} + \frac{1}{3}C_2 d_T^{-1} T\epsilon}\bigg).\end{aligned}$$

Take $\epsilon = C_3 T^{-1/2} d_T^{-1/2} (\log n)^{1/2}$ for a constant $C_3$ which will be determined later. Under the assumption $\frac{\log n}{T d_T} \to 0$, it follows that $\epsilon \to 0$ and the right hand side of the inequality becomes

$$2\exp\bigg(-\frac{1}{2}\frac{(C_3)^2 d_T^{-1}\log n}{C_1 d_T^{-1} + \frac{1}{3}C_2 C_3 d_T^{-1} T^{-1/2} d_T^{-1/2} (\log n)^{1/2}}\bigg) \leq 2\exp\Big(-\frac{1}{4}(C_3)^2 \log n / C_1\Big),$$

where the last inequality holds for $\log n/(T d_T)$ sufficiently small. Then, we have

$$\begin{aligned}\mathbb{P}\bigg(\sup_k \sup_i \Big|\frac{1}{T}\sum_{t=1}^{T} M_{it,k}\Big| > \epsilon\bigg) \leq & \sum_k \sum_i \mathbb{P}\bigg(\Big|\frac{1}{T}\sum_{t=1}^{T} M_{it,k}\Big| > \epsilon\bigg) \\ \leq & 2np\exp\bigg(-\frac{(C_3)^2}{4C_1}\log n\bigg) \to 0\end{aligned}$$

by taking $(C_3)^2 > 4C_1$. Hence, we obtain

$$\sup_i \bigg\|\frac{1}{T}\sum_{t=1}^{T} M_{it}\bigg\|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{\log n}{T d_T}}\bigg).$$

Combining this with (56), the fact that $\sup_i \big\|\big|B_i^{-1}\big|\big\|_2 = \mathcal{O}(1)$, and Assumption 3.9 gives

$$\sup_{i,s} |\widehat{e}_{is} - e_{is}| = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{\log n}{T d_T}} + \sqrt{\frac{\log n}{T d_T^2}}\bigg) = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{\log n}{T d_T^2}}\bigg)$$

as desired. □

*Proof of Theorem 3.5(ii).* The following bound follows by the same arguments as Lemma 8 of Galvao et al. (2020) (note that the condition $n(\log T)^2/T = o(1)$ made in that paper is not used in their proof of Lemma 8):

$$\sup_i \big\|\big|\mathbb{E}[B_{iT}] - B_i\big|\big\|_2 = o(1). \tag{57}$$

In addition, we will prove the following bounds

$$\sup_i \left|\!\left|\!\left| \widehat{B}_{iT} - B_{iT} \right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}(b_{n,T}) \tag{58}$$

with $b_{n,T} = \sqrt{\frac{\log n}{T d_T^2}}$,

$$\sup_i |||B_{iT} - \mathbb{E}[B_{iT}]|||_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right), \tag{59}$$

and

$$\sup_i \left|\!\left|\!\left| \widehat{H}_{iT}^{-1} - H_i^{-1} \right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{T}}\right). \tag{60}$$

The remaining proof follows from similar arguments as the proof of Lemma 10 of Galvao et al. (2020), but modifications are needed to take into account that $n(\log T)^2/T = o(1)$ made in that paper is replaced by $\log n = o(T)$ and that the rate changes accordingly. We note that

$$\widehat{B}_{iT} - B_{iT} = \frac{1}{T}\sum_{t=1}^{T} (\widehat{f}_{it} - f_{it}) \mathbf{z}_{it}\mathbf{z}_{it}^\top .$$

Using Taylor expansion, we have

$$\widehat{f}_{it} - f_{it} = \widehat{e}_{it}^{-1} - e_{it}^{-1} = \frac{e_{it} - \widehat{e}_{it}}{e_{it}^2} + \mathcal{O}\Big(|\widehat{e}_{it} - e_{it}|^2\Big), \tag{61}$$

where the remainder term is uniform in $i, t$ since under Assumption 3.7, it holds that

$$\begin{aligned} \inf_{i,t} e_{i,t} &= \inf_{i,t} \frac{q_{i,\tau+d_T}(\mathbf{z}_{it}) - q_{i,\tau-d_T}(\mathbf{z}_{it})}{2d_T} \geq \inf_{i,t} \inf_{|\eta-\tau|\leq d_T} \frac{1}{f_{Y|\mathbf{z}}(q_{i,\eta}(\mathbf{z}_{it}) \mid \mathbf{z}_{it})} \\ &= \frac{1}{\sup_{i,t} \sup_{\eta,\mathbf{z}} f_{Y|\mathbf{z}}(q_{i,\eta}(\mathbf{z}_{it}) \mid \mathbf{z}_{it})} \geq 1/f_{max} \,, \end{aligned} \tag{62}$$

almost surely. By Assumption 3.6 and Lemma 9.6, the bound in (58) follows.
Next we prove (59). Define the matrix $N_{it} \in \mathbb{R}^{(p+1)\times(p+1)}$ via

$$N_{it} := f_{it}\mathbf{z}_{it}\mathbf{z}_{it}^\top - \mathbb{E}[f_{it}\mathbf{z}_{it}\mathbf{z}_{it}^\top] \,.$$

It then follows that $\mathbb{E}[N_{it}] = \mathbf{0}$. We denote by $N_{it,j,\ell}$ the $(j,\ell)$-th entry of the matrix $N_{it}$. By Assumption 3.6 and the inequality (62), we have $\sup_{i,t,j,\ell} |N_{it,j,\ell}| \leq C_5$ and $\sup_{i,t,j,\ell} \mathbb{V}\mathrm{ar}\,[N_{it,j,\ell}] \leq$

$C_6$ for some constants $C_5$, $C_6 > 0$. Applying the Bernstein inequality gives, for any $\epsilon_2 > 0$,

$$\begin{aligned}\mathbb{P}\bigg(\Big|\sum_{t=1}^{T} N_{it,j,\ell}\Big| > T\epsilon_2\bigg) \leq & 2\exp\bigg(-\frac{\frac{1}{2}T^2\epsilon_2^2}{\sum_{t=1}^{T}\mathbb{E}[N_{it,j,\ell}^2] + \frac{1}{3}C_5 T\epsilon_2}\bigg) \\ \leq & 2\exp\bigg(-\frac{\frac{1}{2}T^2\epsilon_2^2}{TC_6 + \frac{1}{3}C_5 T\epsilon_2}\bigg).\end{aligned}$$

Take $\epsilon_2 = C_7 T^{-1/2}(\log n)^{1/2}$ for some constant $C_7 > 0$ to be determined later, and the right hand side of the inequality becomes, for $\log n/T$ sufficiently small,

$$2\exp\bigg(-\frac{1}{2}\frac{(C_7)^2\log n}{C_6 + \frac{1}{3}C_5 C_7 T^{-1/2}(\log n)^{1/2}}\bigg) \leq 2\exp\bigg(-\frac{(C_7)^2\log n}{4C_6}\bigg).$$

Choosing $(C_7)^2 > 4C_6$, then for every $j, \ell$, it holds that

$$\mathbb{P}\bigg(\sup_i\Big|\frac{1}{T}\sum_{t=1}^{T} N_{it,j,\ell}\Big| > \epsilon_2\bigg) \leq \sum_{i=1}^{n}\mathbb{P}\bigg(\Big|\frac{1}{T}\sum_{t=1}^{T} N_{it,j,\ell}\Big| > \epsilon_2\bigg) = 2n\exp\Big(-\frac{(C_7)^2\log n}{4C_6}\Big) \to 0\,.$$

Thus, $\sup_i \big|\!\big|\!\big|\frac{1}{T}\sum_{t=1}^{T} N_{it}\big|\!\big|\!\big|_2 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{T}}\Big)$. This implies (59). Finally, we prove the bound (60). By Assumption 3.6, it holds that $\sup_i \big|\!\big|\!\big|H_i^{-1}\big|\!\big|\!\big|_2 < \infty$. Moreover, we have

$$\begin{aligned}\widehat{H}_{iT}^{-1} - H_i^{-1} = H_i^{-1}(H_i\widehat{H}_{iT}^{-1} - I) = H_i^{-1}(H_i - \widehat{H}_{iT})\widehat{H}_{iT}^{-1} \\ = H_i^{-1}(H_i - \widehat{H}_{iT})H_i^{-1} + \mathcal{O}\Big(\big|\!\big|\!\big|H_i^{-1}\big|\!\big|\!\big|_2^2\,\big|\!\big|\!\big|\widehat{H}_{iT} - H_i\big|\!\big|\!\big|_2^2\Big), \quad (63)\end{aligned}$$

where

$$\sup_i \big|\!\big|\!\big|\widehat{H}_{iT} - H_i\big|\!\big|\!\big|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{\log n}{T}}\bigg)$$

holds by an application of the Bernstein inequality which is similar to the one given above. This completes the proof. □

*Proof sketch of Theorem 3.7* Both parts follow by simple computations provided that we can establish the bound

$$\sup_{|\eta-\tau|\leq\varepsilon}\sup_{i\in\{1,\ldots,n\}}|\hat{\alpha}_i(\eta) - \alpha_i^*(\eta)| = \mathcal{O}_{\mathbb{P}}\bigg(\sqrt{\frac{\log n}{T}}\bigg).$$

for some $\varepsilon > 0$. This can be established by following the arguments given in Step 1–Step 3 in the proof of Theorem 3.2 in Kato et al. (2012). Note that all empirical processes appearing in those steps retain the same complexity (in terms of VC dimension and envelope functions). Note also that the assumption that $T$ grows at most polynomially in $n$ made in their Theorem 3.2 can be dropped at the cost of replacing $\log n$ by $\log(T \vee n)$, see also

the discussion in the latter paper following Theorem 3.2. □

### 9.4.1 Proof for quantile regression in the dependent case (Theorem 3.6)

Before proving Theorem 3.6 we collect some preliminary technical results. Let $\mathcal{S}^{p+1} := \{\mathbf{a} \in \mathbb{R}^{p+1}, \|\mathbf{a}\|_2 = 1\}$. Let $\tilde{\mathcal{T}} = [\tau - \varepsilon, \tau + \varepsilon]$ where $\varepsilon > 0$ is such that $\tilde{\mathcal{T}} \subset \mathcal{T}$ for $\mathcal{T}$ from Assumption 3.8. Define the function classes

$$\mathcal{G}_1 := \Big\{(\mathbf{z}, y) \mapsto \mathbf{a}^\top \mathbf{z}(\mathbb{1}\{y \leq \mathbf{z}^\top \mathbf{b}\} - \tau)\mathbb{1}\{\|\mathbf{z}\|_2 \leq \kappa\} : \mathbf{b} \in \mathbb{R}^{p+1}, \tau \in \tilde{\mathcal{T}}, \mathbf{a} \in \mathcal{S}^{p+1}\Big\}. \quad (64)$$

$$\mathcal{G}_2(\delta) := \Big\{(y, \mathbf{z}) \mapsto \mathbf{a}^\top \mathbf{z}(\mathbb{1}\{y \leq \mathbf{b}_1^\top \mathbf{z}\} - \mathbb{1}\{y \leq \mathbf{b}_2^\top \mathbf{z}\})\mathbb{1}\{\|\mathbf{z}\|_2 \leq \kappa\} \Big| \|\mathbf{b}_1 - \mathbf{b}_2\|_2 \leq \delta, \mathbf{a} \in \mathcal{S}^{p+1}\Big\}. \quad (65)$$

Further, define the functions

$$g_{b,k,\ell}(\mathbf{z}_1, \mathbf{z}_2, y_1, y_2) := \mathbf{z}_{1,k}\mathbf{z}_{2,\ell}(\mathbb{1}\{y_1 \leq \mathbf{z}_1^\top \mathbf{b}\} - \tau)(\mathbb{1}\{y_2 \leq \mathbf{z}_2^\top \mathbf{b}\} - \tau), \quad k, \ell = 1, \ldots, d, \mathbf{b} \in \mathbb{R}^d.$$

With this notation, let

$$\mathcal{G}_{3,k,\ell} := \Big\{(\mathbf{z}_1, \mathbf{z}_2, y_1, y_2) \mapsto g_{\boldsymbol{b},k,\ell}(\mathbf{z}_1, \mathbf{z}_2, y_1, y_2) : \boldsymbol{b} \in \mathbb{R}^{p+1}\Big\},$$

and

$$\mu_{3,k,\ell}(\boldsymbol{b}, i, j) := \mathbb{E}[g_{\boldsymbol{b},k,\ell}(\mathbf{z}_{i1}, \mathbf{z}_{i,1+j}, y_{i1}, y_{i,1+j})].$$

Consider the empirical measures $\mathbb{P}_{i,j,T}$ corresponding to $\big\{(\mathbf{z}_{i,t}, \mathbf{z}_{i,t+j}, y_{i,t}, y_{i,t+j})\big\}_{t=1,\ldots,T}$ and denote by $\tilde{\mathbb{P}}_{i,j}$ the distribution of $(\mathbf{z}_{i,1}, \mathbf{z}_{i,1+j}, y_{i,1}, y_{i,1+j})$. Note that for $j \neq 0$ this includes "observations" outside of the observable sample. This quantity only appears in the proofs and is not used to compute any of the estimators. With this notation we have the following technical result.

**Lemma 9.7.** *Assume the conditions of Theorem 3.6(i). For $n$ sufficiently large we have for all $s > 0$ and all $1 \ll q_{n,T}$ with $q_{n,T}^2 \log q_{n,T} = o(T)$ for constants $C_{\mathcal{G}_1}, \tilde{C}_{\mathcal{G}_1}$*

$$\mathbb{P}\left(\|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_1} \geq C_{\mathcal{G}_1}\left(\sqrt{\frac{\log q_{n,T}}{T}} + \sqrt{\frac{s}{T}} + \frac{s q_{n,T}}{T}\right)\right) \leq 2e^{-s} + 2T\beta(q_{n,T}), \quad (66)$$

$$\mathbb{P}\Big(\max_{i=1,\ldots,n} \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_1} \leq \tilde{C}_{\mathcal{G}_1}\sqrt{\frac{\log(nT)}{T}}\Big) \geq 1 - \frac{1}{nT}. \quad (67)$$

*Further, we also have for all $s > 0$ and all $1 \ll q_{n,T}$ with $q_{n,T}^2 \log(m_T \vee q_{n,T}) = o(T)$ and a*

*constant $C_{\mathcal{G}_3}$ for $n$ sufficiently large*

$$\mathbb{P}\left(\|\mathbb{P}_{i,j,T} - \tilde{\mathbb{P}}_{i,j}\|_{\mathcal{G}_{3,k,\ell}} \geq C_{\mathcal{G}_3}\left(\sqrt{\frac{m_T \log q_{n,T}}{T}} + \sqrt{\frac{sm_T}{T}} + \frac{sq_{n,T}}{T}\right)\right) \leq 2e^{-s} + 2T\beta(q_{n,T}) \tag{68}$$

$$\max_{i,j,k,\ell} \|\mathbb{P}_{i,j,T} - \mathbb{P}_i\|_{\mathcal{G}_{3,k,\ell}} = O_{\mathbb{P}}\Big(\sqrt{\frac{m_T \log(nT)}{T}}\Big). \tag{69}$$

*Next, let*

$$\sigma^2_{q,i}(g) := \mathbb{V}ar\Big(\frac{1}{\sqrt{q}}\sum_{t=1}^{q} g(\mathbf{z}_{it}, Y_{it})\Big)$$

*and assume that*

$$\sigma^2_{n,T}(\delta) \geq \sup_i \sup_{g \in \mathcal{G}_2(\delta)} \sigma^2_{q,i}(g)\,. \tag{70}$$

*Then for any $s > 0$ and any $q_{n,T}$ satisfying*

$$q^2_{n,T} \log\Big(\frac{q_{n,T}}{\sigma^2_{n,T}(\delta)}\Big) \leq \tilde{C} T \sigma^2_{n,T}(\delta)\,, \tag{71}$$

*for a certain constant $\tilde{C}$ depending only on $\kappa$ and the dimension $p$ of $\mathbf{z}_{it}$ we have*

$$\mathbb{P}\left(\|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_2(\delta)} \geq C\left(\sqrt{\frac{\sigma^2_{n,T}(\delta)}{T}\log\Big(\frac{q_{n,T}}{\sigma^2_{n,T}(\delta)}\Big)} + \sqrt{\frac{\sigma^2_{n,T}(\delta)s}{T}} + \frac{sq_{n,T}}{T}\right)\right) \leq 2e^{-s} + 2T\beta(q_{n,T}). \tag{72}$$

*In particular, for $\delta = \delta_{n,T} := (CT^{-1}\log(nT))^{1/2}$ with $C > 0$ arbitrary but fixed we obtain*

$$\max_{i=1,\ldots,n} \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_2(\delta_{n,T})} = O_{\mathbb{P}}\Big(\frac{(\log(nT))^{5/4}}{T^{3/4}}\Big). \tag{73}$$

*Finally, letting $U_{it} := F_{Y_{it}|X_{it}}(Y_{it}|X_{it})$,*

$$\sup_i \left\| \frac{1}{2Td_T}\sum_{t=1}^{T} \mathbf{z}_{it}\Big(\mathbb{1}\{U_{it} \leq \tau + d_T\} - \mathbb{1}\{U_{it} \leq \tau - d_T\} - 2d_T\Big)\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{Td_T}}\right). \tag{74}$$

*Proof of Lemma 9.7* The proof strategy for many parts is similar to that in the proof of Lemma 5 in Galvao et al. (2020) and we will only point out the relevant differences. We will repeatedly apply Proposition C.2 from Kato et al. (2012). That result requires the corresponding function classes to be centered. Assume that $\mathcal{F}$ is a class of functions that are not centered and such that $\sup_Q N(\mathcal{F}, L_1(Q), \epsilon) \leq (A/\epsilon)^\nu$ for some constants $A, \nu$ and let $\tilde{\mathcal{F}} := \{f - \mathbb{P}f : f \in \mathcal{F}\}$. The it is easy to see that $N(\tilde{\mathcal{F}}, L_1(Q), \epsilon) \leq N(\mathcal{F}, L_1(Q), \epsilon/2)$

and $\|f-\mathbb{P}f\|_\infty \leq 2\|f\|_\infty$ so that Proposition C.2 from Kato et al. (2012) can be applied to non-centered function classes to obtain bounds on $\|\mathbb{P}_T f-\mathbb{P}f\|_{\mathcal{F}}$. This fact will be repeatedly used throughout the proofs that follow.

**Proof of** (66) **and** (67) By the proof of Lemma 5 in Galvao et al. (2020), it holds for any $g \in \mathcal{G}_1$ that

$$\|g\|_\infty \leq U_1, \quad \text{and} \quad \sup_i \sup_{g\in\mathcal{G}_1} \mathbb{V}\text{ar}\left(g(\mathbf{z}_{i1}, Y_{i1})\right) \leq U_2$$

with some positive universal constants $U_1$ and $U_2$. Moreover, it holds for any probability measure $Q$ and any $0 < \epsilon < 1$ that

$$N(\mathcal{G}_1, L_1(Q), \epsilon) \leq (A/\epsilon)^\nu$$

with some positive constants $A, \nu < \infty$. The claim in (66) follows by Proposition C.2 of Kato et al. (2012). For (67), let $q_{n,T} := C_1 \log(nT)$ with the constant $C_1 \geq 1$ satisfying $b_\beta^{C_1} \leq e^{-2}$ and $s = 2\log(nT)$. Clearly $q_{n,T} \gg 1$, $q_{n,T}^2 \log(q_{n,T}) = o(T)$, so (67) follows from the union bound and simple calculations.

**Proof of** (68) **and** (69) Observe that any function in $\mathcal{G}_{3,k,\ell}$ can be expressed as through sums and products of functions from the classes $\mathcal{H}_1 := \{(y_1, \mathbf{z}_1, y_2, \mathbf{z}_2) \mapsto \mathbf{z}_{1,k}\mathbf{z}_{2,\ell} | 1 \leq k, \ell \leq p+1\}, \mathcal{H}_2 := \{(y_1, \mathbf{z}_1, y_2, \mathbf{z}_2) \mapsto \tau - \mathbb{1}\{y_1 \leq \mathbf{z}_1^\top \mathbf{b}\} | \mathbf{b} \in \mathbb{R}^{p+1}\}$, $\mathcal{H}_3 := \{(y_1, \mathbf{z}_1, y_2, \mathbf{z}_2) \mapsto \tau - \mathbb{1}\{y_2 \leq \mathbf{z}_2^\top \mathbf{b}\} | \mathbf{b} \in \mathbb{R}^{p+1}\}$ and that each of the three classes satisfies

$$N(\mathcal{H}_j, L_2(Q), \epsilon) \leq (\widetilde{A}/\epsilon)^{\widetilde{v}},$$

for all $0 < \epsilon \leq 1$ and some constants $\widetilde{A}, \widetilde{v} < \infty$. Hence, by the Cauchy-Schwarz inequality and Lemma 23 in Belloni et al. (2019) (note that the proof of this Lemma continues to hold for arbitrary probability measures, discreteness is not required), we find that

$$N(\mathcal{G}_{3,k,\ell}, L_1(Q), \epsilon) \leq (A/\epsilon)^v,$$

for some $A, v < \infty$. Next, note that under Assumption 3.5 the series of random vectors $\{\xi_{i,t} := (Y_{it}, \mathbf{z}_{it}, Y_{it+j}, \mathbf{z}_{it+j})\}_{t\in\mathbb{Z}}$ is $\beta$-mixing with mixing coefficients $\widetilde{\beta}(t)$ satisfying $\widetilde{\beta}(t) \leq \beta(0 \vee (t-j))$. Since the functions in $\mathcal{G}_{3,k,\ell}$ are uniformly bounded, Lemma C.1 in Kato et al. (2012)(applied with $\delta = 1$ in the notation of that Lemma) yields

$$|Cov(g(\xi_{i,t}), g(\xi_{i,t+j}))| \leq C\tilde{\beta}(j)^{1/2}$$

for a constant $C$ independent of $n, T, i$. For $g \in \mathcal{G}_{3,k,\ell}$ let

$$\sigma_{q,i,j}^2(g) := Var\Big(\frac{1}{\sqrt{q}} \sum_{t=1}^{q} f(Y_{it}, \mathbf{z}_{it}, Y_{it+j}, \mathbf{z}_{it+j})\Big).$$

We have

$$\begin{aligned}\sigma^2_{q,i,j}(g) &= Var(f(\xi_{i,t})) + 2\sum_{j=1}^{q-1}\Big(1-\frac{j}{q}\Big)Cov(f(\xi_{i,1}), f(\xi_{i,1+j})) \\ &\leq C + 2C\sum_{j=1}^{m_T}\Big(1-\frac{j}{q}\Big) + 2C\sum_{j=m_T}^{q-1}\tilde{\beta}(j)^{1/2} \\ &\leq 2(m_T+1)C + 2C\sum_{j=1}^{\infty}\tilde{\beta}(j)^{1/2} \\ &\leq \tilde{C}(m_T+1)\end{aligned}$$

for a constant $\tilde{C}$ independent of $i, n, T$. The claim in (68) follows by an application of Proposition C.2 in Kato et al. (2012). To obtain (69), set $s = 4\log(nT)$ and $q_{n,T} = C\log(nT)$ with $C$ chosen such that $\beta(C) \leq e^{-4}$.

**Proof of** (72) **and** (73) By the proof of Lemma 5 in Galvao et al. (2020), it holds for any $g \in \mathcal{G}_2(\delta)$ that

$$\|g\|_\infty \leq U_2$$

with some constant $U_2 > 0$, and it also holds for large $n, T$ satisfying $\frac{1}{nT} \leq \delta \leq 1$ that

$$\sigma^2_{q,i}(g) = \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{q}}\sum_{t=1}^{q} g(\mathbf{z}_{it}, Y_{it})\right) \leq C_{\sigma,2}\delta\log(nT), \quad i = 1, \ldots, n\,,$$

where $C_{\sigma,2}$ is a constant. Moreover, by the first display in the proof of Lemma 5 in Galvao et al. (2020), it holds for any probability measure $Q$ and any $0 < \epsilon < 1$ that

$$N(\mathcal{G}_2(\delta), L_1(Q), \epsilon) \leq (A/\epsilon)^\nu$$

with some positive constants $A, \nu < \infty$. Invoking Proposition C.2 of Kato et al. (2012) gives (72). To prove (73) pick

$$q_{n,T} := C_1\log(nT)$$

with the universal constant $C_1 \geq 1$ satisfying $b_\beta^{C_1} \leq e^{-2}$, and set

$$\sigma^2_{n,T} := C_{\sigma,2}\log(nT)\delta\,.$$

With this choice (70) holds by definition and we have

$$\frac{q^2_{n,T}}{\sigma^2_{n,T}}\log\left(\frac{q_{n,T}}{\sigma^2_{n,T}}\right) \lesssim \sqrt{T}\log(nT)^{1/2}\log(T)^{1/2} = o(T)$$

so that (71) holds for $n, T$ large enough. Let $s_{n,T} := 2\log n$. By elementary computations using the fact that $\log(n)^3 = o(T)$ by assumption the claim follows by applying the union

bound.

**Proof of** (74) Denote the $k$-th element of vector $\mathbf{z}_{it}$ by $\mathbf{z}_{it,k}$. Define the function $f_k$ via

$$\begin{aligned} f_k : \mathbb{R}^d \times \mathbb{R} &\to \mathbb{R} \\ (\mathbf{z}_{it}, Y_{it}) &\mapsto \mathbf{z}_{it,k}\Big(\mathbb{1}\{U_{it} \le \tau + d_T\} - \mathbb{1}\{U_{it} \le \tau - d_T\} - 2d_T\Big)\,. \end{aligned}$$

By Lemma 11 in Galvao et al. (2020), we have

$$\sup_{i,t,k} |f_k(\mathbf{z}_{it}, Y_{it})| \le C_1, \quad \mathbb{E}[f_k(\mathbf{z}_{it}, Y_{it})] = 0\,,$$

and

$$\sigma^2_{q,i}(f) := \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{q}}\sum_{t=1}^{q} f_k(\mathbf{z}_{it}, Y_{it})\right) \le C_2 d_T |\log(d_T)|\,,$$

where $C_1, C_2$ are constants independent of $i, T$. Note that the constants are independent of $i$, throughout the proof we drop the dependence of $\sigma^2_{q,i}(f)$ on $i$, and denote it by $\sigma^2_q(f)$ instead. Applying Corollary C.1 in Kato et al. (2012), we have for some constant $C$ independent of $i, T, k$ and $q_{n,T} \in [1, \frac{T}{2}]$ and for some $s_{n,T} > 0$

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} f_k(\mathbf{z}_{it}, Y_{it})\right| \ge C\left(\frac{\sqrt{(s_{n,T} \vee 1)}}{\sqrt{T}}\sigma_q(f) + \frac{s_{n,T} q_{n,T}}{T}\right)\right) \le 2e^{-s_{n,T}} + 2T\beta(q_{n,T})\,.$$

Set $s_{n,T} := 2\log n$ and let

$$q_{n,T} := C_1 \log(nT)$$

where $C_1 > 1$ is a constant satisfying $b_\beta^{C_1} \le e^{-2}$. Then, it holds for some large $n$ and $T$ and a small $d_T$ that

$$\frac{\sqrt{(s_{n,T} \vee 1)}}{\sqrt{T}}\sigma_q(f) + \frac{s_{n,T} q_{n,T}}{T} \le \sqrt{\frac{2\log n}{T}}\sqrt{d_T}\sqrt{\log\Big(\frac{1}{d_T}\Big)} + \frac{C_1 \log n \log(nT)}{T}\,.$$

By Assumption 3.11 and $T$ grow at most polynomial in $n$, we have

$$\sqrt{\frac{2\log n}{T}}\sqrt{d_T}\sqrt{\log\Big(\frac{1}{d_T}\Big)} + \frac{C_1 \log n \log(nT)}{T} \lesssim \frac{\log n}{\sqrt{T}}\sqrt{d_T} + \frac{(\log n)^2}{T}\,.$$

Moreover, note that

$$2e^{-s_{n,T}} + 2T\beta(q_{n,T}) \lesssim \frac{1}{n^2} + \frac{1}{n^2 T}\,.$$

Taking the union bound over $i = 1, \ldots, n$ then gives

$$\sup_i \left\| \frac{1}{2Td_T}\sum_{t=1}^{T} \mathbf{z}_{it}\Big(\mathbb{1}\{U_{it} \le \tau + d_T\} - \mathbb{1}\{U_{it} \le \tau - d_T\} - 2d_T\Big)\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{Td_T}}\right).$$

This completes the proof of (74). Now the proofs of all results in Lemma 9.7 are complete. □

### 9.4.2 Proof of Theorem 3.6 (i)

The Lemma C.2 in Chao et al. (2017) can be used in our framework by setting $t = 2, n = T, \xi_m = \kappa, g_n = 0$, which implies the following for each $i \in \{1, \ldots, n\}$

$$\left\{ \sup_{\tau \in \mathcal{T}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 \leq \frac{4\,\|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_1}}{\inf_{\tau \in \mathcal{T}} \lambda_{\min}(\widetilde{J}_i)} \right\} \supseteq \left\{ \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_1} < \frac{\inf_{\tau \in \mathcal{T}} \lambda_{\min}^2(\widetilde{J}_i)}{8\kappa \overline{f'} \lambda_{\max}\big(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\big)} \right\}, \quad (75)$$

where $\widetilde{J}_i := \mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top f_{Y_{it}|\mathbf{z}_{it}}(\mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^* | \mathbf{z}_{it})\big]$, with the function class $\mathcal{G}_1$ defined in (64). By the assumption that $(\log n)^3 = o(T)$ and Assumptions 3.6-3.8, it holds for sufficiently large $n, T$ that

$$\tilde{C}_{\mathcal{G}_1} \sqrt{\frac{\log(nT)}{T}} \leq \frac{\inf_{\tau \in \mathcal{T}} \lambda_{\min}^2(\widetilde{J}_i)}{8\kappa \overline{f'} \lambda_{\max}\big(\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^\top\big]\big)}\,. \quad (76)$$

Define the event

$$\Omega_{\mathcal{G}_1} := \Big\{ \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_1} \leq \tilde{C}_{\mathcal{G}_1} \sqrt{\frac{\log(nT)}{T}} \Big\}. \quad (77)$$

By the relation (75), we obtain that on the event $\Omega_{\mathcal{G}_1}$, it holds that

$$\sup_{\tau \in \mathcal{T}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 \leq C_3 \sqrt{\frac{\log(nT)}{T}}\,,$$

where $C_3 > 0$ is a constant independent of $i, n, T$. Combined with (67) we find that for all sufficiently large $n, T$

$$\mathbb{P}\Big( \sup_{\tau \in \mathcal{T}} \|\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i^*\|_2 \leq C_3 \sqrt{\frac{\log(nT)}{T}} \Big) \geq 1 - \frac{1}{nT}. \quad (78)$$

This completes the proof of Theorem 3.6 (i). □

### 9.4.3 Proof of Theorem 3.6 (ii).

The assumptions made imply that the smallest eigenvalues of the matrices $B_i$ are bounded away from zero uniformly in $i$. Since we work in fixed dimension, it suffices to show that

$$\max_{i,k,\ell} |\widehat{B}_{iT,k,\ell} - B_{i,k,\ell}| + \max_{i,k,\ell} |\widehat{H}'_{iT,k,\ell} - \widetilde{H}_{i,k,\ell}| = o_{\mathbb{P}}(1).$$

We will consider the two sums separately, starting with $\widehat{B}_{iT}$. Note that

$$\begin{aligned}\left|\!\left|\!\left|\widehat{B}_{iT}-B_i\right|\!\right|\!\right|_2 \leq & \left|\!\left|\!\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}(\hat{f}_{it}-f_{it})\right|\!\right|\!\right|_2+\left|\!\left|\!\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}-\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}\big]\right|\!\right|\!\right|_2 \\ & +\left|\!\left|\!\left|\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}\big]-B_i\right|\!\right|\!\right|_2.\end{aligned}$$

The boundmax$_i$ $\left|\!\left|\!\left|\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}\big]-B_i\right|\!\right|\!\right|_2=o(1)$ follows from standard Taylor expansions similarly to the proof of Lemma 8 in Galvao et al. (2020). Further, we have

$$\max_i\left|\!\left|\!\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}(\hat{f}_{it}-f_{it})\right|\!\right|\!\right|_2\leq\kappa^2\max_{i,t}|\hat{f}_{it}-f_{it}|=o(1)$$

by Lemma 9.8. To bound $\max_i\left|\!\left|\!\left|T^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}-\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}\big]\right|\!\right|\!\right|_2$ note that the entries of $\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}$ are uniformly bounded. Thus an application of Lemma C.1 from Kato et al. (2012) shows that $\mathbb{V}\mathrm{ar}\,(T^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it})\leq C_1$ for a constant $C_1$. Now apply Corollary C.1 from Kato et al. (2012) with $s=2\log n, q=c\log(nT)$ for a suitable constant $c$ to obtain $\max_i\left|\!\left|\!\left|T^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}-\mathbb{E}\big[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}f_{it}\big]\right|\!\right|\!\right|_2=o_{\mathbb{P}}(1)$.

Next we proceed to bound $\left|\!\left|\!\left|\widehat{H}'_{iT}-\widetilde{H}_i\right|\!\right|\!\right|_2$. Recall the notation from the paragraph before Lemma 9.7. Observe the decomposition

$$\begin{aligned}[\widehat{H}'_{iT}]_{k,\ell}-[\widetilde{H}_i]_{k,\ell} = & \tau(1-\tau)\frac{1}{T}\sum_{t=1}^{T}\Big\{[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}]_{k,\ell}-\mathbb{E}\big[[\mathbf{z}_{it}\mathbf{z}_{it}^{\top}]_{k,\ell}\big]\Big\} \\ & +\sum_{1\leq j\leq m_T}\Big(1-\frac{j}{T}\Big)\Big(\mathbb{P}_{i,j,T}g_{\hat{\boldsymbol{\gamma}}_i(\tau),k,\ell}-\mu_{3,k,\ell}(\hat{\boldsymbol{\gamma}}_i(\tau),i,j)\Big) \\ & +\sum_{1\leq j\leq m_T}\Big(1-\frac{j}{T}\Big)\Big(\mu_{3,k,\ell}(\hat{\boldsymbol{\gamma}}_i(\tau),i,j)-\mu_{3,k,\ell}(\boldsymbol{\gamma}_i^*(\tau),i,j)\Big) \\ & +\sum_{1\leq j\leq m_T}\Big(1-\frac{j}{T}\Big)\mu_{3,k,\ell}(\boldsymbol{\gamma}_i^*(\tau),i,j)-\sum_{j=1}^{\infty}\mu_{3,k,\ell}(\boldsymbol{\gamma}_i^*(\tau),i,j) \\ & +R_{n,T,k,\ell,i}(\tau) \\ =: & \sum_{j=1}^{4}\Delta_{i,k,\ell,n,T}^{(j)}(\tau)+R_{n,T,k,\ell,i}(\tau).\end{aligned}$$

where $R_{n,T,k,\ell,i}(\tau)$ arises due to the summation range over $T_j$. Note that

$$\sup_{i,k,\ell}|R_{n,T,k\ell,i}|\leq\frac{2m_T^2\kappa^2}{T}=o(1)$$

since $T\geq|T_j|\geq T-m_T$ and $\|g_{\hat{\boldsymbol{\gamma}}_i(\tau),k,\ell}\|_\infty\leq 2\kappa^2$. The bound $\max_{i,k,\ell}\sup_{\tau\in\mathcal{T}}|\Delta_{i,k,\ell,n,T}^{(1)}(\tau)|=o_{\mathbb{P}}(1)$ follows by combining Lemma C.1 and Proposition C.2 from Kato et al. (2012) with

$s = 2\log n, q = c\log n$ for a suitable constant $c$. The bound $\max_{i,k,\ell}\sup_{\tau\in\mathcal{T}}|\Delta^{(4)}_{i,k,\ell,n,T}(\tau)| = o(1)$ follows from the arguments in the last paragraph in the proof of Lemma 12 in Galvao et al. (2020). To bound $\max_{i,k,\ell}\sup_{\tau\in\mathcal{T}}|\Delta^{(3)}_{i,k,\ell,n,T}(\tau)| = o(1)$ note that under Assumption 3.10 the maps $\boldsymbol{b}\mapsto\mu_{3,k,\ell}(\boldsymbol{b},i,j)$ are Lipshitz continuous with Lipshitz constant $\kappa^2$ bounded uniformly in $n,T,i,j,k,\ell$. Thus

$$\max_{i,k,\ell}\sup_{\tau\in\mathcal{T}}|\Delta^{(3)}_{i,k,\ell,n,T}(\tau)| \leq \kappa^2 m_T \max_i \sup_{\tau\in\mathcal{T}}\|\hat{\boldsymbol{\gamma}}_i(\tau)-\boldsymbol{\gamma}^*_i(\tau)\|_2 = O_{\mathbb{P}}\Big(m_T\sqrt{\frac{\log(nT)}{T}}\Big) = o_{\mathbb{P}}(1)$$

by the first part of the theorem and Assumption 3.11. Finally, observe that

$$\max_{i,k,\ell}\sup_{\tau\in\mathcal{T}}|\Delta^{(2)}_{i,k,\ell,n,T}(\tau)| \leq m_T \max_{i,k,\ell,j}\|\mathbb{P}_{i,j,T}-\tilde{\mathbb{P}}_{i,j}\|_{\mathcal{G}_{3,k,\ell}} = O_{\mathbb{P}}\Big(\sqrt{\frac{m_T^3\log(nT)}{T}}\Big) = o_{\mathbb{P}}(1)$$

where we used (69) and the assumption on $m_T$. This completes the proof of Theorem 3.6 (ii). □

### 9.4.4 Technical results used in the proof of Theorem 3.6 (ii)

**Lemma 9.8.** *Let the assumptions stated in Theorem 3.6(i) and Assumption 3.11 hold. Then*

$$\sup_{i,t}|\hat{f}_{it}-f_{it}| = o(1).$$

*Proof of Lemma 9.8* The proof strategy follows from Lemma 11 in Galvao et al. (2020), where we employ the Bernstein inequality for $\beta$-mixing sequences (Corollary C.1 in Kato et al. (2012)). Define $\widehat{e}_{it} := \widehat{f}_{it}^{-1}$ and $e_{it} := 1/f_{it}$. By definitions of $\widehat{e}_{is}$ and $e_{is}$ it holds that

$$\widehat{e}_{is}-e_{is} = \mathbf{z}_{is}^\top\Big(\big(\hat{\boldsymbol{\gamma}}_i(\tau+d_T)-\boldsymbol{\gamma}^*_i(\tau+d_T)\big)-\big(\hat{\boldsymbol{\gamma}}_i(\tau-d_T)-\boldsymbol{\gamma}^*_i(\tau-d_T)\big)\Big)/2d_T\,.$$

By Lemma 9.9and the assumptions $\frac{\log(nT)}{Td_T^2} = o(1), \log(n)^3 = o(T)$ we obtain

$$\max_{i,s}|\widehat{e}_{is}-e_{is}| \leq 2\kappa^2\max_i\big\|\big|B_i^{-1}\big|\big\|_\infty\max_i\Big\|\frac{1}{2Td_T}\sum_{t=1}^T\psi_{i,\tau+d_T}(\mathbf{z}_{it},Y_{it})-\psi_{i,\tau-d_T}(\mathbf{z}_{it},Y_{it})\Big\|_2+o_{\mathbb{P}}(1).$$

Letting $U_{it} := F_{Y_{it}|X_{it}}(Y_{it}|X_{it})$ we see that $\mathbb{1}\{Y_{it}\leq\boldsymbol{\gamma}^*_i(\tau\pm d_T)\} = \mathbb{1}\{U_{it}\leq\tau\pm d_T\}$ and hence

$$\begin{aligned}&\frac{1}{2Td_T}\sum_{t=1}^T\psi_{i,\tau+d_T}(\mathbf{z}_{it},Y_{it})-\psi_{i,\tau-d_T}(\mathbf{z}_{it},Y_{it})\\&\qquad=\frac{1}{2Td_T}\sum_{t=1}^T\mathbf{z}_{it}\Big(\mathbb{1}\{U_{it}\leq\tau+d_T\}-\mathbb{1}\{U_{it}\leq\tau-d_T\}-2d_T\Big)\end{aligned}$$

Thus $\max_{i,s}|\widehat{e}_{is}-e_{is}| = o_{\mathbb{P}}(1)$ by (74). Finally, under the assumptions made we have

$\min_{i,t} e_{it} \geq 1/f_{max}$, see (62). The claim follows by a Taylor expansion of $x \mapsto 1/x$. □

**Lemma 9.9.** *Let the assumptions stated in Theorem 3.6(i) and Assumption 3.11 hold. It holds for every $i \in \{1, \dots, n\}$ that*

$$\hat{\boldsymbol{\gamma}}_i(\tau) - \boldsymbol{\gamma}_i^*(\tau) = -\frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \psi_{i,\tau}(\mathbf{z}_{it}, Y_{it}) + R_{n,T,i}(\tau)\,,$$

*where*

$$B_i := \mathbb{E}[f_{Y|\mathbf{z}}(q_{i,\tau}(\mathbf{z}_{i1}) \mid \mathbf{z}_{i1}) \mathbf{z}_{i1} \mathbf{z}_{i1}^\top], \quad \psi_{i,\tau}(\mathbf{z}, Y) := \mathbf{z}(\mathbb{1}(Y \leq q_{i,\tau}(\mathbf{z})) - \tau\,,$$

*and*

$$\sup_i \sup_{\tau \in \mathcal{T}} \|R_{n,T,i}(\tau)\|_2 = \mathcal{O}_{\mathbb{P}}\left( \frac{(\log(nT))^{5/4}}{T^{3/4}} \right).$$

*Proof of Lemma 9.9* Observe the decomposition

$$\widehat{\boldsymbol{\gamma}}_i(\eta) - \boldsymbol{\gamma}_i^*(\eta) = -\frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \mathbf{z}_{it}(\mathbb{1}(Y_{it} \leq q_{i,\eta}(\mathbf{z}_{it})) - \eta) + r_{i,1}(\eta) + r_{i,2}(\eta) + r_{i,3}(\eta),$$

where

$$r_{i,1}(\eta) := \frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \mathbf{z}_{it}(\mathbb{1}(Y_{it} \leq \mathbf{z}_{it}^\top \widehat{\boldsymbol{\gamma}}_i(\eta)) - \eta),$$

$$r_{i,2}(\eta) := -\frac{1}{T} B_i^{-1} \sum_{t=1}^{T} \Big\{ \mathbf{z}_{it} \Big( \mathbb{1}(Y_{it} \leq \mathbf{z}_{it}^\top \widehat{\boldsymbol{\gamma}}_i(\eta)) - \mathbb{1}(Y_{it} \leq \mathbf{z}_{it}^\top \boldsymbol{\gamma}_i^*(\eta)) \Big) - \int z[F_{Y|Z}(z^\top \widehat{\boldsymbol{\gamma}}_i(\eta) \mid z) - F_{Y|Z}(z^\top \boldsymbol{\gamma}_i^*(\eta) \mid z)] dP^{\mathbf{z}_{i1}}(z) \Big\},$$

$$r_{i,3}(\eta) := -B_i^{-1} \Big[ \int z[F_{Y|Z}(z^\top \widehat{\boldsymbol{\gamma}}_i(\eta) \mid z) - F_{Y|Z}(z^\top \boldsymbol{\gamma}_i^*(\eta) \mid z)] dP^{\mathbf{z}_{i1}}(z) - B_i(\widehat{\boldsymbol{\gamma}}_i(\eta) - \boldsymbol{\gamma}_i^*(\eta)) \Big].$$

Let $R_{iT}^{(1)}(\eta) := r_{i,2}(\eta), R_{iT}^{(2)}(\eta) := r_{i,1}(\eta) + r_{i,3}(\eta)$. Following the arguments in the proof of Theorem 5.1 in Chao et al. (2017) with $n = T, m = p + 1, \xi_m = \kappa, g_n = 0$, and $c_n = 0$ we have almost surely

$$\sup_{\eta \in \mathcal{T}} \|r_{i,1}(\eta)\| \lesssim T^{-1}.$$

Moreover, on the event

$$\max_i \sup_{\eta \in \mathcal{T}} \|\widehat{\boldsymbol{\gamma}}_i(\eta) - \boldsymbol{\gamma}_i^*(\eta)\| \leq \delta$$

we have

$$\sup_{\eta \in \mathcal{T}} \|r_{i,2}(\eta)\| \lesssim \max_i \|\mathbb{P}_{i,T} - \mathbb{P}_i\|_{\mathcal{G}_2(\delta)}$$

and $\sup_{\eta \in \mathcal{T}} \|r_{i,2}(\eta)\| \lesssim \delta^2$ where the constants in $\lesssim$ depend on the constants from Assump-

tion 3.6–3.8 only. Letting $\delta = C_3\sqrt{T^{-1}\log(nT)}$ and recalling (78) and (73) completes he proof. □

*Proof sketch of Theorem 3.8.* The results for the first part can be established by following the arguments given in the proof of Theorem 5.1 in Kato et al. (2012), which are parallel to the step 1-3 in the proof of Theorem 3.2 therein. The results for the second part can be proved similarly to those for the second part of Theorem 3.6. □